AFOSR - TR - 77 - 0333

AD A038184

AD No. ____
DDC FILE COPY



Canyon Research Group, Inc.

# ANALYSIS OF HUMAN FACTORS ENGINEERING EXPERIMENTS: CHARACTERISTICS, RESULTS, AND APPLICATIONS

Charles W. Simon

Technical Report No. CWS-02-76

Research sponsored by the Air Force Office of Scientific Research (AFSC), United States Air Force, under ~~Contract No~~ F44620-76-C-0008. The United States Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright notation hereon.

2313 A4 AFOSR TR-77-0333

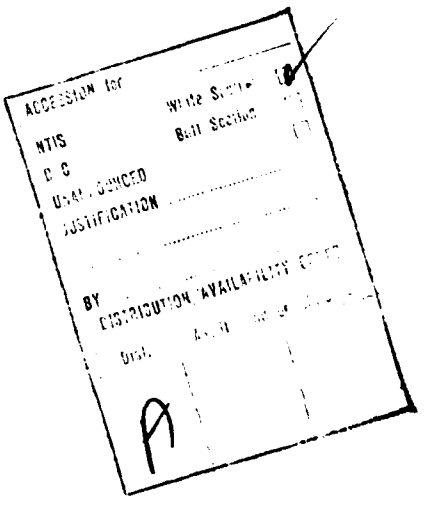
August 1976

Canyon Research Group, Inc.
32107 Lindero Canyon Road, Suite 123
Westlake Village, California 91361

# FOREWORD

How investigators collect their data has an important effect upon the results they obtain. When the president of the Society of Engineering Psychologists, a division of the American Psychological Association, states that research efforts in this field "have been and are insufficient," then it is time to look at what has been done. The data in this report is believed to be the only systematic effort to examine characteristics of a representative group of experiments in human factors engineering, to measure how much information they provided, and to query whether or not the results have ever been used. This information provides a base from which improved experimental methods can be developed. Some of the material presented here has been used in other documents of the advanced methodologies program, particularly in the report entitled: "Economical Multifactor Designs for Human Factors Engineering Experiments."

No effort was made to apply sophisticated statistics to the analyses. In many cases it would not have been warranted and interesting data that did not lend itself to a more elegant treatment might have been discarded. Without considerably more data than was available from the published reports, attempts to examine more complex relationships would have been an unjustified over-analysis of the data. The descriptive data provided in most cases enables some fairly clear-cut conclusions to be drawn.



## ACKNOWLEDGMENTS

The data collection, analysis, and writing for this report has extended over a four-year period, supported in part by the Air Force Office of Scientific Research, Air Force Systems Command, United States Air Force, under prime Contract No. F44620-72-C-0086 with Hughes Aircraft Company and Prime Contract No. F44620-76-C-0008 with Canyon Research Group, Inc.

Drs. Charles Hutchinson, Glen Finch, and Ralph Canter of AFOSR have all been technical monitors during the period in which this report was being produced.

Miss Marilyn A. Wilson (now Mrs. Marilyn Lewis) collected, calculated, and organized much of the basic data extracted from the journals.

Mr. Flynard E. Roberts prepared the computer programs used to list and analyze the data base.

Miss Paula Lintz provided many valuable editorial comments.

Their help is gratefully acknowledged.

TABLE OF CONTENTS

TABLE OF CONTENTS (Cont)

TABLE OF CONTENTS (Cont)

TABLE OF CONTENTS (Cont)

LIST OF TABLES AND FIGURES

LIST OF TABLES AND FIGURES (Cont)

# I. INTRODUCTION

Human factors engineering as a distinct discipline was born during the Second World War. A "human factors specialist," as defined in a System Development Corporation national salary survey, is one who:

> "...establishes, conducts, coordinates and applies major research studies in the social, behavioral or physiological sciences; contributes to design, development and operation of man-machine, weapons or other complex systems concepts; utilizes psychological principles of human behavior, knowledge of human physical and mental characteristics, abilities and limitations, and principles of human engineering."
>
> (Kraft, 1961)

Despite a prolific output during the past 30 years, experiments relating human performance to equipment parameters have shown a relatively low information-to-cost ratio. While human factors practitioners have made significant contributions toward easing the job of the human operator and making system performance more effective, the contributions of the human factors scientists -- the experimenter -- have been modest. Today, one has to search diligently among piles of published papers to find among the trivia and the isolated facts, data that is sufficiently generalizable to answer questions concerning the design of future systems and to do so quantitatively. Jack Adams (1972), in his presidential address to the Society of Engineering Psychologists, American Psychological Association, summarized this condition quite succinctly: "Our research efforts have been and are insufficient. The future of engineering psychology is in jeopardy unless we examine realistically the state of our knowledge and ask what we must do to strengthen it."

Surprisingly, in spite of criticisms leveled against human factors engineering experiments, there has been little serious effort to examine the methodologies used in these experiments to

detect the sources of their weaknesses. On the contrary, even when discontent has been expressed with the usefulness of experimental results, the reaction has been to collect more data in the same way.

## THE PURPOSE AND EXTENT OF THE PRESENT REPORT

In this report, the methods, results, and applications of a large body of human factors experiments are examined. The information obtained not only provides the first known quantitative evaluation of past human factors research but also provides a basis for establishing the methodological requirements for future research.

This report is divided into five sections:

I. Introduction - describes the sample from which the data was drawn and the data structure, plus a classification scheme that enables data from diversified experiments to be normalized.

II. Characteristics of the Experimental Plans - describes characteristics of the experimental data-collection efforts, such as: types of experimental designs used; numbers of factors; levels per factor; allocations of subjects and trials; handling of sequence effects; and estimation of error variance.

III. Experimental Results - analyzes results to determine: how effective the experiments were in accounting for observed variations in performance; and how well empirical evidence supports certain principles of advanced experimental methods.

IV. Research Application Survey - summarizes the results of a survey of the experimenters to determine to what extent the experimental data was ever applied to real systems.

V. Summary and Discussion - reviews the primary results from preceding sections and interprets their significance in terms of an improved experimental methodology.

## THE SAMPLE

The sample of experiments in this report was taken from the journal, Human Factors. This journal informs its readers that it "...publishes original articles which increase and diffuse the knowledge of man in relation to machine and environmental factors in all their ramifications, pure and applied." An analysis was made of 141 papers published in this journal from Volume 1, No. 1, September 1958 through Volume 14, No. 3, June 1972 in which formal experiments were described and the results presented in some summary inferential statistical tables. In 23 of these papers, 34 analyses of variance (ANOVA) were not included here because the analyses fell into one of the following categories:

- A partial analysis of a more complete analysis. (8)
- A reprint of an analysis from a study not described in the article. (3)
- A study of a single factor at two levels. (4)
- No data (7), incomplete data (9), or incorrect data. (2)
- Chi square analysis. (1)

As a result of these exclusions, the data in this report is based on the text of the 118 papers and the 239 analyses of variance tables* in these papers. Although the data for several analysis of variance tables may have been collected at the same time, either the independent variables or the performance measure changed. Therefore, each analysis of variance is treated and referred to here as if it represents the results of a different experiment.

---

*Journal references for the 239 analyses of variance included in the study are listed in Appendix A.

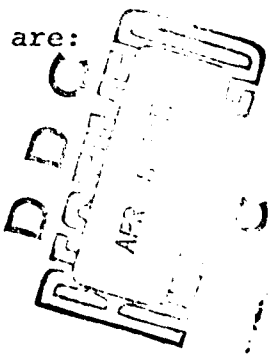
Is this sample representative?

There are a great many human factors experiments produced yearly in industry and government laboratories that are never published in the journal. Many of these are given national security classifications which limit their accessibility. How representative, therefore, is the group of experiments covered in this report? While there is no way to accurately answer this question, the fact that many of the same people are involved in both suggests that those papers published in Human Factors and those published as company reports and government documents do not differ materially insofar as their experimental methodologies are concerned. The human factors community is relatively small, probably fewer than 2,000 people. Members of the Human Factors Society who publish in the journal, as well as those on its editorial staff, are among those doing research in industrial and government laboratories. Although human factors personnel now span two generations, the nature of their formal research training remains essentially unchanged and in many cases, research is still dominated by the first generation. If there is a difference in quality, it would seem that published research might be expected to be better than the unpublished efforts.

Which organizations conducted and funded the research?

Eight types of organizations are identified. These are:

- Army (A)
- Consulting (C)
- Air Force (F)
- Government (non-military) (G)
- Industry (I)
- Navy (N)
- University (U)
- Other (e.g., private research organizations) (X)



The matrix in Table 1 shows how the 239 experiments are distributed among these organizations. The organization to which the principal investigator belonged when he conducted the experiment determines the "organization doing the research." The type of organization which funded the project was taken from published acknowledgments or, when no other source was given, was considered to be the same as the organization where the study was done. Approximately one-third of the experiments were performed in industry, one-fourth in universities, and one-fifth by consulting companies. Less than ten percent of the studies were performed by the three military agencies, and approximately seven percent by non-military government-related agencies. Industry supported approximately two-thirds of its own research, while universities and consulting companies supported approximately one-third of theirs. The remaining research was supported by governmental agencies with the military supporting approximately twice as much as non-military agencies.

How were the 239 experiments distributed in the journals?

In the time-period covered by this analysis, the editor-in-chief of the journal was changed three times. Of the 239 experiments, 16 percent were published between 1958 and 1963, 64 percent between 1964 and 1968, and 20 percent between 1969 and 1972, each period representing a different editorship.

As shown in Table 2, the 239 analyses of variance were not distributed evenly among the 118 papers. Over half of the papers reported only one ANOVA table and 94 percent of the papers contained four or fewer tables. It is recognized that characteristics of experiments designed by the same experimenter in the same study are likely to be correlated. This in turn could distort the results of an analysis of these characteristics. However, since all of this data is used to analyze the experimental results

TABLE 1. ORGANIZATIONS WHICH PERFORMED OR FUNDED THE RESEARCH (N=239)

| Organization which performed the research | PERCENTAGE OF TOTAL IN EACH CATEGORY | Organization which funded the research | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | A | C | F | G | I | N | U | X | Percentage Performing |
| Army (A) | | 5.4 | | | | | | | | 5.4 |
| Consulting (C) | | 2.5 | 6.3 | 2.1 | 6.7 | | 2.5 | | | 20.1 |
| Air Force (F) | | | | 1.3 | | | | | | 1.3 |
| Govt. Non-military (G) | | | | | 2.9 | | | | | 2.9 |
| Industry (I) | | 3.8 | | 6.7 | 0.8 | 22.2 | 1.3 | | | 34.8 |
| Navy (N) | | | | | | | 2.5 | | | 2.5 |
| University (U) | | 2.1 | | 5.0 | 5.1 | | 5.8 | 10.8 | | 28.8 |
| Other (X) | | | | | 4.2 | | | | | 4.2 |
| Percentage Funding | | 13.8 | 6.3 | 15.1 | 19.7 | 22.2 | 12.1 | 10.8 | 0 | 100.0 |

TABLE 2. DISTRIBUTION OF 239 EXPERIMENTS (ANOVAS) AMONG 118 PAPERS

| | Number of ANOVAS | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 |
| Number of papers having indicated number of articles per paper | 64 | 26 | 13 | 8 | 0 | 3 | 0 | 2 | 1 | 1 |
| Percentage of papers | 54 | 22 | 11 | 7 | 0 | 2 | 0 | 2 | 1 | 1 |
| Cumulative percentage | 54 | 76 | 87 | 94 | 94 | 96 | 96 | 98 | 99 | 100 |

(Section II of this report), all 239 ANOVAs were analyzed in this first part. A cursory examination of the parameters in the studies with multiple ANOVAs reveals that collectively they distribute themselves on either side of the median values found in the analysis. When the data in this report was analyzed after removing multiple entries of the four studies containing from 8 to 10 ANOVAs, leaving only a single entry for each one, no critical shift in the results was observed.

## Considerations and qualifications in the analyses

The reader should be aware of some of the decisions that were made in examining the analyses of the experiments. The purpose here is not to confuse or to destroy confidence in the data, but to be candid about the arbitrariness that sometimes existed. At no time did it appear that the major conclusions that could be reasonably drawn from the data would be affected, no matter which choices were made.

Experiment-content data. This report is concerned with methods and the scientific and pragmatic effectiveness of the methods. The factual contents of the experiments were never considered.

Lack of sophisticated and inferential analyses. Only simple, descriptive analyses were performed. Several efforts to make more complex analyses, such as relating results to design characteristics with a regression equation, were aborted. The data was too irregular and imbalanced for such an effort. This does not negate the value of drawing conclusions from the patterns in the data. The danger was in over-analyzing, and this was avoided.

Reliability. Some categories in the tables involve an extremely small number of cases. These are often included for completeness and interpreted with caution. The n is noted.

Working with percentages. In more sophisticated analyses there may be reasons to transform percentage data using arcsine transformations to produce a more normalized distribution. This was considered totally unnecessary for the descriptive data here since most of the information could have been extracted if only an ordered scale had been used.

## THE DATA STRUCTURE

The contents of experiments relate to such a wide variety of topics that before a common analysis could be performed it was necessary to standardize the sources of variance. This standardization is possible since the sources of variance in all human factors engineering experiments can be assigned to one of four general categories (or their interactions), i.e., variances attributable to: 1) equipment, system processes, and environment; 2) the people involved; 3) certain temporal effects; and 4) methodology introduced into the experiment by the investigator (Figure 1). Thus the classification scheme of sources of performance variance include:

### Equipment factors (E)

These are the variables associated with the physical equipment, system, environment, and processes, i.e., the ones that generally make up the bulk of the "independent variables" in a human factors engineering experiment. For convenience, these are all referred to in this report as "equipment" factors.

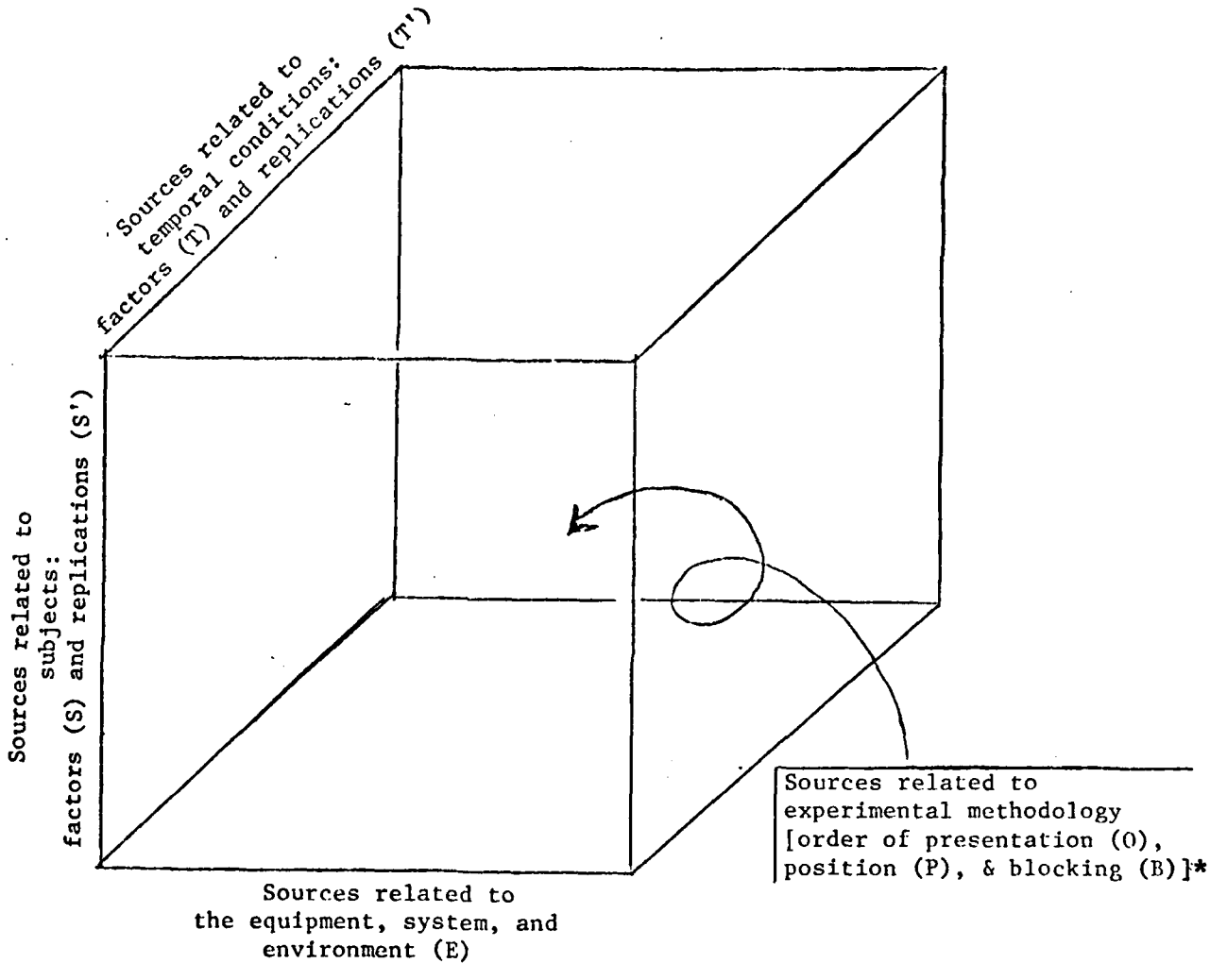



FIGURE 1. SOURCES OF VARIANCE IN HUMAN FACTORS ENGINEERING EXPERIMENTS

*These sources do not comprise a fourth dimension. Instead, they are confounded with the other three dimensions and/or interactions.

Subject sources

These are sources of performance variations stemming from the observers used in the experiment. Subject-related variances are divided into two types:

Subject factors (S). When specific subject characteristics are designated and included as independent variables in the experimental design, these are treated as "subject" factors. For example, the experiment may be designed to answer such questions as: Do pilots perform differently with a new display or control than non-pilots? The pilot/non-pilot conditions are levels of a subject factor. Should equipment be designed differently to compensate for age or sex differences among operators? When age or sex differences are systematically varied, they are subject factors. How much difference is there in operator performance on different devices as a function of prior training? The differences in training -- if it is isolated -- is a subject factor.

Subjects as a form of replication (S'). When no specific subject characteristics are included as an experimental variable, however, the assumption is made here that the effect of subjects is considered by the experimenter to be of secondary importance, their presence in the experiment being a form of design replication to improve the reliability of the data.

In most experiments, if subject variability is estimated, the results are usually presented in the ANOVA table and discussed no further. Under these circumstances, subjects are not considered to be a factor, an intended source of variance. Instead, theoretically they are expected to be a homogeneous sample from some population varying only incidentally in performance.

It is sometimes argued, however, that knowledge of subject variance is a measure of individual differences and will be

important when the results of the experiment are applied to the real-world. In practice, unfortunately, this is seldom the case; the variance attributable to subjects is seldom a useful piece of information. To be useful:

- The subjects in the experiment must be truly representative of the population to which the data is to be extrapolated.
- Representativeness must be based on multiple characteristics.
- The values of the characteristics for the sample and population must be known.

These conditions seldom exist for the typical human factors experiment. When multiple subjects are run as replications, the chances that they are representative of the population are slight for the following reasons:

- The average number of subjects in these human factors experiments run around nine, although the population may be made up of thousands.
- In many cases, no systematic sampling of subjects is, or can be made. Those that are available are used.
- When subjects have been selected, it is often on the basis of a single label (for example, Air Force pilot). Seldom are additional considerations (such as amount of flying time, types of aircraft flown, etc., that can cause wide variations in performance) taken into account.
- Quantitative descriptions of population and samples are seldom available making it impossible to adequately identify to what subportion of the population experimental results refer.

In addition, the artificiality of the experimental situation also influences the performance of individual subjects. Part of the variability between subjects' performance reflects the basic

instability of a mean score for subjects who are often still learning how to handle the experimental situation as the study progresses and do so at different rates. It is highly presumptive to believe that the variance associated with the performance of a small group of subjects used to replicate an experimental design has much permanency or practical validity insofar as the experimental results may be applied to the real world. Thus, when subject variability is in an experiment undimensionalized, it is considered to be an identifiable but relatively uninformative source of variance.

Temporal sources

These are sources of variations associated with changes occurring from trial to trial when a subject is tested on the same experimental condition. Temporal variances are divided into two types:

Temporal factors (T). When the investigator is specifically interested in the effects of repeated trials on performance, the variance associated with trials is considered to be an independent variable, a factor. For example, if the intent of the experiment is to find out operator learning functions while using alternative devices, trial variances would be considered to be a "temporal" factor. In vigilance studies, the detrimental effect of performing over many trials on the same condition would cause the variability associated with trials to be treated as a temporal factor.

Trials as a form of replication (T'). In certain circumstances, an investigator will test the same subject on an experimental condition for several trials primarily to ensure a reliable measure. There is no serious interest in the change in performance from trial to trial, except for the hope that it won't be too great.

At times, under these circumstances, an average performance value over trials per condition is obtained and used in subsequent data analyses to give the outward illusion of stability; at other times, the values for multiple trials are used in the analysis to increase the degrees of freedom and to provide an error estimate for tests of statistical significance. In these cases, trials are not considered to be a factor.

Experiment-related sources

Some investigators introduce systematic variations into an experiment, not to study their effects, but to control unwanted effects and be able to eventually isolate them from the effects of primary interest. Those most commonly found are:

Order-of-presentation effects (O). When human subjects are tested sequentially on a number of experimental conditions, residual effects from a preceding condition may combine with the effects of the condition that follows, distorting performance. To be able to compensate for these condition-to-condition transfer effects, which may be either positive or negative, an investigator may systematically change the order in which the conditions are presented to the subject. If order is varied essentially orthogonal to the experimental factors, its effect can be removed in the data analysis. In experiments not concerned with transfer (or in training research), the hope is that order of presentation has no effect.

Position effects (O). When the order of presentation of conditions to different subjects is systematically varied using a Latin square arrangement, three sources of variance can be removed from the two-dimensional matrix of performance values: experimental conditions, order (and/or subjects), and the effect of the position in a series of trials. Experimenters often refer to position effects as "trial" effects. This differs from the

trials (T) factors described above in that position (Q) is a measure of the variability in mean performance from trial to trial averaged across different combinations of subjects and experimental conditions. The position effect might be thought of as a measure of "learning-to-learn," or a generalized change in the adaptation to the experimental situation independent of particular experimental conditions or individuals.

Blocking effects (B). The experimental conditions are divided into groups or blocks in such a way that conditions within blocks are more homogenous than between blocks due to the presence of unwanted and irrelevant sources of variance (Simon, 1970).

Interaction effects

With the exception of position (Q) and, sometimes, order (O) effects, the above sources of variances may interact with one another. Simple and higher order interaction effects can be isolated as independent sources of performance variance.

The following types of interactions were actually isolated during the analysis whenever the experimenter's original data was sufficient for this purpose:

Equipment variable interactions: EE, EEE, EEEE, etc

Equipment by subject-factor interactions: ES, EES, etc

Equipment by subject-replication interaction: ES', EES', etc

Equipment by temporal factor interaction: ET, EET, etc

Equipment by trial-replication interaction: ET', EET', etc

Subject-replication by trial-replication interaction: S'T'

Equipment by subject-replication by trial-replication interaction: ES'T', EES'T', etc

In general, the notations in this report exclude the repeated letters, e.g., EES' would be written ES'.

Pooled effects (P)

In some analyses when certain effects can be isolated, but are not, the investigator will combine, or pool, them into a single variance. This is most commonly done with interaction effects, but may include trials and subject effects when these sources are considered to be replications. To complete the classification scheme of sources of variance, this pooled category must be included.

## II. CHARACTERISTICS OF THE EXPERIMENTAL PLANS

The information in this section summarizes characteristics of the 239 experiments published in the journal, Human Factors. The information answers the following questions:

- How large is the experimental space (normalized in terms of the number of factors and levels per factor)?
- How large is the data collection effort (based on the total number of observations that were made in each experiment)?
- What types of experimental designs are used (including considerations of the methods for handling order of presentation and the sources used in the denominator of tests of statistical significance)?

### THE SIZE OF THE EXPERIMENTAL SPACE

There is always the question of how much of the real world must be simulated in an experiment in order to explain most of the performance variability present in the task being studied. It seems reasonable to assume that the more characteristics of the real world critical to the particular task being investigated that are included in an experiment, the more generalizable the results are likely to be. However, in the past, both real and fantasized difficulties in collecting the large quantities of data required for truly multifactor experiments have kept the extent of the experimental space relatively small proportionate to the real world space that is being simulated.

Two analyses were made to determine how large an experimental space was covered in the 239 experiments included in this analysis. Data was obtained on:

- The number of factors in the experiments.
- The number of levels per factor in the experiment.

How many factors are studied in each experiment?

The independent experimental factors fall into three categories: equipment (E), temporal (T), and subject (S). These are the ones the experimenter systematically varies to measure their effects on performance. In Table 3, the number (and proportion) of experiments containing each of these types is shown. Eighty percent of the experiments in this study contained only equipment factors. Fifteen percent of the experiments also included a temporal factor and four percent included a subject factor. In only two cases were both a subject and a temporal factor included in experiments with equipment factors. In two other cases, only temporal factors and no equipment factors were studied. That relatively few non-equipment factors have been included in human factors engineering experiments is probably a residue of the fact that historically these experiments have been conducted primarily to find ways of optimizing performance by improving the equipment rather than by selecting or training the operator.

In Figure 2, the number (and percentage) of experiments containing from one to seven factors is shown. This information is analyzed in two ways: 1) the number of equipment factors only in the experiment, and 2) the number of equipment, subject, and temporal factors combined in the experiments. By either breakdown shown, more than 60 percent of the experiments investigated the effects of only one or two factors. Less than three percent of the experiments investigated the effects of five or more factors.

Grouping experiments by the number of equipment factors only. In subsequent analyses, the experiments are usually divided into sub-groups based on the number of factors in the experiment. There is a choice of dividing them in terms of only the number of

TABLE 3. NUMBER OF EXPERIMENTS CONTAINING NONE TO SEVEN EQUIPMENT FACTORS AND NONE TO ONE SUBJECT OR TEMPORAL FACTOR

| Number of subject and temporal factors in each experiment: NUMBER OF EXPERIMENTS IN EACH CATEGORY | Number of equipment factors in each experiment: 0 | 1 | 2 | 3 | 4 | 5 | 7 | TOTAL NUMBER (Proportion of total) |
|---|---|---|---|---|---|---|---|---|
| 0 temporal factor<br>0 subject factor | 0 | 47 | 80 | 49 | 11 | 4 | 1 | 192<br>(.803) |
| 1 temporal factor<br>0 subject factor | 2 | 19 | 12 | 2 | 0 | 0 | 0 | 35<br>(.146) |
| 0 temporal factor<br>1 subject factor | 0 | 3 | 1 | 4 | 2 | 0 | 0 | 10<br>(.042) |
| 1 temporal factor<br>1 subject factor | | 2 | | | | | | 2<br>(.008) |
| TOTAL NUMBER (Proportion of total) | 2<br>(.008) | 71<br>(.297) | 93<br>(.389) | 55<br>(.230) | 13<br>(.054) | 4<br>(.017) | 1<br>(.004) | 239<br>(1.000) |

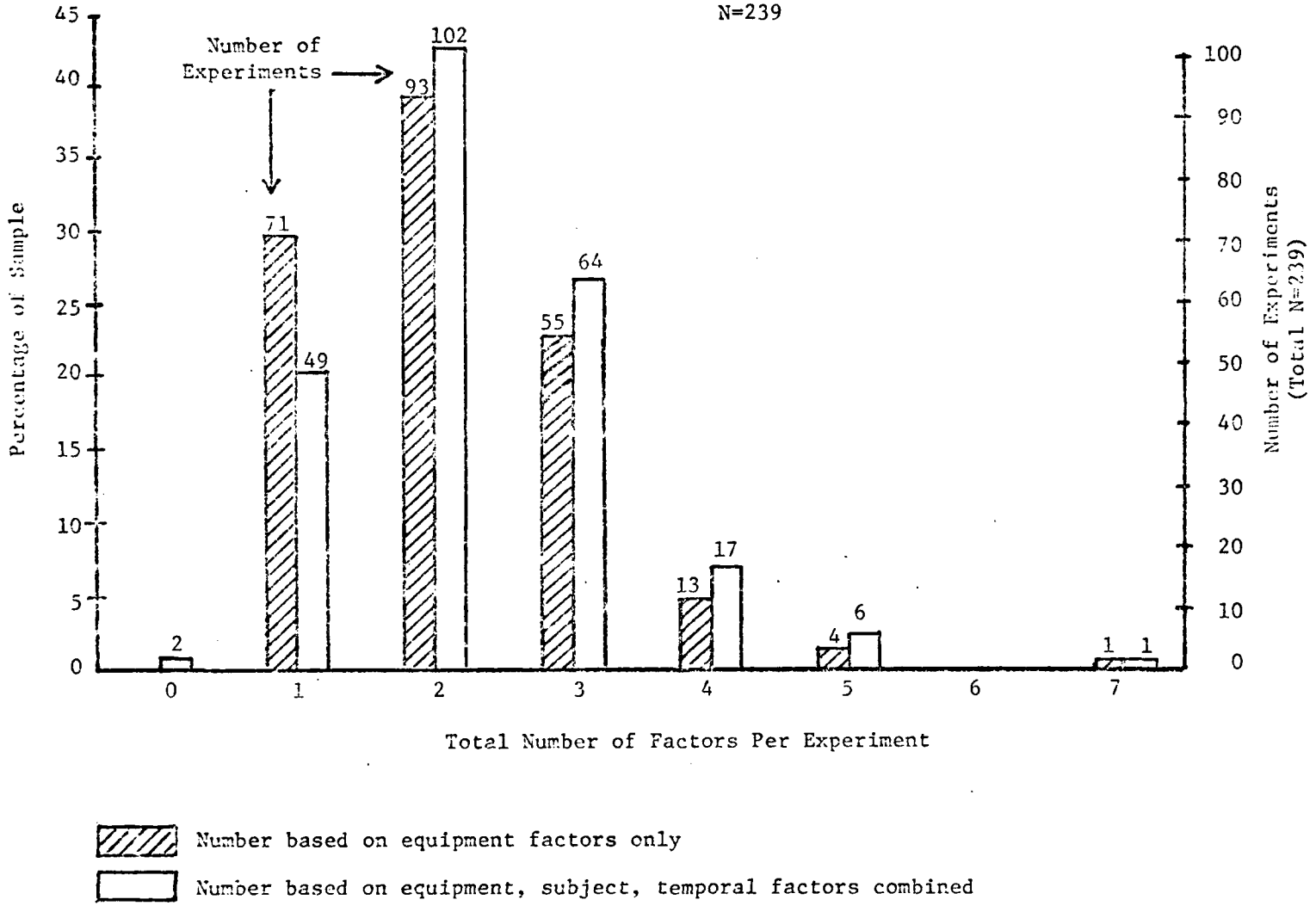



FIGURE 2. DISTRIBUTION OF NUMBER OF FACTORS PER EXPERIMENT

equipment factors in the experiment or in terms of the number of factors from all sources (E, S, and T) in the experiment. Most of the tables in this report use only the first classification scheme. Tabular numbers would be different, of course, when the second classification scheme is used (primarily shifting values from the "one-factor" category to the "two-factor" category, depending, for example, on whether a temporal factor in an experiment was not or was considered as part of the "number of factors" in the experiment). For most analyses, the data was examined both ways even though only the results of the "equipment factors only" scheme were supplied here. Inspection of the two sets of data provided no reasons to believe that the shift in numerical values would change the major conclusions that were drawn from the data.

How many levels of each factor were measured?

The term, "levels", is used to refer to the different conditions per factor on which performance is to be measured. Levels may refer to conditions that are quantitatively or qualitatively different. When the experimental conditions of a single factor can be ordered, they represent the levels of a quantitative factor, and if any of a range of values can be used, they represent a continuous quantitative factor. If only certain values can be used, possibly because only certain values are manufactured, these would be levels of a quantitative, discrete factor. When the experimental conditions cannot be ordered and represent different categories of some factor, they are considered "levels" of a qualitative factor. Ten', twenty', and thirty-foot resolution of a display would represent levels of a quantitative factor; different kinds of military targets (e.g., personnel, tanks, missile sites) would represent levels of a qualitative factor.

With qualitative factors, the number of levels depends on the number of categories, conditions, or types of situations that

the experimenter wishes to investigate. With quantitative factors, theoretically, any reasonable number of levels can be selected and the numbers chosen determine how complex a relationship can be fit between the independent and dependent variables. No careful analysis was made to determine whether the levels considered here were qualitative or quantitative; however, a casual examination suggested that equipment factors with more than five levels were qualitative.

In this sample, in a total of twelve experiments, all but one subject factor had two levels. The one exception had three levels.

Temporal factors had the greatest number of levels. This is not surprising since many of these experiments were interested in the effects of extended work periods on performance. Work periods measured in trials which represent the levels as units of time. The median number of trials for the 37 examples in which the number of temporal factors considered was five while the maximum was 120.

The number of levels per equipment factor was determined for the 501 factors in all experiments, grouped according to the number of E factors per experiment. The frequencies with which from two to ten or more levels per factor were used are shown in Table 4. From this table it can be seen that:

- The median number of levels per factors decreased from four to two as the number of factors in an experiment increased from one to seven.
- Slightly more than two-thirds of the total number of factors occurred at two or three levels.
- Across all factors, the median number of levels studied was three.

TABLE 4. FREQUENCY AND PROPORTION OF EQUIPMENT-FACTOR LEVELS

| Number of equipment factors per experiment | Number of levels per equipment factor: 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | ≥ 10 | [Actual number of levels in cases ≥10] | TOTAL |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 14** (44.8%) | 18 | 20* | 10 | 2 | 3 | 1 | 1 | 2 | [10, 12] | 71 |
| 2 | 83 (68.8%) | 45* | 19 | 20 | 13 | 1 | 0 | 1 | 4 | [10,10,14,31] | 186 |
| 3 | 74 (73.9%) | 48* | 28 | 5 | 6 | 0 | 2 | 1 | 1 | [30] | 165 |
| 4 | 20 (78.8%) | 21* | 7 | 2 | 1 | 0 | 0 | 0 | 1 | [20] | 52 |
| 5 | 14* (85.0%) | 3 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | [15] | 20 |
| 7 | 7* (100.0%) | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | | 7 |
| Total frequency | 212 | 135* | 75 | 38 | 22 | 4 | 3 | 3 | 9 | | 501 |
| Percentabe | 42.3 | 26.9 | 15.0 | 7.6 | 4.4 | .8 | .6 | .6 | 1.8 | | |
| Cumulative percentage | 42.3 | 69.2 | 84.2 | 91.8 | 96.2 | 97.0 | 97.6 | 98.2 | 100 | | |

*Median for the row.

**Four one-factor, two-level studies were not included in this analysis. All others were one equipment factors with two-levels plus some other (S or T) factor.

Although the median value would not change, had four experiments studying four one-factor at two-levels (which had been excluded from the analysis) been included, the proportion of one-factor studies at one and two levels would have risen to 48 percent.

What was the total number of measurements made in these experiments?

Practical as well as scientific considerations affect the number of factors, levels per factor, and replications an experimenter uses in his experimental design. The availability of both time and money will generally place a limit on the total number of measurements that can be made. The experimenter is, therefore, faced with the problem of balancing and fitting the three design components that affect the number of total observations as best he can to optimize the information he will obtain from the study.

The expression "total number of measurements" as used here refers to the total number of degrees of freedom (plus one) supplied in the published ANOVA tables. This operational definition, therefore, will not take into consideration the measurements that were made and then averaged together to create the values used in the ANOVA tables. Some of the experiments in which the total degrees of freedom are few, are of this type. Nor does this definition of total number of measurements fully encompass situations in which a "single", summary measurement is obtained over a period of time. Some examples of this type of measurement are the number of words that are misunderstood while listening to a 15-minute taped conversation under conditions of background noise, or the errors made while driving an automobile over a five-mile course, or in the integrated mean square error during a two-minute tracking task.

The total number of measurements made in the experiments are shown in Table 5 as a function of the number of equipment factors studied. In general, these numbers represent minimum values for the reasons cited above.

That the median number of observations increases as the number of factors being investigated increases is not surprising. What is most striking, however, are the considerable differences in the sizes of experiments investigating the same number of factors and the very large number used in some cases to study a very small number of factors.

A rough indication of the efficiency of an experiment is given by finding the ratio of the number of observations that were made to the minimum number required to approximate a second degree space with a polynomial, Taylor series expansion. This polynomial would contain only the mean plus all terms for the linear, quadratic, and linear-by-linear interaction effects of a $3^n$ experiment. The following ratios were obtained:

| Number of Factors in Experiment | Median Total Observations | Minimum Required | Ratio |
|---|---|---|---|
| 1 | 72 | 3 | 24 |
| 2 | 180 | 6 | 30 |
| 3 | 192 | 10 | 19 |
| 4 | 768 | 21 | 37 |

The weighted mean of the ratios for these experiments yielded a ratio of 26. While there are no absolute standards for evaluating this number, it does show that a tremendous amount of data was collected relative to that which would have been needed to extract the greatest portion of information content in the experiments.

TABLE 5. TOTAL NUMBER OF OBSERVATIONS PER EXPERIMENT

(Percentile Distribution)

| Number of equipment factors per experiment (Number of experiments) | TOTAL NUMBER OF OBSERVATIONS PER EXPERIMENT | Lowest | 1st Quartile | Median | 3rd Quartile | Highest |
|---|---|---|---|---|---|---|
| 1 (71) | | 18 | 25 | 72 | 163 | 7680 (1120)** |
| 2 (93) | | 15*** | 72 | 180 | 250 | 2016 (1944)** |
| 3 (55) | | 24 (48* | 64 | 192 | 480 | 9600 (3024)** |
| 4 (13) | | 48 (256)* | 192 | 768 | 1416 | 3888 (3888)** |

*Fewest total number of observations when the experimental data had not been averaged across subjects or trials before the analysis was made.

**Highest number when experiments with Trial (T) factors are excluded.

***This is correct. Fifteen subjects were distributed unevenly into a 2 x 2 design.

How often is the basic experimental design repeated by measuring different subjects?

The number of replications based on subjects refers to the number of times performance measurements are made on the same experimental condition using a different subject. Since the same or different subjects may be tested on all of the conditions of an experimental design, the number of subject replications is not necessarily related to the number of subjects used in the experiment. For example, in a 3 x 3 factorial design that has been replicated three times, i.e., 26 degrees of freedom in the total design, 3, 9, or 27 subjects might have been used depending on whether the same three subjects were tested on all of the nine conditions; or three groups of three subjects were each tested separately on a different set of three conditions of one factor but all of the conditions of the second factor; or 27 different subjects were tested in sets of three, a different set being assigned to each of the experimental conditions.

In half of the 239 experiments, the basic experimental design was replicated nine or more times using different subjects. In 25 percent of the experiments, the basic design was replicated 12 or more times using different subjects per condition. The maximum number of subject replications was 64; that is to say, 64 different subjects were tested on each condition in the experiment.

How often is the basic experimental design repeated by testing each subject with extra trials on each condition?

"Trial replications" refer to the number of times the same subject made repeated measurements on the same experimental condition. It does not refer to multiple trials employed to study a temporal factor. In over half of the 239 experiments, each subject was tested only once. In 25 percent of the experiments,

each subject was tested two or more times on each experimental condition. The maximum number of trial replications was 70.

The trial replications reported here are those reported in the published ANOVA tables. As such they probably represent a conservative estimate since, in some cases, multiple trials were averaged to arrive at a single value for the conditions used in the analysis of variance.

## EXPERIMENTAL DESIGNS

Practically all human factors engineering experiments in this survey used some form of a factorial design and an analysis of variance model.* The variations among these designs can be conveniently classified -- in addition to the characteristics already discussed -- by the way in which subjects, trials, and experimental conditions were interrelated. In addition, special arrangements of the experimental conditions are used to offset the effects of the order when they are presented serially to the subjects.

---

*Edgington (1974) found a similar emphasis in seven journals of the American Psychological Association "primarily concerned with original empirical research." From 1948 to 1972 inclusively, he found that 91% of the articles involved statistical inference, and by 1972, 71% of those articles used analysis of variance. Simple one-way analysis of variance techniques were used frequently but 88% of the analysis of variance articles employed repeated-measures design (where a subject was tested on more than a single experimental condition) or factorial designs.

The analyses that follow provide:

- The frequency with which different types of experimental designs were used.
- A list of the different methods of handling order-of-presentation effects.
- A list of the different "error" terms used in the tests of statistical significance.

How frequently were different types of experimental designs used?

Experimental designs can be conveniently classified by the way subjects and trials are introduced into the design relative to the experimental conditions. Although each method used can affect the experimental results and the interpretation of those results, no explanations for the choice nor a discussion of possible implications of the choice were given in any of the experiments.

Classifying techniques of subject deployment. Subjects can be introduced into a design in four ways:

- Each of a group of subjects is used as his own control and is tested on every experimental condition. The effect due to subjects can be removed along with the interaction among subjects and other experimental factors. This is the classical factorial design. (49% of the experiments were of this type).
- Each of a group of subjects is tested across all conditions of some experimental factors but not all of them. Variability among subjects is the average variability among subjects within the same conditions. Conversely, the reliability of differences among conditions varied between subjects must be tested independently of those varied within subjects. This is the nested or split-plot design. (28%)

- Different groups of subjects are tested on each experimental condition. Neither subject variability nor the interactions between subjects and experimental conditions can be isolated in these designs. Instead, the average variance within groups is obtained. (17%)

- A single subject -- either the same or a different individual -- is tested on each experimental condition. In several experiments where this occurred, the single measurement for each condition was actually the performance of a crew of men acting as a unit. All effects of subjects are totally confounded with experimental effects. (6%)

Classifying multiple-trial designs. Fifty-seven percent of the experiments tested each subject only once on an experimental condition. The remaining 43 percent of experiments using multiple trials could be subdivided as follows:

- In slightly more than one-third of these experiments, multiple trials were introduced for the purpose of studying temporal factors. (15% of the 239 experiments were of this type).

- Of the remaining 28% of the multiple-trial designs, in which multiple trials were merely a form of replication, over two-thirds were introduced sequentially. This means that when a subject performed on an experimental condition once, he would perform on the same condition on the very next trial or sequence of trials. (19%)

- In the remaining group of multiple-trial designs, repeated measures by the same subject on the same experimental conditions were made periodically, only after other experimental conditions had been tested between replications of the same condition. (9%)

Table 6 shows how the designs of the 239 experiments are proportioned among the four subject and two trial replication plans.

TABLE 6. FREQUENCY AND PERCENTAGE OF TIMES DIFFERENT EXPERIMENTAL DESIGNS WERE USED

| Distribution of subjects across conditions: NUMBER AND PERCENTAGE IN EACH CATEGORY | Number of times subject was tested on each experimental condition: MORE THAN ONCE | ONCE | TOTAL SUBJECTS |
|---|---|---|---|
| Every subject tested on all conditions | 47 (20%) | 63 (29%) | 115 (49%) |
| Some subjects tested on some conditions and other subjects tested on other conditions | 27 (11%) | 40 (17%) | 67 (28%) |
| Different groups of subjects tested on each condition | 16 (77%) | 25 (10%) | 41 (17%) |
| Single subject (or group treated as individual) tested on each condition | 13 ( 5%) | 3 ( 1%) | 16 ( 6%) |
| TOTAL TRIALS | 103 (43%) | 136 (57%) | 239 (100%) |

How was order of presentation handled in these experiments?

In some experiments, provisions were made for systematically controlling the order in which experimental conditions were presented to the same subjects. When the same subject is tested sequentially on a number of experimental conditions, it is necessary to do something to minimize carry-over effects from one condition to the other. In behavioral studies, this carry-over effect may be a general learning-to-learn phenomenon or a general fatigue or boredom effect that develops as the study progresses,* or it may be a specific transfer effect in which performance on one experimental condition is influenced by the characteristics of experimental conditions preceding it. If these sequence effects are not controlled, they become confounded with and thereby distort the effects of interest. Three techniques have commonly been used by behavioral scientists to reduce order-of-presentation effects: randomization, counterbalancing, and procedural controls between runs. The many variations on these basic techniques used in the experiments under review are listed in Table 7.

In describing their experimental designs in Human Factors, the investigators gave little justification for the particular method selected for handling order-of-presentation effects. None of the investigators used a class of experimental designs that would enable first-order transfer effects to be isolated from main effects (Simon, 1974). Among the experimenters who were concerned enough to do something about the order-of-presentation effect (as shown by the way they planned their experiments), only a few actually analyzed their data to remove these or related effects statistically.

---

*There are other causes that are not subject-related, such as drift in electronic equipment or slow changes in the environment or other factors that can affect performance.

TABLE 7. METHODS USED TO HANDLE ORDER-OF-PRESENTATION EFFECTS

1. The problem of order was never mentioned in the descriptions of experimental design and procedures.

2. In three cases, order was determined by the nature of the task rather than by the experimenter (e.g., in a reconnaissance task, the targets were embedded in the radar imagery and for a moving-scene presentation, the order was fixed by their positions on the film).

3. Some investigators' only description of how they handled order effects was by stating they "juggled," "varied," or "semi-randomized" their presentation.

4. When a group of experimental conditions was subsequently repeated in the same study, some investigators "randomized" the order of treatment presentations on the first set of trials and:

    a. Repeated that order on subsequent sets of trials; or

    b. Reversed that order on subsequent sets of trials; or

    c. Used a different randomization on subsequent sets of trials.

5. When more than one subject was tested under the same set of experimental conditions, some investigators "randomized" the order in which the treatments were presented, and:

    a. Maintained the same "randomized" order for all subjects; or

    b. Used a different "randomized" order for each subject; or

    c. Used a different "randomized" order for groups of subjects.

6. Some investigators acknowledged that they used "restricted randomization" to determine their order of presentation. This meant that after some order had been determined essentially by chance, treatments that occurred in a sequence of positions which the investigator believed might create unwanted effects were rearranged.

7. When the same experimental conditions were presented more than once to the same or different subjects, some investigators preferred more systematic methods of determining order of presentation. The concept of "counterbalancing" is frequently referred to by this group. Counterbalancing in these studies took several forms, such as:

    a. Seeing that each condition occurred an equal number of times in each position of a sequence of trials.

    b. Seeing that each condition occurred an equal number of times in every trial position and for every subject and (with the exception of the treatments at each end of the series) appeared once before and after every other condition.

    Familiar experimental designs to handle these systematic methods of counterbalancing are the Latin square and Graeco-Latin square designs. There is insufficient information in most cases to determine when the same or different Latin squares were used within the same experiment.

8. Systematic and random ordering were both used in some studies.

Classifying designs that systematize presentation orders. In some experiments, the designs included provisions for systematically controlling the order in which a sequence of experimental conditions were presented to the same subjects. Subsequently, in the analysis, this and related effects were isolated. The designs of these experiments could be classified into two types in which:

- Order of presentation was varied systematically and the order effect was isolated from the effects of the experimental conditions. (1.7% of the 239 experiments had this feature.)
- Order of presentation was varied systematically using a Latin-square design in such a way that order, experimental conditions, and position effects could be isolated (but not their interactions). Of these designs, there were some that: a) isolated both the order effects and the sequential position effects. (3.3%)
  b) Isolated only the position effects. (1.3%)

In the above situations, when only one subject was tested on each order of presentation, the effects of subjects and order were confounded. When several subjects were tested on each order of presentation, the effect of order could be isolated from the effect of subjects within orders.

When an order-of-presentation factor was included in the experimental design but the experimenter did not isolate it during analysis, the following types of analyses were made instead:

- The effect of positions was removed but the effect of presentation order was left confounded with other effects.
- The effects of presentation order were left confounded with the effects of subjects.
- The effects of presentation order were confounded with interaction effects between equipment factors and trials, equipment factors and subjects, and equipment factors, trials, and subjects - which were isolated.

In very few studies were the results of the effects of presentation order discussed.

What sources of variance were used in the denominator of the tests of significance?

Tests of statistical significance are made by dividing the variance associated with the factor of interest by a variance which the experimenter attributes to chance. One speaks of this chance variance as the "error term." However, the selection of an error term is not always a straightforward procedure in multi-factor designs and in psychological research, seldom a chance effect. Binder (1955) wrote: "Among the various treatments of psychological statistics one finds a good deal of confusion and discrepancy in the recommended procedures for selecting an error term in the analysis of variance. In all too many cases the obtained significance or insignificance of the experimental results depends as much upon the particular test used as upon the sampling data."

In Table 8 are listed the actual sources of variance that were used as an "error" term in tests of significance of the experiments analyzed in this report. No judgment is made here as to their correctness. By making the particular selection, however, each experimenter markedly affected both the results and the interpretation of his data. This was never taken into consideration in any discussion of results.

TABLE 8. SOURCES OF VARIANCE USED IN HUMAN FACTORS EXPERIMENTS AS AN ESTIMATE OF ERROR VARIANCE

- The next higher significant $E_i \times E_j$ was used to test $E_i$.
- The next higher significant $E_i \times E_j \times E_k$ was used to test $E_i \times E_j$.
- The higher order interactions (E x E etc, ) were pooled (without ever calculating their effects).
- The highest E x E x ... E interaction was used.

- All S' x E interactions pooled (residual).
- $E_i$ x S' interaction, to test $E_i$; $E_i \times E_j$ x S' interaction to test $E_i \times E_j$.
- S' combined with S' x E interactions.
- Between S's (within groups).
- Between S's (within groups) x E interaction.

- T' (within groups) x E interaction.
- T' summed with T' x E interaction.
- T' x E interaction.
- T' (within groups).

- E x S' x T' interaction.
- E x T' interaction summed with S's (within groups) effects.
- Any non-statistically significant interactions pooled.
- "Expected mean square."

## III. EXPERIMENTAL RESULTS

How good were these 239 experiments? Simon (1975) proposes to evaluate the "goodness" of an experiment by the information it produces. One method of evaluating information is to determine what proportion of the performance variance is explained by the experimental variables. In this section, this measure is used to evaluate the experiments. The experiments are also analyzed for characteristics of the interactions that relate to assumptions made when advanced experimental designs are used in human factors research (Simon, 1973).

### EVALUATION CRITERION - ETA SQUARED

It was necessary to find some quantitative measure that could be applied across all of the experiments being studied that would be independent of their content, yet would provide relevant data for the evaluation process. Although many psychologists still consider obtaining "statistically significant" experimental variables as at least one indication of the goodness of the experiment (Bozarth and Roberts, 1972), that criterion is almost totally ignored in this report. Hays (1966, p. 300) wrote: "It is a grave error to evaluate the "goodness" of an experiment only in terms of the significance level of its results." Lykken (1968, p. 158) also noted that "statistical significance is the least important attribute of a good experiment."

Too much has been written about the misapplication and misinterpretation of significant tests to discuss it here (Bakan, 1971; Kleiter, 1969; Lykken, 1968; Nunnally, 1960; Rozeboom, 1960). As Nunnally (1960, p. 643) said: "If rejection of the null hypothesis were the real intention of psychological experiments, there usually would be no need to gather data."

For this report, the analyses of variance in the articles reviewed were reanalyzed using another statistic, eta squared. The results of this reanalysis are the basis for much of the discussion that follows. Eta squared is a descriptive measure of the proportion of total variance accounted for by specified components of the experiment.

Eta squared is calculated by dividing the sum of squares for the particular source of variance in question by the total sum of squares. The proportion is a descriptive index of the strength of the relationship between a source of variance and performance, and is meaningful only within that particular sample.

Another measure that might have been used is omega squared. It is an inferential measure of how much of an effect a factor would have in the population based on the results from the experimental sample (Hays, 1966, p. 547). It adjusts the estimate of an effect on the basis of the size of the error variance and the number of degrees of freedom involved. There are several forms of omega squared depending on the experimental design used as well as certain statistical assumptions made in developing the equation (Vaughan and Corballis, 1969). However, for the purposes of this report, eta squared is considered to be the more appropriate statistic to use because:

- It provides a direct measure of the data in each experiment and needs to make no assumptions about a hypothetical population. (This is not the case with omega squared.)
- Since the calculation uses no error term, a decision need not be made as to what should be used to estimate error. Nor is it necessary to recalculate the values used in the published data, if the experimenter failed to use the more technically correct error term.

- Since eta squared is always equal, or greater than, omega squared, it provides the most optimistic estimate of the contribution of each source of variance. Thus, any results of this analysis are biased in favor of the experiments.

- The measure is simple, intuitively understandable, and familiar. Its square root is a correlation between a factor and performance. With a one-degree-of-freedom factor, it is a Pearson product moment correlation. With more than one degree of freedom, it is a correlation ratio, or eta.

Eta squared can be calculated for each component of variance that can be isolated within the experiment. Of the experiments included in this report, only the sources already isolated by each experimenter in his analysis of variance table were used in this analysis. However, in order to make the results from each experiment standardized and capable of being combined, the categories described in Data Structure in Section I of this report -- i.e., equipment, subjects, trials, order, and interactions -- were substituted for the real world names of the variables and even the composition of the source the experimenters designated as "error" was properly identified.

Interpreting eta squared, with qualifications

What distribution of eta squared would we expect in the ideal experiment? Ordinarily E, S, T, ES, and ET as a group would be expected to account for most of the performance variance in good experiments. This is because these are all experimental factors (and their interactions) and are supposed to represent the only conditions that the experimenter purposefully and systematically varied. All other sources are supposed to be held constant or introduced to offset certain unwanted experiment-induced effects.

In a human factors engineering experiment, E and its interactions would be expected to account for most of the performance

variance attributable to the experimental factors, since the majority of these studies are interested primarily in optimizing performance through appropriate equipment designs. When subject and temporal factors are introduced into equipment design studies, it is usually to detect the presence of critical ES and ET.

S' and T' (representing replicated measures of performance on the same experimental condition) should have very small effects. Subjects for replication (S') are supposed to be a homogeneous representation of a single population, unlike the situation in which subjects (S) are treated as an experimental factor and classified into sub-populations each containing one or more specific characteristics. Similarly, when a subject is tested repeatedly on the same condition for several trials to increase data reliability, the effects over time (T') are expected to be negligible, unlike the case when there is interest in a change in performance over time (T) as a result of -- for example -- learning or fatigue. The effects of differences in the order in which experimental conditions are presented to subjects can be isolated from other effects of primary interest; the O should usually be small. Although order has been systematically varied in the same way an experimental factor would be, its introduction into human factors engineering experiments (as opposed to a training study) is more of a precaution if order proves to have a large effect. In practice, the experimenter hopes O will be small as a result of the care with which he planned and ran his experiment.

While experience with eta squared as a measure of experimental quality is still low, certain interpretations seem reasonable. For example, when the equipment variables in a human factors engineering experiment fail to account for a large proportion of the total variance in the experiment in which subjects and trials are introduced for replication purposes only, the experimenter has failed in some way to optimize his study. Because the values are relative, a low proportion could have occurred for a number of

reasons. For example, the experimenter may not have selected a homogeneous sample of subjects and the unidentified variables contributing to their heterogeneity inflated the error term. It may be that the experimental factors were in fact trivial and accounted for little, relative to the random error. It may be that in the data collection, measurement errors occurred that inflated the error term. It may be that the experimenter failed to isolate known sources of variance from his error term.

Thus, when the proportion of variance accounted for by the experimental variables is small relative to the "unexplained" variance, it is reasonable to believe the experiment was a poor one for one reason or another. On the other hand, if the proportion of variance accounted for by the experimental variables is large, while this may indicate either a restricted experimental design* or a well conducted experiment, it does not mean that the experiment was necessarily a good one if the ultimate criterion is that it must produce useful information.

This is because eta squared does not reflect how much of the real world is represented by the experiment. Although equipment factors may account for a large proportion of the total performance variance in a well conducted experiment, when the number of factors in the laboratory experiment are either so few or so unimportant that they represent only a small proportion of the performance variance in that task in the real world, the experimental data actually will be explaining a relatively small proportion of the performance variance under operational conditions.

---

*An experimental design that provides no estimate of error, e.g., an unreplicated factorial or fractional factorial design, can be analyzed so that all sources of variance are accounted for by experimental variables. This may not in fact be true, but any error variance will be hidden within the other effects and non-isolatable.

How small is small? We may wonder how small an eta squared might be in the experiment and still be important in the real-world (where its relative magnitude is bound to be reduced). Cohen (1969) discusses this in his book on the statistical power of tests in the behavioral sciences, pointing out how the proportion of variance accounted for, and the product moment correlation, are mathematically related. The proportion of variance accounted for equals the square of the correlation. Emphasizing that his suggestions are arbitrary, Cohen suggests that a "small" effect of a single variable would correlate .10 with performance (yielding an eta squared of .01) which would not be perceptible on the basis of casual observation. A "medium" effect, he suggests, would correlate .30 with performance, yielding an eta squared of .09. This effect he says "would be perceptible to the naked eye of a reasonably sensitive observer" (p. 77). A "large" effect would be defined as a correlation of .50, yielding an eta squared of .25. However, as Cohen himself notes, he has drawn his examples and based his levels on examples from the mental-personality-social measurement field, i.e., the field of testing. He writes: "One can, of course, find higher values of r in behavioral sciences. Reliability coefficients of tests, particularly of the equivalence variety, will run generally higher. Also, if effects in highly controlled "hard" psychology (e.g., psychophysics) [and engineering psychology] are studied by means of r's, they would frequently be distinctly higher than .50." (p. 77-78). Of course, these are only rules of thumb for there is no theoretical answer to the question posed at the beginning of this paragraph. It all depends on what proportion of the total performance variance in the real world is accounted for by the variables in the experiments, a fact we can only learn empirically.

When an investigator intends to use eta squared to compare the relative effects of his variables, he should be cautious about rejecting a variable as being "unimportant" just because the

proportion of variance it accounts for is small. In screening experiments (Simon, 1973), for example, we do not screen in order to discard small-effect variables; instead, we screen so that early in a program we can expend our efforts in building a data base, i.e., a framework to which other studies can be related, based on the more important variables. We temporarily ignore the small-effect variables with the intent of examining those of interest later [See page 57 of this report for further considerations in the interpretation of small-effect variables.] The real determinant of whether or not a variable is important is not whether it is statistically significant, or even whether its eta squared is large, but whether under operational conditions the observed absolute effect has practical significance.

"Unexplained" variances

Throughout the discussion of this analysis, references will be made to the proportion of data accounted for by the experiment and the proportion not accounted for, or "unexplained." The term, unexplained, has a particular meaning that should be understood in the context in which it is used. Here, unexplained is generally identified with sources of variance that were unintended and unidentified by the experimenter, who delegates them without comment to what he labels "error." Most typical of these are the interactions between subjects and trials and between subjects and/or trials and equipment factors when subjects and trials were treated as replications in the experiment. This is a rather conservative definition of "unexplained variance," since it does not include subject and trial main effects, nor order-of-presentation effects, when actually their presence in any magnitude reflects a failure on the part of the experimenter to control these unwanted sources of variance. To this extent, unexplained as used here is somewhat synonymous with "irrelevant" or "unwanted" sources of variance, neither planned for, nor identified by, the experimenter.

## SYMBOLOGY AND OTHER CONSIDERATIONS IN THESE ANALYSES

The symbology of E, S, T, S', T', O, Q, and some interactions as described in Section I will be used in the discussions of the analyses that follow. At times, E may be used to represent any equipment-only source of variance, whether from main or equipment-by-equipment interaction effects. ES and ET are used as general forms when any number of equipment factors interact with subject or temporal factors and ES', ET', and ES'T' when any number of equipment factors interact with subject and trial replication.

The sources of variance for which eta squared is calculated are those isolated and published by the investigators of each experiment. In each case, the generic name (e.g., E, S, T, etc.) will be substituted for actual content names. Similarly, the investigators' inferences of what were statistically significant are always used (without concern for standards or whether the proper analysis was made).

In tables in which the proportions of variance are shown, a zero proportion actually means that the value was less than one-half of one percent. No effort was made to transform the proportion data; there was no practical reason to. The qualifications cited in Section I apply to this data.

## WHAT DO THE RESULTS OF THESE EXPERIMENTS REVEAL?

The following questions are answered regarding this sample of experiments:

- What proportion of the total performance variability in the experiment is accounted for by the major sources of variance: E, S, T, ES, ET, S', T', O, and Q?

- Do specifically selected subject characteristics, or groups of presumably homogeneous subjects used for replication, account for more of the performance variance?

- What proportion of total performance variability remains "unexplained?"

- How do proportions of variances of main effects distribute themselves?

- What proportion of the total performance variance is accounted for by the main and interaction effects? What degree polynomial is needed to account for the functions relating main effects to performance?

- What percentage of the effects accounting for one percent or less of the total variance in an experiment is statistically significant?

What proportion of the variability in performance is accounted for by major sources of variance: E, S, T, ES, ET, S', T', O, and Q?

The proportions of variance accounted for by each of these major sources of variance are reported in Table 9. The data is divided into sub-groups based on the number of equipment factors (from one through five) in the experiments. The number of experiments on which each group was based is indicated since the number varies from one source of variance to the other (i.e., all sources may not always be included in every experimental design or, because of pooling, some sources were not individually analyzable). For each source and each group, the smallest and largest proportions of variance accounted for are specified, as well as the amount of variance accounted for by 25, 50, and 75 percent of the experiments in each group. The reader is encouraged to study Table 9 and draw his own conclusions, although the discussion below may serve as a guide.

TABLE 9. PROPORTION OF VARIANCE ACCOUNTED FOR BY E, S, T, ES, ET, S', T', O, and Q PER EXPERIMENT

Proportion of variance x 100

| Source of variance | Number of E factors per experiment | Number of experiments | Percentile level | | | | |
|---|---|---|---|---|---|---|---|
| | | | Lowest | 1st Quartile | Median | 3rd Quartile | Highest |
| Equipment | 1 | 71 | 0 | 5 | 16 | 56 | 92 |
| (E) | 2 | 92 | 1 | 16 | 32 | 64 | 98 |
| | 3 | 50 | 7 | 32 | 44 | 66 | 99 |
| | 4 | 10 | 32 | 46 | 68 | 83 | 94 |
| | 5 | 4 | 37 | -- | 66 | -- | 90 |
| Subject | 1 | 5 | 0 | -- | 4 | -- | 9 |
| (S) | 2 | 1 | -- | 1 | -- | -- | -- |
| | 3 | 4 | 0 | -- | 1 | -- | 5 |
| | 4 | 2 | 0 | -- | (1) | -- | 2 |
| | 5 | 0 | -- | -- | -- | -- | -- |
| | 0 | 2 | 3 | -- | (10.5) | -- | 18 |
| Temporal | 1 | 21 | 1 | 3 | 8 | 15 | 53 |
| (T) | 2 | 10 | 1 | 1 | 2 | 5 | 38 |
| | 3 | 0 | -- | -- | -- | -- | -- |
| | 4 | 0 | -- | -- | -- | -- | -- |
| | 5 | 0 | -- | -- | -- | -- | -- |
| Equipment x subject factor interaction (ES) | 1 | 5 | 0 | -- | 3 | -- | 4 |
| | 2 | 1 | -- | -- | 1 | -- | -- |
| | 3 | 4 | 0 | -- | 4 | -- | 5 |
| | 4 | 4 | -- | -- | -- | -- | -- |
| | 5 | 0 | -- | -- | -- | -- | -- |

*Experiments in which any source of variance was pooled were not included in any analysis in this table.

Table 9 (Cont)

Proportion of variance x 100

| Source of variance | Number of E factors per experiment | Number of experiments | Percentile level | | | | |
|---|---|---|---|---|---|---|---|
| | | | Lowest | 1st Quartile | Median | 3rd Quartile | Highest |
| Equipment x temporal factor interactions (ET) | 1 | 21 | 0 | 1 | 3 | 12 | 19 |
| | 2 | 10 | 0 | 0 | 2 | 4 | 5 |
| | 3 | 0 | -- | -- | -- | -- | -- |
| | 4 | 0 | -- | -- | -- | -- | -- |
| | 5 | 0 | -- | -- | -- | -- | -- |
| Subjects for replication (S') | 1 | 68 | 2 | 20 | 38 | 57 | 93 |
| | 2 | 89 | 0 | 16 | 34 | 54 | 83 |
| | 3 | 44 | 1 | 6 | 9 | 24 | 65 |
| | 4 | 13 | 1 | 3 | 9 | 11 | 53 |
| | 5 | 3 | 0 | -- | 1 | -- | 29 |
| Trials for replication (T') | 1 | 16 | 0 | 2 | 3 | 9 | 72 |
| | 2 | 14 | 0 | 0 | 2 | 5 | 17 |
| | 3 | 7 | 0 | -- | 2 | -- | 33 |
| | 4 | 0 | -- | -- | -- | -- | -- |
| | 5 | 1 | -- | -- | 20 | -- | -- |
| Order of presentation (O) | 1 | 7 | 0 | -- | 1 | -- | 23 |
| | 2 | 1 | -- | -- | 7 | -- | -- |
| | 3 | 3 | 6 | -- | 6 | -- | 24 |
| | 4 | 1 | -- | -- | 3 | -- | -- |
| | 5 | 0 | -- | -- | -- | -- | -- |
| Positions (Q) | 1 | 7 | 1 | -- | 2 | -- | 14 |
| | 2 | 0 | -- | -- | -- | -- | -- |
| | 3 | 4 | 1 | -- | 4 | -- | 11 |
| | 4 | 0 | -- | -- | -- | -- | -- |
| | 5 | 0 | -- | -- | -- | -- | -- |

Results and comparisons across all sources. From the data in Table 9, the following observations can be made:

1. The equipment factors (E) and subjects for replication (S') account for most of the variance, on average. As the number of E factors in the experiment increase, the proportion accounted for by the equipment effects increases and by the subjects for replication decreases, on average.

   Interpretation: These numbers are relative values. The decrease in the proportion accounted for by subjects for replications does not necessarily mean that subjects become less variable as more equipment factors are introduced; instead, it is more likely that they are maintaining a constant absolute degree of variability, but accounting for a smaller proportion of an expanding absolute total variance as each critical equipment factor is added. In cases where S' accounts for most of the total variance, either the equipment factors had trivial effects or the subjects were, in fact, not homogeneous.

2. Within any group, based on the number of equipment factors studied, the proportions accounted for by the equipment factors range from relatively little to practically all of the total variance. In some of the experiments in which only three or fewer equipment factors were studied, these factors failed to account for more than an incidental amount of the observed variations in performance. In all groups, there were always some experiments that accounted for practically all of the observed variance.

   Interpretation: In a properly conducted experiment, essentially all of the variance should be accounted for by only those factors that the experimenter systematically

varies. Some of the published experiments were poor in this regard, signifying that either trivial factors had been selected for the investigation or that the data-collection process and experimental design were inferior. As more factors are added, however, the chances of including more non-trivial effects increase.

3. On average, neither subject nor temporal factors accounted for much of the variance in these experiments. There were some experiments, however, where a temporal factor had a sizeable effect on performance.

   Interpretation: Whatever subject characteristics the experimenters thought might make a difference, apparently they did not in these experiments. Not surprisingly, when an extended number of trials were included in the experiment -- that's what a T factor is -- performance decrement in a vigilance study or performance increment in a training study did occur.

4. Interactions between equipment and subject factors (ES) and equipment and temporal factors (ET) were for all practical purposes negligible, on average. There were a few cases where conditions of the equipment factors showed proportionately differential effects when studied over an extended number of trials. These values are based on a relatively small sample.

   Interpretation: These interactions reflect the magnitude of the main effects, suggesting that in general they were ordinal rather than disordinal interactions.

5. The different sources related to changes in performance over trials, on average, accounted for only a small proportion of the variance. Yet for each source, whether it

was trials for replication (T') position (Q), or order of presentation (O), there were some cases for each in which the size of an effect was large.

Interpretation: In the case of the T' and P effects, this suggests that some trend, such as learning or fatigue had not been adequately controlled, and in the case of Q and O effects, it was possible that when they were large, these may have been due to S x E interactions with which they were confounded.

Which have the larger effects on performance -- selected subject characteristics or subjects used for replicating?

Theoretically when an experiment is replicated by running a number of subjects selected from a single population, there is an implicit assumption that the subjects are homogeneous. While the presence of "individual differences" is acknowledged, when subjects are used as a form of replication, in theory, any variability among subjects is primarily a chance effect. On the other hand, when an investigator singles out specific subject characteristics and includes them as factors in his experimental design, it is done because he suspects that they may have a practical influence on the performance of the task under investigation.

A comparison was made between the proportions of variance accounted for in every experiment by S and S' where both sources of variance occurred. In Table 10 the proportions of variance for pairs of S and S' obtained in ten experiments are listed. It is obvious that in this sample, purposefully created heterogeneous groups accounted for only a miniscule proportion of the variance, while presumably homogeneous groups of subjects accounted for an

TABLE 10. COMPARISON OF PROPORTIONS OF VARIANCE ACCOUNTED FOR BY S AND S' IN THE SAME EXPERIMENT

| Number of factors in experiment | Subject factors (S) | Subjects as replications (S') |
|---|---|---|
| 1 | 4 | 63 |
| | 4 | 67 |
| | 0* | 51 |
| 2 | 1 | P** |
| 3 | 5 | 58 |
| | 1 | 7 |
| | 1 | 6 |
| | 0 | 7 |
| 4 | 2 | 53 |
| | 0 | 11 |
| Median | 1 | 51 |
| Mean | 1.8 | 36 |

Legend:

Values in table represent proportion of variance multiplied by 100.

Each line represents a different one of ten experiments.

*A zero value represents a proportion less than 0.005.
**Subject variability was pooled with other sources and not isolatable.

exceptionally large proportion. It is apparent that investigators in these experiments were not identifying the subject characteristics having the largest effects on performance. When subject-as-replications accounted for more than half the variance in an experiment, it suggested that either the equipment factors that had been selected were not very important, or that the subjects were not really homogeneous. In those investigations, it would appear that system performance would have been improved more readily by emphasizing personnel selection over equipment design (if a choice had to be made).

What proportion of the total performance variability remained "unexplained"?

Theoretically, in an experiment, only the sources of variance systematically introduced by the experimenter should account for performance variability. These sources may be equipment factors (E), subject factors (S), and temporal factors (T), and their interactions, along with systematic variations used to compensate for artificially created order (O) and position (Q) effects. Even when subjects (S') and trials (T') are introduced only to replicate the design and to estimate error, realistically, if their effects were isolated in the experimental design we should not consider them "unexplained" sources of variance.

Once the proportions of variance attributable to all of these known sources have been isolated from the total performance variance, if they are present in the experimental design, what is left is defined here as the proportion of "unexplained" variance (UV). Thus,

$$UV = 1 - (E+S+T+ES+ET+ST+S'+T'+O+Q)$$

UV thus represents the combined effects of all left-over sources of variance that were not specifically identified and isolated by the investigator. Ordinarily, they are also the sources of

variance which would be difficult to interpret meaningfully even if they had been isolated.

An ordered distribution (from lowest to highest) was prepared of the proportions of unexplained variance for each experiment, broken into groups based on the number of E factors under investigation. In Table 11, the lowest, median, and highest values are noted along with those at the .90 percentile.

Too detailed an analysis and interpretation of this data would be dangerous; however, some obvious conclusions can be drawn. Perhaps the most startling observation to be made from Table 11 is how large some of the proportions of unexplained variance are. For example, in at least one experiment involving three equipment factors, more than .99 of the total performance variance remained unexplained by any of the factors introduced into the experiment by the investigators. But even the median values (ranging from .18 to .33) make one wonder what was really learned from all of the effort that must have gone into these experiments, for it must be remembered that these values, while not small in themselves, did not include the proportions of variance due to S', T', O and Q, all of which might be explainable but which provide no information insofar as the design of a device or system is concerned. Thus, the proportion accounted for by the unexplained variance is inversely related to experimental quality. Some experiments left little unexplained (or unaccounted for). This may mean that the irrelevant sources of variance were well controlled or, as discussed earlier, this result could have been achieved artificially by averaging out sources of variance associated with subjects and subject interactions.

TABLE 11. PROPORTION OF TOTAL VARIANCE WHICH IS "UNEXPLAINED"*

| Number of E factors per experiment | Number of experiments involved | Proportion of variance x 100 | | |
|---|---|---|---|---|
| | | Lowest | Median | Highest |
| 1 | 71 | 0** | 21 | 84 (67)*** |
| 2 | 93 | 0 | 22 | 99 (53) |
| 3 | 55 | 0 | 30 | 97 (71) |
| 4 | 13 | 3 | 33 | 52 (45) |
| 5 | 4 | 10 | 18 | 47 |

*Unexplained proportion of variance is what's left over when: [1 - Σ (E,S,T, ES,ET,ST,S',T',0,Q]. Thus, what's left over may contain P and all interactions involving S' and T'. Of course, depending on the experimental design, all sources of variance may not be present in a particular experiment.

**A zero value represents a proportion less than 0.005.

***Numbers in parentheses are the proportions at the 90th percentile.

How do the proportions of variance of main effects in multifactor experiments distribute themselves?

Reference is made in the statistical literature to the "principle of maldistribution" (Budne, 1959). This principle states that the proportions of variance accounted for by a large number of factors in an experiment will be distributed exponentially. This is an important principle, if true, because it suggests that for most tasks, a relatively small number of factors will account for most of the performance variance.

In Table 12, the proportion of variances accounted for by each main effect in the 13 four-factor, four five-factor, and one seven-factor experiments are presented, ordered from largest to smallest. While neither four nor five factors are hardly "a large number of factors," an examination of the table shows fairly clearly that, with only a few exceptions, the proportion of variance accounted for by the main effect in a single experiment varied considerably. In those experiments in which this was not the case, that is, when the variability among factors was slight, the total proportion accounted for by the equipment factors was already relatively small.

When the proportions accounted for by each main effect are ordered from the highest to lowest for each four-factor and five-factor experiment and the columns are averaged, the distributions of the mean values approximate exponential curves within the accuracy of the limited amount of data. The purpose here is not to verify the accuracy of the particular mathematical model, but merely to provide what empirical evidence is available to show that proportions are not distributed equally and that relatively few factors seem to account for most of the variance that can be explained by the experimental variables. This effect is even more vividly illustrated when the interaction values are included in the distributions.

TABLE 12. PROPORTION OF VARIANCE ACCOUNTED FOR BY INDIVIDUAL MAIN EFFECTS IN 4-, 5-, AND 7-FACTOR EXPERIMENTS*

A. Four-factor experiments (13 experiments)

| | | | |
|---|---|---|---|
| 69 | 9 | 5 | 0 |
| 43 | 18 | 8 | 2 |
| 34 | 24 | 7 | 0 |
| 56 | 6 | 1 | 0 |
| 27 | 15 | 14 | 1 |
| 33 | 10 | 7 | 4 |
| 37 | 18 | 2 | 0 |
| 29 | 11 | 4 | 0 |
| 19 | 6 | 5 | 3 |
| 12 | 10 | 7 | 3 |
| 11 | 9 | 9 | 2 |
| 16 | 5 | 1 | 0 |
| 5 | 1 | 1 | 1 |
| Mean 29.8 | 10.9 | 5.5 | 1.2 |
| Median 29 | 10 | 5 | 1 |

B. Five-factor experiments (4 experiments)

| | | | | |
|---|---|---|---|---|
| 10 | 8 | 3 | 1 | 1 |
| 56 | 3 | 1 | 0 | 0 |
| 38 | 38 | 5 | 2 | 0 |
| 39 | 5 | 4 | 3 | 1 |
| Mean 35.8 | 13.5 | 4.2 | 1.5 | .5 |
| Median 38 | 6 | 4 | 2 | 0 |

C. Seven-factor experiment (1 experiment)

| | | | | | | |
|---|---|---|---|---|---|---|
| 6 | 4 | 3 | 1 | 1 | 0 | 0 |

*Each line represents a different experiment. Proportions are ordered from highest to lowest. Values listed are porportions multiplied by 100.

Differentiating between small and unimportant effects. The exponential distribution of the effects of variables provides some clues to how their individual magnitudes might be interpreted. Within any experiment in which a sufficiently large number of factors is studied, a relatively few will consistently account for most -- perhaps 70 or 80 percent -- of the performance variance for a particular operational task. The remaining proportion of variance will be accounted for by a great many sources of variance, none of which in the experimental situation account for very much of the variance -- possibly only a few percent. In understanding the effects of these remaining factors, the experimenter must distinguish those that are essentially unimportant though constantly present when the task is being performed and those that are important but occur infrequently. In an experiment in which the measures from many observations are being averaged, the effects of both types of variables may appear numerically equivalent. Quite obviously they are not. For any one performance, the important but infrequent factor could totally disrupt otherwise typical performance. Luckily this type of factor can often be rationally identified.

## How important are higher order effects in human factors experiments?

Considerable economy in data collection can be effected if higher order effects are negligible in human factors experiments. Simon (1973) reviewed a number of experimental designs that would enable a very large number of factors to be studied in the experiment quite economically provided it was not necessary to isolate third order and higher interaction effects from lower order effects. The principle in a simple one. If a main and a four-factor interaction effect are confounded, that is, if their individual effects cannot be isolated, and if the four-factor interaction effect is negligible, then the combined measure must

actually be that of the main effect. While the better designs test the validity of the assumption of negligible higher order effects (and if it is not valid, more data should be taken), an investigator will use these data-collection plans with more confidence if he knows in advance that the likelihood of higher order effects being important is slight.

An analysis of the 239 experiments was made to determine the proportions of variance accounted for by the equipment interaction effects. In Table 13, this data is first separated in column 1 into the order of the interactions (i.e., the number of interacting equipment factors), and in column 2 by the total number of factors in the experiment from which the data was taken. The number of interactions of each particular order in an experiment is shown in column 3, and the number of experiments available for analysis in this sample is shown in column 4. The data, so subdivided, could be examined in two ways. In the first way (columns 5 through 9), the sum of the proportions of variance accounted for by all interactions of the same order in an experiment (column 3) was the basic unit for the analysis; in these cases, the term "combined" was used. In the second way, the proportions of variance for the individual interactions were analyzed (columns 10 and 11).

From the data in Table 13, the following generalizations can be made:

- The more factors studied in a single experiment, the smaller the proportion of variance accounted for by individual interactions.
- The higher the order of interaction, the lower the proportion of variance accounted for by that order.
- Four-factor interactions and higher are for all practical purposes negligible.

TABLE 13. ANALYSES OF THE PROPORTION OF VARIANCE EXPLAINED BY EQUIPMENT-INTERACTION EFFECTS

| Order of Interaction Effect (Number of Factors in the Interaction) | Number of Factors Studied in the Experiment | Number of Interactions of each Order in an Experiment | Number of Experiments in this Category | Combined* Proportion of Total Variance Accounted for by All Interactions of the Same Order Within Each Experiment | | | Percent of Interactions (Combined & Individual) in which the Proportion of Variance Accounted for Exceeds the Indicated Proportion | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | | | | | Combined Interactions of Same Order (n=col. 4) | | Individual Interactions (n=cols. 3x4) | |
| | | | | 50%ile | 75%ile | 100%ile | >.05 | >.10 | >.05 | >.10 |
| Column: (1) | (2) | (3) | (4) | (5) | (6) | (7) | (8) | (9) | (10) | (11) |
| 5th (5 FI) | 5 | 1 | 4 | .00 | | .01 | 0 | 0 | 0 | 0 |
| 4th (4 FI) | 5 | 5 | 4 | .00 | | .01 | 0 | 0 | 0 | 0 |
| | 4 | 1 | 13 | .00 | .01 | .0[illegible] | 0 | 0 | 0 | 0 |
| 3rd (3 FI) | 5 | 10 | 4 | .02 | | .02 | 0 | 0 | 0 | 0 |
| | 4 | 4 | 13 | .03 | .03 | .11 | 13.2 | 7.6 | 1.9 | 0 |
| | 3 | 1 | 55 | .01 | .02 | .19 | 10.9 | 5.4 | 10.9 | 5.4 |
| 2nd (2 FI) | 5 | 10 | 4 | .08 | | .1[illegible] | 50.0 | 50.0 | 5.0 | 2.5 |
| | 4 | 6 | 13 | .09 | .10 | .24 | 76.9 | 7.6 | 9.0 | 0 |
| | 3 | 3 | 55 | .05 | .10 | .36 | 43.6 | 20.0 | 13.9 | 4.8 |
| | 2 | 1 | 93 | .02 | .04 | .65 | 24.0 | 11.3 | 24.0 | 11.3 |

*The term "combined" as used throughout this table refers to the sum of the proportions of variance accounted for by every interaction of the same order within each experiment.

- In over 75 percent of the experiments, three-factor interaction effects can be considered to be negligible. However, as the number of variables studied in an experiment decreased, some three-way interaction effects were large enough to require further examination.

Three-factor interactions. From Table 14, it can be seen that when five factors were studied in an experiment, the three-factor interaction effects were negligible. However, this is based on the results from only four experiments. Three-factor interaction effects also appear to be negligible for all practical purposes in the four-factor studies. The maximum combined value of four interactions accounted for only .11 of total variance. Of the four interactions that were summed to make that amount, only one accounted for more than .05 of total variance -- it accounted for .06.

All of the experiments in which the combined three-factor interactions accounted for more than .05 of the total variance are listed along with some descriptive data in Table 14. This was the case in only eight of the 72 experiments which could be analyzed for three-factor interactions; two were the combined value of four effects. Only four of the eight accounted for more than ten percent of the total variance. Two (No. 4 and No. 8) were the combined value of four individual three-way interaction effects of which only one of the six individual interactions accounted for .06 of the total variance. Two (No. 2 and No. 3), although accounting for .18 and .16 of the total variance in each experiment, were used in lieu of an error term. That means that the experimenter treated these effects as if they were due to pure chance, i.e., negligible. One case (No. 7) was not reliable, i.e., statistically significant. The factors making up this group of three-factor interactions were primarily qualitative variables; there was only one exception (No. 4). Only one (No. 1) of these three-factor interactions (among those for which it could be

TABLE 14. ANALYSES OF THREE-FACTOR INTERACTION EFFECTS ACCOUNTING FOR MORE THAN .05 OF THE TOTAL VARIANCE

| | Number of Factors in Experiment | Proportion of Total Variance Accounted for by Combined 3FIs | Number of Interaction Effects Summed | Proportion Accounted for by Individual 3FIs | Number of Levels | Type of Variable* | Type of Interaction |
|---|---|---|---|---|---|---|---|
| 1 | 3 | .19 | 1 | .19 | 2,2,2 | LLL | Disordinal |
| 2 | 3 | .18 | 1 | .18 | 3,2,30 | LLL | ** |
| 3 | 3 | .16 | 1 | .16 | 3,4,2 | LLL | ** |
| 4 | 3 | .11 | 4 | .06,.04,.00,.01 | 3,2,2 | NNN | Ordinal |
| 5 | 3 | .10 | 1 | .10 | 2,3,2 | LLL | *** |
| 6 | 3 | .09 | 1 | .09 | 3,3,5 | LLN | Ordinal |
| 7 | 3 | .08 | 1 | .08 | 3,3,5 | LLN | Not Signif. |
| 8 | 4 | .06 | 4 | .04,.01,.01,.00 | 20,3,2 | LNN | Ordinal |

*L=qualitative; N=quantitative; LLN=2 qualitative, 1 quantitative; LNN=qualitative, 2 quantitative.

**Used as error term.

***Insufficient data to decide.

determined) was of the disordinal type. A disordinal interaction is one in which the performance at different levels of a factor will be ordered differently depending on the level of a second factor which is operating when the performance is measured. The others were the ordinal type of interaction, which could probably have been eliminated had a different measurement scale been used, or if the performance scores had been appropriately transformed. It is of interest to note that in the worst case, that is the case in which the three-factor interaction accounted for .19 of the total variance, the absolute difference between the worst and the best of the eight experimental conditions in that experiment was 1.44 bits/second of transmitted information from display to control. In reaction time alone, the difference amounted to .78 parts of a second. This probably were of no practical significance.

It is apparent that a tentative assumption that three-factor interactions are negligible is the most parsimonious one to make. In very few cases it may be wrong. However if the measurement scales are selected from the beginning to linearize the data as much as possible, the number of critical three-factor interactions will be reduced. Non-negligible effects are more likely with qualitative factors.

Cochran and Cox (1957, p. 219) suggest watching the two-factor interactions for clues that three-factor interactions might be important. They suggest that if the main effects and two-factor interactions of a set of factors are large, it is likely that some three-factor interactions might also be large, If the two-factor interactions are small, it is less likely (but not impossible) that the three-factor interactions are large.

Two-factor interactions. While most economical multifactor designs are constructed so as not to ignore two-factor interactions, it still is of interest to obtain quantitative information on how

important these effects are likely to be. From the data in Table 13, the following generalizations can be made about the two-factor interaction effects:

- The more factors studied in an experiment, the more likely it is that an individual two-factor interaction will be negligible.
- If all of the data from experiments with three or more factors were combined, only 36 out of 72 experiments had the combined effects of the two-factor interactions in the studies accounting for more than .05 of the total variance. Only 11.3 percent of the individual two-factor interactions in the studies involving three or more factors accounted for more than .05 of the total variance. Only 3.2 percent of the individual two-factor interactions in the studies involving three or more factors accounted for more than .10 of the total variance.
- Two-factor interactions, in general, cannot *a priori* be assumed negligible.

In general, interaction effects tended to be somewhat higher when qualitative factors were involved than quantitative.

Higher order terms of the polynomial. The functions relating quantitative factors to performance can be approximated by a graduated polynomial. Each term of the polynomial will represent a single degree of freedom. Thus the main effect of a three-level factor with two degrees of freedom in the analysis of variance, will be represented by two terms in the equation -- a linear and a quadratic term. The interaction of two three-level variables with four degrees of freedom in the analysis of variance would be represented by the following four terms, each with a single degree of freedom, in the polynomial:

| | | |
|---|---|---|
| $x_i x_j$ | (linear-by-linear interaction) | 2nd degree term |
| $x_i^2 x_j$ | (quadratic-by-linear interaction) | 3rd degree term |
| $x_i x_j^2$ | (linear-by-quadratic interaction) | 3rd degree term |
| $x_i^2 x_j^2$ | (quadratic-by-quadratic interaction) | 4th degree term |

The degree of the term is equal to the sum of the exponents in the term; the order of the equation is equal to the highest degree of any term in the equation. The majority of economical multifactor designs that can be used with quantitative factors, e.g., central composite designs (Simon, 1970), limit the data collection to that required for a first or second degree model. In the above example of the two-factor interaction, this would mean that only the linear-by-linear component of the interaction would be estimated, and the other three components would be assumed negligible.

Similarly, if a factor contained five experimental levels, its relation to performance could be represented by four terms:

$$x_i,\ x_i^2,\ x_i^3,\ x_i^4$$

of which the cubic and quartic terms would be assumed negligible. The question is: How likely is it that these higher order effects are really negligible?

Because the analysis of variance model dominated the analyses of the experiments published in the journal, Human Factors, between 1958 and 1972, there was little data available for checking this assumption. Only nine of the 118 papers included regression components, i.e., linear, quadratic, or cubic terms, in their ANOVA tables. However, whenever the means of every level of a quantitative main effect were published, it was possible to determine how well equations containing from first- to fifth-order terms would fit these main effects. An analysis was performed on the main effects of all quantitative variables with three, four, five, or six levels that had accounted for .25 or more of the total performance variance in the experiment. The results are shown in Table 15.

TABLE 15. PROPORTION OF VARIANCES OF MAIN EFFECTS ACCOUNTED FOR AS A FUNCTION OF THE ORDER OF THE POLYNOMIAL

| Number of Levels Involved | Order of the Polynomial | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| | 1st | | 2nd | | 3rd | | 4th | 5th |
| | Median | Range | Median | Range | Median | Range | Median | Median |
| 3 (20)* | .96 | .71-1. | 1.0 | - | | | | |
| 4 (10) | .76 | .55-1. | .98 | .92-1. | 1.0 | - | | |
| 5 (4) | .97 | .80-1. | .99 | .95-1. | 1.0 | .99-1. | 1.0 | |
| 6 (2) | .60 | - | .98 | - | 1.0 | - | 1.0 | 1.0 |

*Numbers in parentheses indicate the number of main effects included in the analysis. Only main effects that accounted for .25 or more of the total variance were included.

Table 15 shows the proportion of the variance of quantitative main effects that is accounted for when represented by polynomials of different orders. Obviously an equation of order (d - 1) will account for all of the variance of any main effect with d levels. For each group of data, the median, and range from lowest to highest proportions accounted for, are given. One can conclude from the data in this table that for the sample involved, the inclusion of higher than second-order terms in the polynomial will account for a negligible proportion of the main effects.

What proportion of the "small effect" factors were statistically significant?

The statistical significance of a factor has too often been a major criterion for eliciting concern for its effect on performance. This procedure, however, has been subjected to criticism since, quite often, the magnitude of an effect on the performance has been found to be trivial. Since statistical significance only implies the probability that an effect might be reliable (and being a probability, might be unreliable), the importance of examining the strength of the effect of a factor is emphasized. This was discussed earlier in this paper.

An analysis was made of all main and interactions effects (up to and including four-factor interactions) for all experiments studying from one to five factors to determine what percent of the effects accounted for .01 or less of the variance in the experiment and what percent of these were statistically significant.

In Table 16-A, the percentage of effects in each group accounting for one percent or less of the variance is shown. The total number of effects in each group on which each percentage is based is shown in parentheses. In general, the percentage of effects accounting for one percent or less of the experimental variance increases as the number of equipment factors in the experiment increases. A similar increase appears in the percentage of one-percent effects as the order of the effect increases; however, interactions appear to increase at a more rapid rate than the main effect in this regard. This "increase" means that more effects are becoming more negligible.

In Table 16-B, the percentage of only the effects accounting for one percent of the variance or less in each group that was statistically significant is shown. In this case, as the number

TABLE 16-A. PERCENTAGE OF ALL EFFECTS ACCOUNTING FOR ONE PERCENT OR LESS OF THE TOTAL VARIANCE

| Number of equipment factors per experiment | Sources of variance: Main | | 2FI* | | 3FI | | 4FI | |
|---|---|---|---|---|---|---|---|---|
| 1 | 8 | (71)** | | | | | | |
| 2 | 24 | (186) | 39 | (93) | | | | |
| 3 | 30 | (165) | 54 | (165) | 47 | (55) | | |
| 4 | 23 | (52) | 60 | (78) | 65 | (52) | 69 | (13) |
| 5 | 35 | (20) | 92 | (40) | 98 | (40) | 100 | (20) |

TABLE 16-B. PERCENTAGE OF EFFECTS IN TABLE 16-A THAT WERE STATISTICALLY SIGNIFICANT

| Number of equipment factors per experiment | Sources of variance: Main | | 2FI | | 3FI | | 4FI | |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | (6)** | | | | | | |
| 2 | 13 | (45) | 8 | (36) | | | | |
| 3 | 12 | (49) | 17 | (90) | 12 | (26) | | |
| 4 | 25 | (12) | 29 | (47) | 26 | (34) | 22 | (9) |
| 5 | 71 | (7) | 24 | (37) | 15 | (39) | 10 | (20) |

TABLE 16-C. PERCENTAGE OF ALL EFFECTS THAT WERE STATISTICALLY SIGNIFICANT BUT ACCOUNTED FOR ONE PERCENT OR LESS OF THE TOTAL VARIANCE [TABLE 16-A x TABLE 16-B]

| Number of equipment factors per experiment | Sources of variance: Main | | 2FI | | 3FI | | 4FI | |
|---|---|---|---|---|---|---|---|---|
| 1 | 0 | (71)** | | | | | | |
| 2 | 3 | (186) | 3 | (93) | | | | |
| 3 | 4 | (165) | 9 | (165) | 6 | (55) | | |
| 4 | 6 | (52) | 18 | (78) | 17 | (52) | 15 | (13) |
| 5 | 25 | (20) | 22 | (40) | 15 | (40) | 10 | (20) |

*FI stands for "factor interaction".
**Numbers in parentheses indicate total number of effects in each category.

of experimental factors in an experiment increases, the more likely a one-percent or less effect is statistically significant. This increase in reliability could be accounted for by the larger number of degrees of freedom generally found in the error terms of the larger experiments. On the other hand, the higher the order of the effect, the less chance that a one-percent or smaller effect will be significant.

Finally, in Table 16-C, (which is the product of Tables 16-A and 16-B) the percentage of all effects in all of the experiments that account for one-percent or less of the variance and are statistically significant is shown. For all effects combined, whether main or interaction, 18 percent are of this type.* As the number of equipment factors in an experiment increases so does the probability that one percent or less effects will be statistically significant. Interactions accounting for one percent of the variance or less also have a higher chance of being statistically significant than main effects accounting for one percent or less of the total variance. With the same error term, interactions have the edge on "significance" over main effects since they generally provide more degrees of freedom in the numerator of the F-ratio.

*In this same sample, 28 percent of the effects accounting for four percent or less of the variance are statistically significant.

## SECTION IV. RESEARCH APPLICATION SURVEY

Two hundred thirty nine experiments reported in 118 papers published over a 14-year period have been described and evaluated. On the whole, in these experiments, much more data was collected than was needed to supply the limited amount of information that was obtained. Good, bad, or indifferent, however, this work is characteristic of that being produced by human factors engineers in universities, industry, private institutions, and government (including the military) laboratories. Whatever their intermediate goals, directly or indirectly these investigators performed their experiments to obtain information that would be used to optimize equipment, systems, and environments and ultimately improve man-machine performance.

A lot of time, a lot of effort, and a lot of money have gone into this research. Just how effective has it been? To what extent have the results of these experiments actually improved the performance of an operational man-machine system? To obtain answers and information on these questions, the investigators were surveyed.

### SURVEY RESPONDENTS

The questionnaire shown in Table 17 was sent to as many of the authors of the 118 papers for whom it was possible to find addresses. However, no addresses for the authors of 15 papers could be found. The questionnaire was sent to at least one author of the remaining 103 papers. Completed questionnaires were returned by 114 authors of 94 of the papers. The authors-to-papers distribution among the respondents was: 76 papers, one author each; 16 papers, two authors each; two papers, three authors each. Thus, 91 percent of those queried responded. Of all the papers that were analyzed, 84 percent were represented in this applications survey; 19 percent of them were represented more than once.

TABLE 17

QUESTIONNAIRE ON THE APPLICATION OF DATA FROM HUMAN FACTORS ENGINEERING EXPERIMENTS

Please answer the following questions concerning your experiment published in the journal, Human Factors, entitled:

(94 studies)

Check the appropriate answer(s) in the space provided. If you would care to comment about any item, use the back of this sheet or use additional pages.

Your prompt reply would be greatly appreciated. Please return this questionnaire before 15 November 1973 to Charles W. Simon, Bldg. 6, MS D-120, Hughes Aircraft Company, Aerospace Group, Culver City, California 90230. Thank You.

| | | % |
|---|---|---|
| 1. Who originally requested (by posing the question to be answered) that this experiment be conducted? N = 82 AQ = 66% | a. The experimenter himself | a(76)7 |
| | b. Someone within the organization that employed the experimenter | b(17)8 |
| | c. A customer outside of the experimenter's organization | c( 7)9 |
| 2. What kind of answers was the experiment intended to supply? N = 90 AQ = 22% | a. Answers relevant to the design of a specific system (at least) | a(42)10 |
| | b. Answers having only general applicability | b(58)11 |
| 3. Did the results of the experiment directly influence the design of a real system? N = 87 AQ = 39% | a. Yes (Which? _____) | a(31)12 |
| | b. *No or don't know* | b(69)13 |
| 4. If there were any measurable benefits to a real system, what were they? (More than one may be checked.) N = 78 AQ = 61% | a. No known measurable benefits | a(51)14 |
| | b. Reduced cost of building or operating system | b( 4)15 |
| | c. Resulted in saving life and/or property | c( 3)16 |
| | d. Enhanced system performance (time, error, etc.) | d(43)17 |
| 5. What do you estimate the total cost of this experiment to be? N = 89 AQ = 22% | a. $ 30,000 or less | a(87)18 |
| | b. Between $ 30,000 and $ 60,000 | b(12)19 |
| | c. Between $ 60,000 and $ 90,000 | c( 1)20 |
| | d. $ 90,000 or more | d( 0)21 |
| 6. Were any experiments performed specifically as a follow-up to this one? N = 86 AQ = 44% | a. Yes | a(37)22 |
| | b. No | b(63)23 |
| If "yes", why? (More than one may be checked.) N = 30 of the 32 answering "yes" | a. To clarify questions arising from the first study | a(18)24 |
| | b. To add to the information obtained in the first study | b(72)25 |
| If more information was obtained in a follow-up study, how was this done? N = 28 of the 29 answering "yes" and also "to add information" | a. Repeated original study but changed experimental space | a( 5)26 |
| | b. Examined some old and some new variables | b(82)27 |
| | c. Examined only new variables | c(13)28 |
| 7. In retrospect, would you have done your experiment differently were you to repeat it today? If "yes", briefly indicate why and how. N = 89 AQ = 27% | a. Yes | a(22)29 |
| | b. No | b(78)30 |

## RESULTS

Responses to the questionnaire were analyzed in terms of the number of papers (studies) rather than the number of authors. Consequently, in summarizing the answers to each question in this report, two groups of data are available for each question: one, the data based on those papers in which every author (be it one or more) agreed on the answer to the question; two, the data based on those papers in which all authors of a paper did not agree on the answer to the question. Since the basis for tabulation is the paper, not authors, and since only 18 of the papers had multiple authors, at most only 18 disagreements were possible.

Some of the results of the survey are summarized on the questionnaire in Table 17. For each question, values for N and AQ are given. N equals the number of papers in which there was internal agreement among the authors of each paper for the answer of that question. AQ represents an ambiguity quotient for each question, or the percentage of the 18 multiple-author papers wherein the responding authors of the same paper failed to agree on the answer. The higher the percentage, the more ambiguous the question is considered to be. The values in parentheses at the right of each answer are the percentages of the N papers that answered the question with that particular answer. When the questions were ordered from least to highest AQ percentage, the least to most ambiguous questions were:

2 and 5; 7; 3; 6a; 4; 1 and 6b; 6c.

## DISCUSSION

Comments made by the respondents to qualify their answers provided clues in some cases as to why multiple authors did not agree. These comments were consolidated and are reported for each question. Conclusions are drawn from this data and interpreted and discussed.

No attempt was made to use inferential statistics on the data. It would be dangerous to overinterpret the information from this questionnaire, i.e., to seek information to a depth or precision that isn't there. The questionnaire represents a quick and inexpensive means of finding out what the investigators knew and thought about their own studies. Their responses, therefore, were examined primarily for gross generalities, the discovery of which made statistical treatments unwarranted.

Question 1. Who originally requested (by posing the question to be answered) that this experiment be conducted?

Number of papers in which all authors agreed on the answer . . . . . 82

76% indicated that "the experimenter himself" posed the question.
17% indicated that the question was posed by "someone within the organization that employed the experimenter."
7% indicated that the question was posed by "a customer outside the experimenter's organization."

Number of papers in which authors disagreed on the answer . . . . . 12

(66% ambiguity, based on 18 possible articles with multiple authors.)

Some respondents qualified their answers with comments, reflecting possible reasons why disagreements occurred. These can be summarized into the following types:

- Junior authors interpreted the word "experimenter" to mean themselves alone and therefore indicated that someone other than the "experimenter" (themselves) originated the problem, when in fact the senior author did.

- Junior authors really didn't know who originated the problem.
- On at least three studies, there was collaboration between the experimenter and the customer in posing the question.
- "Posing a question" was interpreted by some to mean defining the specific variables to be investigated rather than raising the general question to be answered by the experiment.
- It was difficult to decide what constituted the "experimenter's organization," e.g., a department, division, or company.

There is little question that the majority of the investigators believed that they originally proposed the studies included in this survey. The extent to which these proposals may have been tempered by outside interests and requests for proposals cannot be judged. But since a great many of the studies were supported outside the agency that performed the experiments, from the investigator's viewpoint it seems he at least believed that he proposed, and the contracting agency disposed.

Question 2. What kind of answers was the experiment intended to supply?

Number of papers in which all authors agreed on the answer . . . . . 90

42% indicated that "answers were to be relevant to the design of a specific system (at least)."
58% indicated that they sought "answers having only general applicability."

Number of papers in which authors disagreed on the answer . . . . . 4

(22% ambiguity, based on 18 possible.)

The comments made to this question suggest that some disagreements among multiple authors may have arisen for the following reason:

- When experimental results were relevant to a specific class of systems, such as air defense systems, intelligence systems, automobiles in general, production lines, but not to a specific system, some respondents considered this as being relevant to a specific system (while others did not).

The answers were split approximately in half between those who believed their experiments were planned to answer questions relevant to specific systems, at least, particularly if "systems" could be interpreted as "systems of a certain class," and those that were planned to supply answers with general applicability. From the tenor of many comments, those who did studies to obtain answers of general applicability believed the information that they generated would be relevant to specific systems if anyone cared to use it. The real difference affecting the selection of one answer or the other appeared to be whether or not a particular application for the data had been anticipated while the study was being planned.

This question and the alternative answers generated some relatively heated comments. The phrase, "having only general applicability," and particularly the word "only" in the second answer was interpreted as being judgmental. For example, respondents made such comments as: "Is it bad for an experiment to have only general applicability?" and "'Only' is a biasing word." Conversely, in spite of the presence of the words "at least" qualifying the first answer (that the results were applicable to specific systems), several respondents wanted to check both answers to show that these results had general applicability as well.

Actually, as the alternatives were written, answering that results were applicable to "a specific system (at least)" did not deny their general applicability, while saying that they had "only general applicability" did not evaluate the desirability of the research, but merely indicated that the study had been planned without a particular system in mind.

That this question and its alternative answers produced the kinds of comments they did, attributing motives to the wording of the questionnaire that were not there, neither in fact nor in intention, suggested that for some human factors investigators this is a sensitive issue. The comments of several respondents who took the time to express their ideas on this matter in some detail are quoted below and commented upon. One respondent wrote:

> I think you miss the point in your analysis of the experiments. Experiments are almost always science not engineering. Rarely is there time during design to do experiments. If experiments are done, they usually are too specific to be "publishable". Research from a university is rarely "applied science" much less "design" or "development". Only in industry would you find research "to improve the operation of a man machine system." If you are interested in finding research devoted "to improve the operation of a man machine system" you should look in an engineering oriented journal such as Applied Ergonomics, or better yet, company and govt. technical reports.

The comment seems to suggest that human factors experiments are conducted in the university, but that they are not "applied science" nor intended to improve the operation of man-machine systems, and conversely, that experiments for this purpose, if conducted in the industrial environment, are too specific (due to time and other constraints) to be publishable. Of course, neither statement is true.

There are any number of universities today doing research which is intended to aid in the design of man-machine systems. In fact, relevancy in this regard is almost a necessity if one hopes to get governmental support. It is when the academicians or any experimenter forgets that while it is his inalienable right (as long as he is supporting his own work) and a pleasant thing to do his experiment any way he wishes, our status as a profession depends ultimately on being able to show that we can collect information that will help someone make a decision or solve a problem. Human behavior is "real world" and, except in some rare circumstances, should be studied involving tasks with the complexity of the real-world counterparts. No elaborate experimental design or exotic statistical analysis will suffice to justify the existence of any research program -- basic or applied -- unless it produces information previously unknown and ultimately useful. The oversimplified tasks, so often used in what is erroneously labeled "basic" research, have not and will not produce this kind of data. University research need not be criticized because it has no immediate application, but only when it may never have an application. As such, it is not basic, which means that it could be used to answer many problems, but is irrelevant to any real world problem.

That relatively few of the human factors experiments performed by engineering organizations today have enough generality to be published is probably a true assessment of the situation. A second respondent made a similar point about Armed Forces research which "is typically so designed that it only applies to the system for which it was done -- and hence never gets published in any general literature if at all." But this criticism of industrial research was not always so. Not too many years ago, industry supported "basic" research in

human factors engineering on a considerably greater scale than it does today, but when this research failed to produce the kind of information needed by the organizations paying for the work, the support gradually eroded to almost nothing. As one respondent wrote:

> Although I have done a fair amount of publishing, nothing has received the attention within the engineering community as this extremely simple study designed to provide an "engineering answer to an engineering question." In general, I would say that as a first condition a research study designed to serve the engineering community must be "definitive" in the sense that all loose ends must be chased down prior to rushing into print; secondly, the discussion must provide an in-depth treatment of the total problem designed specifically to enable the reader to initiate his own creative activities.

The criticism to be leveled at research performed in industry (often in support of military programs) should not be that it seeks to solve specific problems, but that quite often it solves no problems at all insofar as it fails to supply valid and previously unknown information. Too often these studies involving specific real world systems fail to investigate performance in a broad enough context to match that which is likely to occur under operational conditions. As a result, the data produced is no more likely to be correct than in the case of an experiment conducted in the university where neither the task nor the context is relevant.

Question 3. Did the results of the experiment directly influence the design of a real system?

Number of papers in which all authors agreed on this answer . . . . . 87

31% said "yes."

69% said "no" or "don't know." Ten percent of the 69% specifically indicated that their answer was "don't know."

Number of papers in which authors disagreed on the answer . . . . . 7

(22% ambiguity, based on 18 possible.)

Some respondents qualified their answers with the following kinds of comments. Some of these may explain the disagreements in the answers given on papers with multiple authorship:

- Junior investigators may not have been apprised of where the results were applied.
- Some investigators stated that although application had been the original intention, they did not know if the results were ever used.
- Some stated that the study was done in support of a real system but that they were never applied because the system had not yet been developed, that the project had been canceled, or that the study had been done in support of a proposal that did not result in a contract.
- In one case, while the results favoring a particular system was definitive, the system was not changed accordingly because to do so would have created "administrative problems."
- Some investigators, answering either question, stated that they didn't really know if their results were ever applied but that systems had appeared subsequently that were operated differently from the methods proposed by the experimental results.

Though studies were performed with the intention of applying the results to real man-machine systems, for a variety of reasons the results of only a minority of the studies were actually known to be applied. This suggests,

at worst, that in spite of the time, effort, and money involved in producing a study, its results are not used. At best, it suggests that if results are used, this fact is not often fed back to the investigators. In this regard, the survey is limited in that it provides no information on the number of persons who may seek (and find) answers to their design problems from the results of the experiments published in the Human Factors journal. Investigators are not always aware of this although one respondent wrote that the request for reprints of his paper suggested others may have used the data. On the other hand, another respondent wrote:

> In answer to your letter, the experiment was fun and we learned a few things. The customer's interest flagged and funds disappeared. You are the first one to indicate that he has as much as read the title of the article, and I doubt if anyone in the governmental organization that sponsored the study even knows about our report.

Certainly one aspect of seeing that research results are applied is in getting the information to those who need it. Merely publishing the results in a company report or in a professional journal is no guarantee that the person who could or would use it will ever see it.

Question 4. If there were any measurable benefits to a real system, what were they? (More than one answer may be checked.)

| | |
|---|---|
| Number of papers in which all authors agreed on the answer . . . . . | 78* |

*The author(s) of five others gave no answers at all.

51% stated there were "no known measurable benefits."
4% stated that they "reduced cost of building or operating system."
3% stated that they "resulted in saving life and/or property."
43% stated that they "enhanced system performance."

Number of papers in which authors disagreed on the answer . . . . . 11

(61% ambiguity, based on 18 possible.)

Multiple answers were permitted to this question. However, in ten of the 11 papers in which multiple authors did not agree on which answer was correct, answers were divided between the first alternative that stated there were no known measurable benefits or one of the other alternatives that indicated there was a benefit.

Some respondents qualified their answers with the following kinds of comments, which may help explain some of the disagreements in answers given on papers with multiple authors:

- Benefits could not be measured since experimental results were applied to systems not yet operational.
- Results enabled an "estimation" to be made of an improvement but no test of a real system was possible.
- Although the solution was found to enhance system performance in the experiment, it was never used since its implementation would have actually increased system cost.

While the results from more studies were believed to have enhanced system performance rather than to have reduced costs or to help save lives, these benefits are not necessarily independent nor antagonistic. Equipment designed to make performance less error-prone can save lives. Good research should find ways of redesigning a system to enhance performance at no increase in cost.

There are indications that in spite of the way the question was worded, some respondents were only indicating that results would be beneficial if used. Respondents would refer to the "potential" benefits of their results, or that the results "would have helped" but the author didn't know if they did, or that the system "would be" improved if the results were used. In one case, the comment was made that system benefits did not derive directly from the results of the experiment but from the method developed for the study, which was then applied to the study of other problems.

There was a fairly even split between the papers that were not known to have provided measurable benefits and those that investigators believed did result in measurable benefits. However, from the answers to other questions and the comments made on this one, there is reason to suspect that many of the "measurable benefits" may never have been made on operational systems but were assumed, inferred, or extrapolated from the experimental results. Furthermore, since operational systems are generally built only one way, in many cases no comparison data would be possible to arrive at an absolute measure of benefit for the single system nor absolute trade-off values between proposed solutions and costs.

Question 5. What do you estimate the total cost of this experiment to be?

Number of papers in which all authors agreed on the answer . . . . . 89*

87% indicated the cost to be "$30,000 or less."
12% indicated the cost to be "between $30,000 and $60,000."
1% indicated the cost to be "between $60,000 and $90,000."

Number of papers in which authors disagreed on the answer . . . . . 4

(22% ambiguity, based on 18 possible.)

In the cases where multiple authors disagreed on the cost of the study, the differences were between adjoining alternatives.

On the whole, the experiments published in Human Factors that were included in this study cost less than $30,000. Because the questionnaire was not sufficiently sensitive at the lower end of the scale, we can't know how much less it may have cost; a number of respondents who checked that answer also indicated that the amount was one, two, or three thousand dollars. One estimated the cost to be $100," but one must assume that this didn't include an investigator's salary, at least. On the other hand, the one study that was estimated to cost between $60,000 and $90,000 was actually a thesis being done by a student who included in his estimate what the time of 1,000 graduate students used as subjects might cost. Obviously, the basis on which these costs were estimated varied considerably for different persons.

---

*The author(s) of one other gave no answer.

While $30,000 is a lot of money from one point of view, in today's marketplace, particularly outside the educational institution, it does not buy a great deal after overhead and administrative costs have been removed. If the amount includes not only the salaries of those who conduct or participate in the experiment, but also the cost of equipment, then on the whole the experiments included in this survey were not particularly expensive.

However, mere dollar value is not sufficient to judge whether a study is expensive or not. The dollar costs must be traded against the degree to which the experiments provided the necessary answers, and whether those answers or more complete answers could be obtained for less money. Investigators seldom use these latter criteria to evaluate their research, yet either directly or indirectly the customer does. The earlier sections of this report have already shown that there is a tendency to collect too much data and to get too little information for the effort. Experiments, to be of value, must supply answers to questions that have been asked, or will be asked, and to do so as inexpensively as possible. Statistical criteria must be traded against pragmatic and economic criteria when evaluating research effectiveness.

Question 6-a. Were any experiments performed specifically as a follow-up to this one?

Number of papers in which all authors agreed on the answer . . . . . 86

37% said "yes."
63% said "no."

Number of papers in which authors disagreed on the answer . . . . . 8

(44% ambiguity, based on 18 possible.)

Some respondents, explaining why there were no follow-ups to their experiment, said the published article included the entire series. Some said it was at the end of a series. No comments explained the few disagreements among authors, although comments to earlier questions suggested that junior authors were not always aware of what was happening outside the conduct of study in which they participated.

Question 6-b. If follow-up experiments were performed, why?

Number of papers (of the 32 answering "yes" to 6-a) in which authors agreed . . . . . 30

7% said "to clarify questions arising from the first study."
70% said "to add to the information obtained in the first study."
23% marked both answers. (Multiple answers were allowed).

Number of papers in which authors disagreed on the answer . . . . . 2

Multiple answers were allowed, so the disagreements between two groups of authors were incidental. In each case, one author indicated both answers were true and the other indicated only one answer was true.

Question 6-c. If more information was obtained in the follow-up, how was this done?

Number of papers (out of 29 who answered "yes" on 6-a and also "add information" on 6-b) in which all authors agreed on answer . . . . . 28

4% said they "repeated original study but changed the experimental space."
85% said they "examined some old and some new variables."
11% said they "examined only new variables."

Number of papers in which authors disagreed on the answer . . . . . 1

In the one case in which the authors disagreed, one said the follow-up study investigated only new variables while others said it had investigated both old and new variables.

There were no follow-up studies for the majority of the experiments. Except for the few cases in which the experiment was the end of a series (or when the entire series was published in the paper), no comments for this were given. When follow-up studies were done, they were primarily to add to the information obtained in the first by repeating some old and adding some new variables.

Two observations seem warranted: First, in view of the limited size and degree of inconclusiveness (of most of these studies as described in earlier sections of this report), it is difficult to imagine that an extension of the work might not have proved informative.

The "one-shot" experiment has become a sign of the times -- a school project, a PhD. dissertation, a government contract, a one-experiment publication -- generally to be finished in an academic semester or a fiscal year without regard for the scope of the problem or the requirements for answering a question. Second, those that did the extra studies might have considered a more effective and efficient approach from the beginning. Since most of the studies examined only three or fewer factors in their first study, applications of economical multifactor designs (Simon, 1973) can save time and money in the first place. These designs enable the experimenter to first determine the most important factors out of 15 or more and <u>then</u> study in depth the most important six or seven.

Question 7. In retrospect, would you have done your experiment differently were you to repeat it today, and if so, why and how?

Number of papers in which all authors agreed on the answer . . . . . 89

22% said "yes."
78% said "no."

Number of papers in which authors disagreed on the answer . . . . . 5

(27% ambiguity, based on 18 possible.)

The kinds of changes mentioned by those who would have done their experiments differently (in retrospect) fell into the following categories:

Predictor variables:

- Would expand the range of each variable
- Would introduce new variables
- Would better define their variables (if given the time to do pre-tests)

Dependent variable:

- Would add additional tasks

Experimental design:

- Would improve experimental design
- Would use more economical designs (such as fractional factorials) to reduce the data collection

Subject sample:

- Would use a larger sample
- Would use smaller groups but more replications
- Would draw from a different subject population (e.g., use non-students)

AD-A038 184 CANYON RESEARCH GROUP INC WESTLAKE VILLAGE CALIF F/G 5/5
ANALYSIS OF HUMAN FACTORS ENGINEERING EXPERIMENTS: CHARACTERIST--ETC(U)
AUG 76 C W SIMON F44620-76-C-0008
UNCLASSIFIED CWS-02-76 AFOSR-TR-77-0333 NL

2 OF 2
AD
A038184

END
DATE FILMED
4-77

Equipment:

- Would use better, more modern, and more sophisticated equipment (e.g., computers, better light source, etc.)

Administration:

- Would have closer cooperation with those interested in the experimental problem

The majority of experimenters said they were satisfied with the way they had done this research. A few however qualified this by saying that that was so if they had to repeat it "under the same circumstances." Another expressed satisfaction with what they had done "as far as it went," since it was done in support of a proposal. One respondent stated: "One always would perform an experiment more effectively once experience has been obtained," but it would seem that the greater number of experimenters in this sample did not necessarily agree. In fact, if one examines the quality of information that has been produced over the past 14 years by these studies, there is little indication that we profited much from that experience, since there has been little change in the methods of collecting such information over this extensive period.

The question had been posed to test the investigator's evaluation of his own methods. One respondent, however, suggested that in retrospect there were things other than method that might have been different. In answer to whether or not he'd do the study differently, the investigator wrote: "Yes and no. The experiment was OK for the problems stated, but was too expensive [Note: It was less than $30,000] for the results achieved. The same problem could be replicated with a good . . . simulator at less cost -- and with probably the

same or worse indicative (sic.) results. What is needed is better problem definition, more pilot studies and then some good, controlled experiments, but what is probably even more needed is recognition of, and experiment with, more significant problems."

# V. SUMMARY AND CONCLUSIONS

Two hundred thirty-nine experiments published in Human Factors during the period from 1958 to 1972 were analyzed for the purpose of discovering the characteristics of their experimental plans, the character and quality of their results, and the degree to which these results had been applied to real systems.

## SUMMARY

The following summarizes the major findings regarding this particular sample.

Regarding characteristics of design and methodology, the typical experiment:

- Used an ANOVA model for both its design and analysis.
- Used some form of a factorial design, most commonly with each subject being tested on every experimental condition (49%) or different groups on different sets of conditions (28%).
- Investigated the effects of two, and seldom (<10%) more than three, equipment factors.
- Infrequently (20% of the experiments) systematically studied a subject or a temporal factor along with equipment factors.
- Examined three levels of each factor, on average.
- Used nine, presumably homogeneous, subjects to replicate the entire basic design.
- Infrequently (25% of the experiments) tested the same subjects on the same experimental conditions more than once merely to replicate the basic design.
- Made, on average, approximately 26 times more observations per experiment than were needed to approximate a second-order space.

- Showed concern for sequence effects by testing experimental conditions in a systematic or random order, but generally failed to remove these effects in the analysis.
- Showed little agreement in the selection of the sources of variance to be used as the "error" term for significance testing.

The ranges of some of these characteristics were quite large at times, with some experiments being quite extreme in some cases. Numbers relating to the size of the experiment in many cases were correlated with the number of equipment factors being studied, which ranged from zero to seven in this sample.

Regarding the character and quality of the experimental data from this sample, the following results were obtained:

- Equipment factors accounted for only .31 (median) proportion of the total performance variance (increasing from .16 to .65 as the number of equipment factors per experiment increased from 1 to 5). The proportion of total variance accounted for by equipment factors ranged from practically none (less than .01) to practically all (up to .99).
- The variability of presumably homogeneous subjects, introduced only to replicate the experiment, accounted for more variance than the equipment factors in the one and two factor experiments, and considerably more of the variance than subject characteristics that the investigators had introduced as experimental factors.
- Between a third and a fifth of the variance, on average, in these experiments could be attributed to no interpretable source.
- The magnitude of main effects tend to distribute themselves exponentially, supporting the "principle of maldistribution."
- In this sample, three-factor interactions seldom showed more than a negligible effect and all higher-order interactions showed only negligible effects.

- Main effect functions could ordinarily be approximated by first or second degree polynomials.
- Nearly one-fourth of the main effects accounting for less than one percent of the total variance were statistically significant. More than forty percent of the main effects accounting for less than four percent of the total variance were statistically significant.

Results from a questionnaire sent to the investigators of these experiments revealed that:

- A majority of experimenters initiated their own experiments.
- The papers divided 58-42 as to whether they had been done to find answers of general applicability or answers relevant to a specific system (at least).
- Only 31 percent of the experiments were believed or known to have influenced the design of a real system. Most of the anticipated benefits were in enhancing system performance rather than in saving lives or dollars.
- A majority of experiments cost less than $30,000.
- One-third of the studies were followed up by another experiment that examined some old and some new variables.
- Investigators on 78 percent of the studies said they would not do their experiment differently if they were to repeat it today.

## CONCLUSIONS

In spite of the unsophisticated nature of the analyses in this report, some rather firm conclusions can be drawn regarding the methodological weaknesses found in this fourteen-year sample and what must be done to alleviate them. There is little evidence for thinking that these conclusions are not generalizable to similar research outside this sample or that any major changes have taken place in experimental strategies in the four years since the

sample was taken.* The methodological implications that can be drawn from these analyses are summarized briefly below.

Extent of experimental space

Generalizable experimental data that will predict performance quantitatively and with reasonable accuracy is not likely to be generated from experiments that examine only a few factors. The world is far more complex than any two-, three-, or four-factor study is likely to approximate. More factors must be examined before predictive precision can be achieved and at least three or more levels per factor must be studied to permit nonlinear relationships to be identified. Studying more factors may also increase the generalizabililty of the data. But even experiments with a large number of factors will not achieve the desired goals unless the factors that are included have a high probability of being the ones most critical for the task under consideration.

Size of the experimental effort

The considerable variability in the size of experiments studying the same number of factors suggests that many investigators may be spending more time, money, and effort than the amount of information being obtained justifies. One source of waste may be in the amount of work that goes into measuring the same information over and over again (replications) rather than using that same effort to investigate many more factors or an expanded experimental space. Another source of waste arises when data is collected to isolate higher-order effects that are ordinarily trivial.

---

*Results of these analyses were used in earlier publications to support the applications of new experimental techniques and strategies (see Simon, 1973;1974). Results have also been quoted in seminars on "advanced methodologies" given to several military organizations. Directly as a result of this exposure, a few investigators have begun to change their experimental methods.

Quality of experimental results

That there was so much of the experimental data that remained unexplained within the experiment suggests that in addition to studying unimportant variables there were also deficiencies in the cleanliness of the data collection and the thoroughness of the analysis. Particularly evident was the inadequate control of sequence effects that could occur when subjects were tested on a series of experimental conditions. Also the uneven selection of what would be called the error variance degraded the value of significant tests, which offer relatively little new information to most applied experiments, once the magnitude -- both relative and absolute -- of an effect had been determined.

Limitations of data collection plans and strategies

Overemphasis on analysis of variance models and factorial designs led to the traditional significance testing that at best helps identify critical variables instead of providing the multi-factor functions needed by the engineers. This stops short of where informative experiments should begin and is one of the reasons why the data is so frequently not useful in the design of real-world systems. When the data collection strategy begins with the assumption that every cell in a factorial design must be filled and the entire design replicated many times, it is understandable why so few experiments have studied as few as ten or twenty factors, or isolated out of fifteen to thirty candidate factors the ones having the most important effects on the experimental task to study in depth. Yet there are other designs and strategies that would enable the investigator to look at a great many variables quickly and relatively inexpensively; these take advantage of the relatively simple relationships found in human factors data and collect data in small increments which are analyzed to determine whether or not more data is required (Simon, 1973). Were regression analysis used more frequently, even the existing results would frequently be more informative.

## Application of results

The results suggest that the experiments may have been done more to satisfy the investigator than a customer. While the questionnaire could not answer how often the results have been used by others than the investigator or his sponsor, it did show a disappointingly low percentage of investigators that knew that their data had been applied to real systems. No questions were asked to determine whether quantitative or qualitative decisions were made on the basis of this data, although the common complaint regarding human factors experimental results is that they must be qualified to the point of making the original data unrecognizable. A sizeable increase in the number of critical factors being studied and a systematic decrease in the anomalies from careless data collection, along with more informative and economical experimental designs and analyses could result in a marked improvement in the quality, quantity, and usefulness of the data from human factors research.

## VI. REFERENCES


Adams, J. A. Research and the future of engineering psychology. American Psychologist, 1972, 27, 615-622.

Bakan, D. The test of significance in psychological research. In Steger, J. A. (Ed.), Readings in statistics for the behavioral scientist. New York: Holt, Rinehart and Winston, 1971.

Bozarth, J. D. and Roberts, R. R., Jr. Signifying statistical significance. American Psychologist, 1972, 27, 774-775.

Bunde, T. A. The applications of random balanced designs. Technometrics, 1959, 1, 139-155.

Cochran, W. G., and Cox, G. M. Experimental designs. New York: Wiley, 1957 (2nd edition).

Cohen, J. Statistical power analysis for the behavioral sciences. New York: Academic Press, 1969.

Edgington, E. S. A new tabulation of statistical procedures used in APA journals. American Psychologist, 1974, 29, 25-26.

Hays, W. L. Statistics. New York: Holt, Rinehart, and Winston, 1963.

Human Factors (1958-1972). Published by the Human Factors Society, Box 1369, Santa Monica, California 90406. (Back volumes are available on microfilm from Xerox University Microfilms, 300 N. Zeeb Road, Ann Arbor, Michigan 48106.

References (Cont)

Kleiter, G. The crisis of significance tests in psychology. Jahrbuch für Psychologie, Psychotherapie, und Medizinische Anthropologie, 1969, 17, 144-163. (Library Translation No. 1649, Royal Aircraft Establishment; The R.A.E. Library, Q. 4 Building, R.A.E. Farnborough Hants, England, June 1972.

Kraft, J. A. A 1961 compilation and brief history of human factors research in business and industry. Human Factors, 1961, 3, 253-283.

Lykken, D. T. Statistical significance in psychological research. Psychological Bulletin, 1968, 70, 151-159.

Nunnally, J. The place of statistics in psychology. Educational and Psychological Measurement, 1960, 20, 641-650.

Rozeboom, W. W. The fallacy of the null-hypothesis significance test. Psychological Bulletin, 1960, 57, 416-428.

Simon, C. W. The use of central-composite designs in human factors engineering experiments. Hughes Aircraft Company, Technical Report No. AFOSR-70-6, December 1970, (AD 748-277).

Simon, C. W. Economical multifactor designs for human factors engineering experiments. Hughes Aircraft Company, Technical Report No. P73-326, June 1973, 171 pp., (AD 767-739).

Simon, C. W. Methods for handling sequence effects in human factors engineering experiments. Hughes Aircraft Company, Technical Report No. P74-451, December 1974, (AD A006-240).

Simon, C. W. Methods for improving information from "undesigned" human factors experiments. Hughes Aircraft Company, Technical Report No. P75-287, July 1975, (AD A018-455).

References (Cont)

Vaughan, G. M. and Corballis, M. C. Beyond tests of significance: estimating strength of effects in selected ANOVA designs. Psychological Bulletin, 1969, 72, 204-213.

# APPENDIX A

## LOCATIONS IN HUMAN FACTORS OF 239 ANALYSES OF VARIANCE USED FOR THIS REPORT

| YEAR | MONTH | VOLUME NUMBER | FIRST *PAGE | ANOVA *PAGE | TABLE NUMBER |
|---|---|---|---|---|---|
| 59 | 8 | 1 | 72 | 74 | 1A |
| 59 | 11 | 1 | 1 | 4 | 3A |
| 59 | 11 | 1 | 1 | 5 | 4A |
| 60 | 2 | 2 | 28 | 32 | 1A |
| 60 | 8 | 2 | 141 | 145 | 4B |
| 60 | 8 | 2 | 141 | 145 | 4C |
| 60 | 8 | 2 | 141 | 145 | 4D |
| 60 | 8 | 2 | 141 | 145 | 4E |
| 60 | 8 | 2 | 141 | 145 | 4G |
| 60 | 8 | 2 | 141 | 145 | 4H |
| 60 | 8 | 2 | 141 | 145 | 4I |
| 60 | 8 | 2 | 141 | 145 | 4J |
| 60 | 11 | 2 | 228 | 231 | 1A |
| 60 | 11 | 2 | 228 | 233 | 2A |
| 61 | 3 | 3 | 53 | 56 | 3A |
| 61 | 3 | 3 | 53 | 57 | 4A |
| 61 | 7 | 3 | 93 | 96 | 2A |
| 61 | 7 | 3 | 99 | 104 | 1A |
| 61 | 7 | 3 | 108 | 114 | 2A |
| 61 | 7 | 3 | 120 | 127 | 4A |
| 61 | 7 | 3 | 131 | 140 | 5A |
| 61 | 12 | 3 | 229 | 231 | 2A |
| 62 | 6 | 4 | 125 | 129 | 2A |
| 62 | 8 | 4 | 193 | 194 | 1A |
| 62 | 8 | 4 | 193 | 194 | 1B |
| 62 | 8 | 4 | 193 | 197 | 2A |
| 62 | 8 | 4 | 193 | 197 | 2A |
| 62 | 8 | 4 | 201 | 203 | 1A |
| 62 | 8 | 4 | 201 | 204 | 2A |
| 62 | 8 | 4 | 201 | 204 | 3A |
| 62 | 8 | 4 | 201 | 205 | 4A |
| 62 | 12 | 4 | 367 | 369 | 1A |
| 63 | 4 | 5 | 109 | 111 | 1A |
| 63 | 4 | 5 | 109 | 114 | 4A |
| 63 | 4 | 5 | 117 | 121 | 4A |
| 63 | 6 | 5 | 335 | 336 | 1A |
| 63 | 6 | 5 | 335 | 336 | 2A |
| 63 | 8 | 5 | 379 | 382 | 3A |
| 64 | 2 | 6 | 3 | 8 | 2A |
| 64 | 2 | 6 | 39 | 45 | 1A |
| 64 | 2 | 6 | 39 | 45 | 2A |

| YEAR | MONTH | VOLUME NUMBER | FIRST *PAGE | ANOVA *PAGE | TABLE NUMBER |
|---|---|---|---|---|---|
| 64 | 2 | 6 | 57 | 60 | 2A |
| 64 | 2 | 6 | 71 | 75 | 2A |
| 64 | 2 | 6 | 71 | 76 | 3A |
| 64 | 2 | 6 | 111 | 112 | 1A |
| 64 | 4 | 6 | 127 | 131 | 1A |
| 64 | 4 | 6 | 157 | 160 | 2A |
| 64 | 4 | 6 | 165 | 173 | 8A |
| 64 | 4 | 6 | 165 | 173 | 9A |
| 64 | 4 | 6 | 165 | 173 | 10A |
| 64 | 4 | 6 | 179 | 181 | 1A |
| 64 | 4 | 6 | 179 | 181 | 2A |
| 64 | 4 | 6 | 209 | 214 | 2A |
| 64 | 4 | 6 | 209 | 214 | 2B |
| 64 | 4 | 6 | 209 | 214 | 2C |
| 64 | 4 | 6 | 209 | 214 | 2D |
| 64 | 4 | 6 | 209 | 214 | 2E |
| 64 | 4 | 6 | 209 | 214 | 3A |
| 64 | 4 | 6 | 209 | 214 | 3B |
| 64 | 4 | 6 | 209 | 214 | 3C |
| 64 | 4 | 6 | 209 | 214 | 3D |
| 64 | 4 | 6 | 209 | 214 | 3E |
| 64 | 6 | 6 | 233 | 236 | 1A |
| 64 | 6 | 6 | 233 | 237 | 2A |
| 64 | 6 | 6 | 233 | 237 | 2B |
| 64 | 6 | 6 | 253 | 255 | 1A |
| 64 | 6 | 6 | 253 | 255 | 2A |
| 64 | 6 | 6 | 257 | 260 | 1A |
| 64 | 6 | 6 | 257 | 260 | 2A |
| 64 | 6 | 6 | 257 | 261 | 5A |
| 64 | 6 | 6 | 257 | 261 | 6A |
| 64 | 8 | 6 | 319 | 321 | 3A |
| 64 | 8 | 6 | 327 | 330 | 2A |
| 64 | 8 | 6 | 327 | 330 | 2B |
| 64 | 8 | 6 | 327 | 330 | 2C |
| 64 | 8 | 6 | 327 | 330 | 3A |
| 64 | 8 | 6 | 327 | 330 | 3B |
| 64 | 8 | 6 | 327 | 330 | 3C |
| 64 | 8 | 6 | 343 | 347 | 2A |
| 64 | 10 | 6 | 475 | 476 | 2A |
| 65 | 2 | 7 | 28 | 33 | 2A |
| 65 | 2 | 7 | 28 | 33 | 3A |

| YEAR | MONTH | VOLUME NUMBER | FIRST *PAGE | ANOVA *PAGE | TABLE NUMBER |
|---|---|---|---|---|---|
| 65 | 2 | 7 | 28 | 33 | 4A |
| 65 | 2 | 7 | 28 | 33 | 5A |
| 65 | 2 | 7 | 28 | 33 | 6A |
| 65 | 2 | 7 | 28 | 33 | 7A |
| 65 | 2 | 7 | 54 | 59 | 2A |
| 65 | 2 | 7 | 54 | 59 | 4A |
| 65 | 2 | 7 | 54 | 59 | 5A |
| 65 | 2 | 7 | 71 | 72 | 2A |
| 65 | 4 | 7 | 107 | 113 | 2A |
| 65 | 4 | 7 | 129 | 134 | 2A |
| 65 | 4 | 7 | 129 | 135 | 4A |
| 65 | 4 | 7 | 155 | 159 | 2A |
| 65 | 4 | 7 | 155 | 159 | 2B |
| 65 | 6 | 7 | 185 | 193 | 5A |
| 65 | 6 | 7 | 185 | 194 | 6A |
| 65 | 6 | 7 | 185 | 194 | 7A |
| 65 | 6 | 7 | 207 | 210 | 1A |
| 65 | 6 | 7 | 207 | 211 | 2A |
| 65 | 6 | 7 | 207 | 212 | 3A |
| 65 | 6 | 7 | 219 | 222 | 3A |
| 65 | 6 | 7 | 219 | 222 | 3B |
| 65 | 6 | 7 | 231 | 235 | 4A |
| 65 | 6 | 7 | 231 | 236 | 7A |
| 65 | 6 | 7 | 237 | 241 | 1A |
| 65 | 6 | 7 | 245 | 249 | 1A |
| 65 | 6 | 7 | 245 | 249 | 1B |
| 65 | 6 | 7 | 245 | 252 | 3A |
| 65 | 10 | 7 | 483 | 487 | 1A |
| 65 | 10 | 7 | 483 | 487 | 2A |
| 65 | 10 | 7 | 493 | 498 | 5A |
| 65 | 10 | 7 | 493 | 498 | 7A |
| 65 | 10 | 7 | 493 | 499 | 9A |
| 65 | 12 | 7 | 513 | 516 | 2A |
| 65 | 12 | 7 | 513 | 516 | 3A |
| 65 | 12 | 7 | 513 | 517 | 4A |
| 65 | 12 | 7 | 513 | 517 | 5A |
| 65 | 12 | 7 | 521 | 524 | 3A |
| 65 | 12 | 7 | 527 | 533 | 1A |
| 65 | 12 | 7 | 569 | 576 | 4A |
| 65 | 12 | 7 | 569 | 576 | 4B |
| 65 | 12 | 7 | 569 | 576 | 4C |

| YEAR | MONTH | VOLUME NUMBER | FIRST *PAGE | ANOVA *PAGE | TABLE NUMBER |
|---|---|---|---|---|---|
| 66 | 2 | 8 | 41 | 44 | 2A |
| 66 | 4 | 8 | 121 | 125 | 1A |
| 66 | 4 | 8 | 121 | 125 | 2A |
| 66 | 4 | 8 | 121 | 125 | 3A |
| 66 | 4 | 8 | 121 | 125 | 4A |
| 66 | 4 | 8 | 147 | 150 | 1A |
| 66 | 4 | 8 | 147 | 152 | 3A |
| 66 | 6 | 8 | 245 | 255 | 9A |
| 66 | 6 | 8 | 245 | 256 | 11A |
| 66 | 6 | 8 | 245 | 256 | 11B |
| 66 | 6 | 8 | 245 | 256 | 12A |
| 66 | 6 | 8 | 245 | 258 | 13A |
| 66 | 6 | 8 | 245 | 258 | 14A |
| 66 | 6 | 8 | 245 | 260 | 16A |
| 66 | 6 | 8 | 245 | 260 | 17A |
| 66 | 10 | 8 | 407 | 411 | 1A |
| 66 | 10 | 8 | 407 | 412 | 2A |
| 66 | 10 | 8 | 427 | 431 | 1A |
| 66 | 10 | 8 | 441 | 444 | 1A |
| 66 | 12 | 8 | 481 | 486 | 2A |
| 66 | 12 | 8 | 481 | 486 | 3A |
| 66 | 12 | 8 | 563 | 565 | 2A |
| 66 | 12 | 8 | 569 | 571 | 2A |
| 66 | 12 | 8 | 569 | 571 | 3A |
| 66 | 12 | 8 | 569 | 571 | 4A |
| 67 | 2 | 9 | 45 | 48 | 2A |
| 67 | 2 | 9 | 45 | 48 | 3A |
| 67 | 2 | 9 | 45 | 48 | 4A |
| 67 | 2 | 9 | 45 | 49 | 5A |
| 67 | 2 | 9 | 45 | 51 | 7A |
| 67 | 2 | 9 | 45 | 51 | 8A |
| 67 | 4 | 9 | 93 | 99 | 1B |
| 67 | 4 | 9 | 93 | 99 | 2A |
| 67 | 4 | 9 | 93 | 101 | 3A |
| 67 | 4 | 9 | 93 | 101 | 3B |
| 67 | 4 | 9 | 105 | 111 | 3A |
| 67 | 4 | 9 | 105 | 111 | 3B |
| 67 | 4 | 9 | 105 | 114 | 5A |
| 67 | 4 | 9 | 105 | 114 | 5B |
| 67 | 4 | 9 | 119 | 123 | 2A |
| 67 | 4 | 9 | 175 | 177 | 1A |

| YEAR | MONTH | VOLUME NUMBER | FIRST *PAGE | ANOVA *PAGE | TABLE NUMBER |
|---|---|---|---|---|---|
| 67 | 4 | 9 | 175 | 178 | 2A |
| 67 | 6 | 9 | 239 | 245 | 2A |
| 67 | 6 | 9 | 239 | 245 | 3A |
| 67 | 6 | 9 | 239 | 248 | 12A |
| 67 | 6 | 9 | 257 | 260 | 1A |
| 67 | 10 | 9 | 409 | 416 | 2A |
| 67 | 10 | 9 | 419 | 423 | 3A |
| 67 | 10 | 9 | 427 | 431 | 2A |
| 67 | 10 | 9 | 455 | 458 | 1A |
| 67 | 10 | 9 | 455 | 459 | 5A |
| 67 | 10 | 9 | 461 | 465 | 1A |
| 67 | 10 | 9 | 479 | 481 | 2A |
| 68 | 2 | 10 | 27 | 29 | 1A |
| 68 | 2 | 10 | 27 | 29 | 1B |
| 68 | 6 | 10 | 201 | 205 | 1A |
| 68 | 6 | 10 | 201 | 207 | 2A |
| 68 | 6 | 10 | 201 | 208 | 3A |
| 68 | 6 | 10 | 217 | 219 | 1A |
| 68 | 6 | 10 | 259 | 261 | 2A |
| 68 | 6 | 10 | 283 | 290 | 4A |
| 68 | 8 | 10 | 303 | 308 | 2A |
| 68 | 8 | 10 | 333 | 337 | 2A |
| 68 | 8 | 10 | 333 | 339 | 5A |
| 68 | 10 | 10 | 489 | 491 | 1A |
| 68 | 10 | 10 | 497 | 501 | 1A |
| 68 | 10 | 10 | 497 | 501 | 1B |
| 68 | 10 | 10 | 505 | 509 | 2A |
| 69 | 4 | 11 | 189 | 192 | 1A |
| 69 | 6 | 11 | 239 | 242 | 1A |
| 69 | 6 | 11 | 245 | 247 | 1A |
| 69 | 6 | 11 | 251 | 253 | 1A |
| 69 | 6 | 11 | 251 | 253 | 2A |
| 69 | 6 | 11 | 251 | 255 | 3A |
| 69 | 6 | 11 | 257 | 264 | 3A |
| 69 | 6 | 11 | 257 | 266 | 6A |
| 69 | 6 | 11 | 257 | 268 | 9A |
| 69 | 8 | 11 | 321 | 323 | 2A |
| 69 | 8 | 11 | 331 | 335 | 2A |
| 70 | 2 | 12 | 13 | 18 | 2A |
| 70 | 2 | 12 | 13 | 19 | 4A |
| 70 | 6 | 12 | 261 | 265 | 3A |

| YEAR | MONTH | VOLUME NUMBER | FIRST *PAGE | ANOVA *PAGE | TABLE NUMBER |
|---|---|---|---|---|---|
| 70 | 6 | 12 | 331 | 337 | 1A |
| 70 | 6 | 12 | 341 | 346 | 4A |
| 70 | 8 | 12 | 391 | 397 | 3A |
| 70 | 8 | 12 | 391 | 397 | 3B |
| 70 | 8 | 12 | 391 | 397 | 3C |
| 70 | 8 | 12 | 391 | 397 | 3D |
| 70 | 8 | 12 | 391 | 397 | 3E |
| 70 | 8 | 12 | 391 | 397 | 3F |
| 70 | 8 | 12 | 391 | 397 | 3G |
| 70 | 8 | 12 | 391 | 397 | 3H |
| 70 | 8 | 12 | 391 | 397 | 3I |
| 70 | 10 | 12 | 485 | 491 | 5A |
| 70 | 10 | 12 | 493 | 495 | 2A |
| 70 | 12 | 12 | 553 | 555 | 3A |
| 70 | 12 | 12 | 559 | 560 | 2A |
| 70 | 12 | 12 | 599 | 602 | 1A |
| 71 | 2 | 13 | 59 | 61 | 1A |
| 71 | 4 | 13 | 163 | 166 | 2A |
| 71 | 4 | 13 | 173 | 175 | 2A |
| 71 | 4 | 13 | 177 | 180 | 3A |
| 71 | 6 | 13 | 247 | 252 | 2A |
| 71 | 6 | 13 | 283 | 287 | 1A |
| 71 | 6 | 13 | 283 | 288 | 3A |
| 71 | 8 | 13 | 363 | 366 | 1A |
| 71 | 10 | 13 | 435 | 443 | 2A |
| 71 | 10 | 13 | 435 | 443 | 3A |
| 71 | 12 | 13 | 503 | 507 | 2A |
| 72 | 2 | 14 | 65 | 68 | 1A |
| 72 | 4 | 14 | 181 | 183 | 1A |
| 72 | 6 | 14 | 199 | 203 | 1A |
| 72 | 6 | 14 | 227 | 230 | 2A |
| 72 | 6 | 14 | 227 | 231 | 4A |
| 72 | 6 | 14 | 227 | 232 | 6A |
| 72 | 6 | 14 | 227 | 234 | 8A |

DISTRIBUTION LIST

LCdr. James Ashburn, MSC, USN
NAMRL, Bldg. 1953
Pensacola, FL 32512

Dr. L. E. Banderet
SGDR-UE-CR
Dept. of the Army
U. S. Army Research Institute
of Environmental Medicine
Natick, Mass. 01760

Mr. Vernon E. Carter
Pilot Training Systems
Orgn 3750/62
Northrop Corp./Aircraft Div.
3901 W. Broadway
Hawthorne, CA 90250

Dr. Julien M. Christensen
Chairman, Dept. of Industrial Engr.
Wayne State University
Detroit, Michigan 48202

Mr. James Duva (N-215)
Naval Training Equipment Ctr.
Orlando, FL 32813

Dr. Gordon A. Eckstrand
AFHRL/AS
Wright-Patterson AFB OH 45433

Mr. Ronald A. Erickson, Code 3175
Head, HF Branch, Weapons Devel. Dept.
U. S. Naval Weapons Center
China Lake, California 93555

Dr. Marshall J. Farr
ONR, Code 458
800 N. Quincy Street
Arlington, VA 22217

Terrence W. Faulkner
Health & Safety Laboratory
Bldg. 56, Kodak Park Division
Eastman Kodak Co.
Rochester, N.Y. 14650

Mr. Charles A. Gainer
Chier, Army Research Unit
Bldg. 502, P.O. Box 428
Ft. Rucker, Alabama 36360

Dr. Robert A. Goldbeck
Mail Station S-32
Western Development Laboratories
Division
Philco-Ford Corporation
3939 Fabian Way
Palo Alto, California 94303

James E. Goodson, CDR MSC USN
Head, Aerospace Psychology Dept.
Code 15, Naval Aerospace Med. Research Lab.
Pensacola, Florida 32508

Dr. Tom Gray
AFHRL/FT
Williams AFB, AZ 85224

G. C. Helmstadter, Director
University Testing Services
Payne Hall, B302
Arizona State University
Tempe, Arizona 85281

Dr. Charles O. Hopkins
Head, Aviation Research Lab
University of Illinois
Willard Airport
Savoy, Illinois 61874

Dr. Richard Jagacinski
Human Performance Center
330 Packard Road
Ann Arbor, Michigan 48104

Dr. Edgar M. Johnson
U. S. Army Research Institute for the
Behavioral and Social Sciences
1300 Wilson Blvd.
Arlington, VA 22209

AFHRL/ASM (Patricia A. Knoop)
Wright-Patterson AFB OH 45433

Dr. Richard L. Krumm
P. O. Box 2706
Main Post Office
Washington, D.C. 20013

Mr. Robert G. Mills
6570th AMRL/HEB
Wright-Patterson AFB, OH 45433

Distribution List (Continued)

Dr. Howard L. Parris
AFHRL/CCS
Brooks AFB, Texas 78235

Dr. Wallace W. Prophet
Director, HumRRO Cent. Div.
400 Plaza Bldg.
Pensacola, FL 32505

Dr. James J. Regan
Navy Personnel R&D Ctr.
San Diego, CA 92152

Dr. Clyde R. Replogle
6750 AMRL/EME
Wright-Patterson AFB OH 45433

Charles V. Riche
School of Psychology
Georgia Institute of Technology
Atlanta, GA 30332

John E. Robinson, Jr.
Human Factors Staff
Building 606, M.S. G-233
Hughes Aircraft Co.
Fullerton, CA 92634

Dr. Marty Rockway
Technical Director
AFHRL/TT
Lowry AFB, CO 80230

Dr. Stanley N. Roscoe
Bldg. 6, MS D-120
Hughes Aircraft Co.
Culver City, CA 90230

Dr. Mark S. Sanders
Department of Psychology
California State University
Northridge, CA 91324

Dr. Dennis E. Smith
Mathematical Statistician
Desmatics, Inc.
P. O. Box 863
State College, PA 16801

Dr. Margaret J. Smith
Naval Education and Training
Program Development Center
Ellyson Field
Pensacola, FL 32509

H. C. Strasel
Chief, ARI Field Unit
U. S. Army Research Institute
P. O. Box 2086
Ft. Benning, GA 31905

Dr. Martin A. Tolcott
Human Engineering Div., ONR
800 N. Quincy Street
Arlington, VA 22217

Dr. Donald A. Topmiller
AMRL/HES
Wright-Patterson AFB OH 45433

AMRL/HE (Dr. Melvin J. Warrick)
Wright-Patterson AFB, OH 45433

Dr. R. Young, Director
Human Resources Office, ARPA
1400 Wilson Blvd.
Arlington, VA 22209

HQ AFSC/DLS
Andrews AFB, MD 20334

ERIC
Processing and Reference Facility
4833 Rugby Ave., Suite 303
Bethesda, MD 20014

HQ AFHRL/CC
Brooks AFB, TX 78235

Director, Behavioral Sciences Dept.
USAF Academy
Colorado Springs, CO 80840

Flight Dynamics & Control Division
Mail Stop 152
NASA - Langley Research Center
Hampton, VA 23665
Attn: Gary P. Beasley

Department of the Air Force
Air Force Human Relations Lab. (AFSC)
Lackland AFB, TX 78236
Attn: Mark Nataupsky, Capt., USAF
Chief, Evaluation Section
Personnel Research Division

Distribution List (Continued)

Military Asst. For Human Resources
OAD (E&LS)
ODDR&E
Pentagon, Washington, D.C. 20330

Executive Editor
Psychological Abstracts
American Psychological Assn.
1200 17th St. N.W.
Washington, D.C. 20036

HQ USAF/RDPS
Washington, D.C. 20330

AFFDL/CC
Wright-Patterson AFB, OH 45433

Director
USAF Avionics Laboratory
Wright-Patterson AFB, OH 45433

AMD/RDH (Col. George C. Mohr)
Brooks AFB, TX 78235

Dr. Chriss Dixon
10201 S. Cedar Lake Rd., #208
Minnetonka, MN 55343

Defense Documentation Center
Cameron Station
Alexandria, VA 22314

Education Research Information Center
Processing & Reference Facility
4833 Rugby Ave., Suite 303
Bethesda, MD 20014

NASA - Scientific & Technical Information Facility
P. O. Box 33
College Park, MD 20740

National Technical Information Services (NTIS)
Operations Division
5285 Port Royal Road
Springfield, VA 22151

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

| REPORT DOCUMENTATION PAGE | | READ INSTRUCTIONS BEFORE COMPLETING FORM |
|---|---|---|
| 1. REPORT NUMBER<br>AFOSR - TR - 77 - 0333 | 2. GOVT ACCESSION NO. | 3. RECIPIENT'S CATALOG NUMBER |
| 4. TITLE (and Subtitle)<br>ANALYSIS OF HUMAN FACTORS ENGINEERING EXPERIMENTS: CHARACTERISTICS, RESULTS, AND APPLICATIONS | | 5. TYPE OF REPORT & PERIOD COVERED<br>INTERIM SCIENTIFIC<br>6. PERFORMING ORG. REPORT NUMBER<br>CWS-02-76 |
| 7. AUTHOR(s)<br>Charles W. Simon | | 8. CONTRACT OR GRANT NUMBER(s)<br>F44620-76-C-0008 |
| 9. PERFORMING ORGANIZATION NAME AND ADDRESS<br>Canyon Research Group, Inc.<br>32107 Lindero Canyon Rd., Suite 123<br>Westlake Village, California 91361 | | 10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS<br>61102F<br>2313A4 |
| 11. CONTROLLING OFFICE NAME AND ADDRESS<br>Air Force Office of Scientific Research (NL)<br>Bldg. 410, Bolling Air Force Base<br>Washington, D. C. 20332 | | 12. REPORT DATE<br>August 1976<br>13. NUMBER OF PAGES<br>104 |
| 14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office) | | 15. SECURITY CLASS. (of this report)<br>Unclassified<br>15a. DECLASSIFICATION/DOWNGRADING SCHEDULE |

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Human Factors (Engineering)
Engineering Psychology
Experiments
Methodology
Research Strategy
Research applications
Experimental designs

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Two hundred thirty-nine experiments published in the journal Human Factors during the period from 1958 to 1972 were analyzed for the purpose of discovering the characteristics of their experimental plans, the quality and character of their results, and the degree to which these results had been applied to real systems. The analysis revealed that these experiments investigated too small an experimental space, showed essentially no diversity in their selection of a

-over-

20. ABSTRACT (Cont)

basic experimental design, collected far more redundant data than was needed, and failed to properly handle the irrelevant variance arising from sequence effects.

When the experimental results were reanalyzed to discover what proportion of the observed performance variance was accounted for by experimental factors, on average, the proportion was much less than half. This proportion increased, on average, as more factors were studied in an experiment, although for any size experiment, there were always some experiments in which the experimental factors accounted for practically none of the observed variance and some in which they accounted for most of it. There was empirical evidence from these experiments to show that third-order or higher interactions had only negligible effects on performance; this was so even though many of these sources had been found to be "statistically significant." Nearly a quarter of the main effects accounting for less than one percent of the total variance were statistically significant. Subjects used to replicate an experiment (and thus presumed to be homogeneous) generally accounted for much more of the performance variance than specifically selected factors of subject characteristics.

When a survey was made of those who conducted the experiments, it was discovered that slightly more than half of the experiments had been done to find answers of general applicability; less than a third ~~of the experiments~~ were known or believed to have influenced the design of a real system. A majority of the investigators said they would not do their experiments any differently if they were to repeat them today.

Numerical data in support of these and other results are supplied along with some limited discussion on the implications of this analysis for an improved experimental methodology.