EOTR-86-7

FINAL REPORT FOR PHASE I

Meta-Analysis of Human Factors Engineering Studies
Comparing Individual Differences, Practice Effects
and Equipment Design Variations

Marshall B. Jones
Robert S. Kennedy
Janet J. Turnage
Lois Ann Kuntz
Sherrie Ann Jones

DTIC
SELECTED
JUN 1 7 1986
D

February 21, 1985

REPORT DOCUMENTATION PAGE

1a. REPORT SECURITY CLASSIFICATION UNCLASSIFIED		1b. RESTRICTIVE MARKINGS			
2a. SECURITY CLASSIFICATION AUTHORITY		3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited			
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE					
4. PERFORMING ORGANIZATION REPORT NUMBER(S) EOTR-86-7		5. MONITORING ORGANIZATION REPORT NUMBER(S)			
6a. NAME OF PERFORMING ORGANIZATION Essex Corporation	6b. OFFICE SYMBOL *(If applicable)*	7a. NAME OF MONITORING ORGANIZATION Department of the Air Force Air Force Systems Command			
6c. ADDRESS *(City, State, and ZIP Code)* 1040 Woodcock Road, Suite 227 Orlando, FL 32803		7b. ADDRESS *(City, State, and ZIP Code)* Wright Patterson AFB OH 45433-6503			
8a. NAME OF FUNDING/SPONSORING ORGANIZATION Dept. of Air Force	8b. OFFICE SYMBOL *(If applicable)*	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER Contract F33615-85-C-0539			
8c. ADDRESS *(City, State, and ZIP Code)* Wright-Patterson AFB, OH 45433-6503		10. SOURCE OF FUNDING NUMBERS			
		PROGRAM ELEMENT NO. -	PROJECT NO. -	TASK NO. -	WORK UNIT ACCESSION NO. -

11. TITLE *(Include Security Classification)*
Meta-Analysis of Human Factors Engineering Studies Comparing Individual Differences, Practice Effects and Equipment Design Variations

12. PERSONAL AUTHOR(S)
M. B. Jones, R. S. Kennedy, J. J. Turnage, L. A. Kuntz, and S. A. Jones

13a. TYPE OF REPORT Final	13b. TIME COVERED FROM 7/30/85 TO 2/22/86	14. DATE OF REPORT *(Year, Month, Day)* February 21, 1985	15. PAGE COUNT 40

16. SUPPLEMENTARY NOTATION

17.	COSATI CODES		18. SUBJECT TERMS *(Continue on reverse if necessary and identify by block number)*
FIELD	GROUP	SUB-GROUP	Human factors engineering, manpower & personnel, human performance; individual differences; training; systems engineering, MANPRINT & HARDMAN, cost-effectiveness

19. ABSTRACT *(Continue on reverse if necessary and identify by block number)*
An approach which has as its theme the trade-offs between parameters wherein different combinations of human and equipment variations result in the same (Iso) performance outcome is described. It emphasizes Individual Differences and Practice effects as contributors to improved systems performance subject to the influence of Equipment design. Bringing them together and quantitatively in one model as contributors to better performance is one of the innovations of the Isoperformance model. A quantitative analysis was conducted of the human factors literature which compared the effects on human performance of equipment features, practice, and individual differences. Of over 10,000 citations scanned, only 10 permitted sufficient detail for calculation of Omega Squared. This final yield (0.01%) is a sobering commentary on the raw material that serves as our human factors engineering technological data base. The Isoperformance model is offered as technical framework in which to make extrapolations from the existing literature to real-world situations. They will be exemplified by experiments carried out for the purpose and implemented under this technical framework.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☑ UNCLASSIFIED/UNLIMITED ☐ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION
22a. NAME OF RESPONSIBLE INDIVIDUAL Steve A. Heckert	22b. TELEPHONE *(Include Area Code)* (513) 255-7593 22c. OFFICE SYMBOL AAMRL/HED

DD FORM 1473, 84 MAR   83 APR edition may be used until exhausted.   SECURITY CLASSIFICATION OF THIS PAGE
All other editions are obsolete.   UNCLASSIFIED

ISOPERFORMANCE FINAL REPORT

Table of Contents

PROLOGUE

The authors recognize that not all readers who consult this Final Report will have the same purpose. Therefore, after the Executive Summary which covers the program in general, we have attempted to accommodate these differing needs by organizing the paper in sections. Not all sections need to be read by all readers. The rationale for our proposed isoperformance approach is in the Introduction (Section II). This is mainly a modern version of our previously submitted Phase I proposal and can be omitted if one is familiar with the original document or is already convinced that Individual Differences, Practice Effects and Equipment Features need to be combined into a single man/machine performance model for human engineering. Section III is intended chiefly as the final report of the present effort and covers the literature review we conducted in Phase I along with a description of the Omega Squared meta-analysis and findings. However, while these conclusions conprise the chief impetus for the for the approach to be followed in Phase II Section III does not include the conceptualizations of Isoperformance model nor address the key technical problems and proposed solutions. To some extent we attempt to bridge this gap in Section IV where we provide a description of the analytic conceptualization between the isoperformance model and the broader notion of Performance Reckoning. For the interested reader, or the one who requires more graphic and analytic detail of how we hope to accomplish our goals, and overcome the key hurdles, we have included as Appendix A, the pertinent technical sections of what will be the Phase II proposal. A formal submission of Phase II proposal will be submitted within 30 days. Section V, the reference list, is for all sections.

Accesion For	
NTIS CRA&I	☑
DTIC TAB	☐
Unannounced	☐
Justification	
By	
Distribution /	
Availability Codes	
Dist	Avail and / or Special
A-1	

EXECUTIVE SUMMARY

Problem

Twenty years ago we were taught that the future of human factors engineering would be in gathering human input/output data and reporting transfer functions. The plan was to use these data to generate standards and specifications whereby requirements could be imposed on design engineers. It was believed that design engineers were waiting for these data and would incorporate them, once available, into new systems. In retrospect, this goal, while lofty, was naive. Admittedly, the availability of such data, properly implemented, can improve systems performance. But human engineering specification data which limit the design engineeer without at the same time providing alternatives are probably not going to be implemented and may not be welcome. Such an approach fails to recognize that the engineering design process is accustomed to trading-off between various hardware and software parameters in order to obtain the desired outcome.

Isoperformance

We advocate an alternative approach which has as its theme the trade-offs between engineering (hardware and software) parameters and human engineering parameters. In our original formulation, we proposed that there were various combinations of equipment sophistication, training time, and human capabilities that could, when combined, still permit the total systems performance to be at the same level. We named the approach isoperformance. In following this method one selects an operationally determined performance outcome. Then one determines the different combinations of Training, Equipment and Individual Difference variations that would result in that same (Iso) performance outcome. The value of this approach is that one could then compare the relative costs of the differing combinations which lead to the same performance and thereby select the less expensive approach. Alternatively, if one were confronted with a given level of one or more of the three variables, then one would be able to determine the suitable level for the other variable(s) in order to achieve the predetermined performance outcome. We now recognize that there are other uses of such a paradigm, and we will refer to the broader utilization as Performance Reckoning".

Definitions

Several synonyms are used in this report for the three dimensions of interest ("Equipment", "Practice", and "Individual Differences"). For example, if the source of the Individual Difference variations emerged from an analysis of variance table, it will probably be called Subjects. On the other hand, when using the language of military manpower management·specialists we may employ the terms, People or Personnel. If reporting on a specific experiment, we may also use the term Aptitude or Ability or Sensory Capability to convey this dimension. So, too, by Equipment comparisons, we may mean features that can be varied on a single piece of hardware (e.g., brightness, resolution or contrast) or several disparate engineering options (e.g., artificial intelligence versus unaided displays) or different software

modifications (e.g., rate aiding, predictor display). Finally, Practice can also be viewed under several different rubrics such as the number and length of trials, instructional systems (e.g., lecture, on-the-job, text), or the type of practice (massed versus distributed). In the discussions which follow, when we employ these and other denotations to report the specific outcomes, it is always our intention to connote the more general notion of the broader class of the dimension.

Opportunity

The opportunity for this effort was suggested to us by a series of recent reports in the literature on the successful application of meta-analysis as a method for integrating large bodies of technological data. This prompted us to take a retrospective look at an experimental program that we have been involved in over the past several years where the experimental focus was to examine the effects of equipment features on transfer of training from ground based flight simulators. All of these simulator studies employed multivariate analyses of the data and most combined examination of the effects of Equipment, Practice, and Individual variations on performance in a single study, although the influence of the latter two dimensions was not appreciated originally in that program. From such an approach, one can determine a breakdown of the total variance attributable to each of the main effects plus some interactions of these. Although Equipment features were our main thrust in these studies, our general finding was that the Individual Differences or Subject variables accounted for a very substantial proportion of the total explained variance and more than either Practice or Equipment variations. Furthermore, as a rule, Practice factors accounted for more variance than Equipment factors. Interactions were small. These findings are consistent enough to suggest lawful relationships, but since they had all been conducted in one laboratory, following similar paradigms, we wondered about the generalizability of our findings to the human factors literature in general. The _isoperformance_ model was formulated as a framework. We then sought to study the literature to determine whether numerical estimates of the relative contributions of each of these three variations were available.

Phase I Findings

A literature review was conducted on the human factors literature and a meta-analysis was performed for those studies identified as suitable for calculating Omega Squared. This calculation is a normalized measure of relationship which permits quantitative comparison between experiments with widely differing characteristics in sample size, training methods, and equipment options. Of over 10,000 citations scanned; 276 involved experimental studies of training and performance as a function of equipment variations; 68 implied an analysis of variance; 30 reported ANOVA data; and only 10 permitted sufficient detail for calculation of Omega Squared. This final yield was a miniscule 0.01% of the original number, an important and somewhat sobering commentary on the raw material that serves as our human factors engineering technological data base.

The meta-analysis of the 10 studies for which sufficient data were available was revealing, but disappointing. It showed that Omega Squared is a computation that can be performed if the experimental outcomes are fully

reported and the designs adequately conceptualized. Unfortunately 10 studies are too few and the data are too irregular to permit sufficient generalizations about trends in these studies. Certainly there is insufficient regularity in published studies to implement in an _isoperformance_ model. For example, three of the five studies with high omega-squared values for Subjects involved no Equipment variation, but two did. Thus the absence of an Equipment variation did not explain the high value of Omega Squared. A similar situation prevailed among the four studies with low values of Omega Squared: two involved several important Equipment variations but the other two did not. Therefore, we found it impossible to integrate the findings of these reports even when they contained the necessary ANOVA information, because of the multiplicity and noncomparability of fixed-effect measures. This result carries the clear implication that meta-analysis of the existing literature will _not_ suffice to implement the _isoperformance_ approach. If extrapolations from the existing literature to real-world situations are to be made, they are going to have to be exemplified by formal experiments carried out for the purpose and implemented under an innovative technical framework. Such a framework is described in summary form in Section IV and in greater detail with experimental exemplification of the framework and application to a real-world situation in Appendix A.

The expanded framework is called Performance Reckoning and one form of such an approach is Isoperformance. Since the technique involves estimating in advance the likely consequences of particular combinations of Personnel, Training, and Equipment for military systems performance, it is like a front-end analysis, and it bears a resemblance to other comprehensive approaches such as MANPRINT and HARDMAN. It is different from other models we know of (e.g, SAINT, THERP, HOS, etc) because it includes an emphasis on Individual Differences and Practice effects as contributors to improved systems performance subject to the influence of Equipment design. Indeed, under specific circumstances, all three variables can singly or together be more important or less. Bringing them together and dealing with them quantitatively in one model as contributors to better performance, not just as occasions for increased error or dispersion, is one of the innovations of the Isoperformance model. Thinking of them as possible elements to be traded-off against each other is the main theme of our approach.

Future Directions

To further the model, in Phase II, we propose to conduct two formal experiments which will provide the necessary data to validate the model and the model's assumptions. These studies are not just experiments. Rather they are designed to be programmatically related to each other and directed at providing numerical estimates for entering into the _isoperformance_ model. In order further to carry out the Isoperformance model a series of technical problems must be addressed and met. Some of these include: methods for segmentalizing practice; how to handle the decreasing correlations that may occur with practice; changing nature of factor structure with practice, discrete versus continuous equipment variations, and one- and two-way interactions. These are the chief obstacles to implementing the _isoperformance_ model, and these problems and attendant solutions will comprise the major undertaking of our work plan.

These studies will examine the effects to performance of varying (in one study) Aptitude, Practice, and Equipment and (in the other) Aptitude, Practice, and Training method. The resulting data will be analyzed for adequacy of the isoperformance model in deriving isoperformance curves in both experiments. Then we will set out to apply this model to a real-world military performance system using the existing scientific literature and expert opinion.

Other plans include detailing plans for Phase III, which will include specific potential uses by the Department of Defense. The goal for Phase III will be the broader production of these software programs for use elsewhere in DOD and the private sector. In the future we intend to take these methods one step further by providing an interactive computer-based simulation subroutine which will permit resource and program managers, systems engineers, instructional developers, and human factors practitioners to ask "what if" questions. It could be used as a training device for these people and for students whereby practical and theoretical "what if" questions may be posed on a desktop microcomputer which will contain real and simulated data bases for all three of the dimensions which interact within the model (Individual Differences, Training, and Equipment Features).

Provided the trade-off approaches described are feasible, conventional methods of systems design could change significantly. While a full exploration of Performance Reckoning space could take many years and considerable resources, the resultant information would be directly and credibly understandable by program managers or other makers of decisions about the allocation of system resources, since the format of presentation would more closely match that used for selecting configuration of hardware and software resources.

SECTION II

INTRODUCTION

Problem

Human factors engineering as a discipline with a formal name has a brief history (Taylor, 1963). It is accepted that human performance in complex systems is a function of human-machine interaction. Such interaction is the focus of attention of design engineers on the one hand, and of human performance scientists on the other. We have been taught as human factors practitioners that it is our role to gather human input/output data (transfer functions) with the hope that these data would generate standards and specifications which could be used by design engineers and thereby improve systems performance. It was believed that design engineers were eagerly waiting for these data to incorporate into new systems which would permit efficient allocation of functions between man and machines.

One of the intentions of this report is to call attention to the fact that we need to make greater efforts at preparing for trade-offs. This is not the only problem, however. For example, it is believed that the field of human factors has failed sufficiently to take into account the reliable variations in the human operator(s) who must operate the hardware. The consequences of this lack are broad, because it is not the simple omission of a limitation, but rather a lack of inclusion of a dimension which properly harnessed can result in IMPROVED systems performance. Some man-machine performance models do include individual differences (cf., Pew, Baron, Feehrer & Miller, 1977, for a review) but they tend to treat them as dispersion, not to our knowledge as correlates. Recent experiences with military weapons systems, underscore this problem and show for example that, "...the performance of the Stinger system is closely linked to the capabilities and training of the gunner." (Tice, 1986, p. 6). Indeed AFQT or ASVAB category scores have been related to tank commander (Wallace, 1982), gunnery (Tice, 1986) and an entire series of military occupational specialty performances (Carter & Biersner, 1985).

Additionally, although elaborate metrics (Shingledecker, 1982; Reid, 1982; Wickens, 1984) have been developed to evaluate the workload which is required for different systems, little account has been taken of the way that such a metric would need to be modified to allow for the change which takes place after the extended Practice which invariably accompanies military use of systems (cf. Hagman & Rose, 1983; Schendel, Shields, & Katz, 1978; Lane, 1986 for reviews). Therefore, in addition to Individual Differences, differing levels of Practice with the system is a second variation which we believe is poorly indexed in the field of human factors engineering design and is largely absent from man/machine models. These notions form the genesis of the Isoperformance model which will be described in this report. It should be pointed out that these issues have been alluded to from the very beginnings of the description of our discipline (cf., e.g., Fitts, 1963), but to our knowledge formal mechanisms for incorporating these mechanisms into one paradigm and techniques for implementing such an approach have not been available.

Background

The first serious comprehensive compilation of technical data on human factors engineering was a series of published lectures conducted at the U.S. Naval War College (Chapanis, Garner, Morgan & Sanford, 1947), which were later revised into a text (Chapanis, Garner & Morgan, 1949). There are other related documents of historical interest (e.g., McFarland, 1953; Committee on Undersea Warfare, 1949; Armstrong, 1943) but they are less pointedly directed at the relationship of human capabilities to equipment design. As early as 1952, there was interest within the DOD in assembling a design handbook, but it was not published until 1963 (Morgan, Cook, Chapanis & Lund) and somewhat after a book with a similar purpose by Woodson (1955). The different agencies within DOD were more self-conscious about picking up on the new human factors technology and made better early progress.

So far as we can tell, the initial proposal for human engineering standards by a federal agency first appeared in the Air Force with the publication of WDT Exhibit 57-8, released August 1957, updated March 1958, and revised November 1958, as AFBM Exhibit 57-8A, "Human Engineering Design Standards for Missile System Equipment" (U.S. Air Force, 1958). In November 1959 MILSTD-803 (Department of the Air Force, 1959) superseded AFBM Exhibit 57-8A and represented the first military standard for human engineering design.

At about the same time, Army and Navy documents were created paralleling the Air Force standards . In the Navy in 1959 (MIL-H-22174 AER) "Human Factors Data for Aircraft and Missile Systems" (Department of Defense, 1959) appeared. Army human engineering standards probably began in 1961 with the ABMA-STD-434 "Weapons System Human Factors Engineering Criteria" (Army Ballistic Missile Agency, 1961). This document descended directly from ABMA-XPD-844, "PERSHING Weapon System Human Factors Engineering Criteria" (Army Ballistic Missile Agency, 1959) which only applied to ballistic missile development and free flight rocket systems. It was superseded by MILSTD 803A1 (USAF), "Human Engineering Design Criteria for Aerospace Systems and Equipment" (Department of the Air Force, 1964). Three additional parts were added and these were directed toward aerospace facilities and vehicles and attempts were made to quantify criteria. December 1964 saw the expansion of MILSTD-8031A to MILSTD-803A2 (Department of the Air Force, 1964) and included were dimension additions, a table on display lighting, a section on hazards and safety and an environment section.

The second epoch (of human factors and systems engineering) began in 1964, where the Department of Defense was studying the possibility of creating a minimum package of human engineering requirements for tri-service use. This study (Chaikin & Chaillet, 1965) was completed 1 October 1965 and became the now well-known MILSTD-1472 which is applicable to all military systems, equipment, and facilities (Chaikin, 1978). Concurrently, when Military Specification MIL-H-22174 was revised as MIL-H-81444(AS) (Department of Defense, 1966) data for systems analyses were generated. These sources only documented analyses conducted during the design phase and were not intended to rationalize the design. MILSTD 803A (Department of the Air Force, 1967) was a further expansion of MILSTD 803A1 and an update of MILSTD 803A2 dealing with aerospace vehicles and vehicle equipment. Later, MIL-H-46855, "Human

Engineering Requirements for Military Systems, Equipment and Facilities"
(Department of Defense, 1970a) was issued.

Several methodologies now exist for the implementation of human factors
engineering design criteria and standards; and modern manuals and handbooks
are available for guidance (viz., Woodson, 1955; Boff, 1984; Morgan et al.
1963; Department of Defense, 1981; Malone, Shenk, & Moroney, 1976; Perkins,
Binel, & Avery, 1983). Human performance models for man-machine systems
evaluation are available (cf., Pew et al., 1977, for a review). Perhaps
because it is easier to criticise that to create, over the past 20 years, much
of the improvement in these systems approaches has been in an emphasis on test
and evaluation rather than on design (Kearns, 1982). "Reverse engineering"
(Marcus & Kaplan, 1984) is an attempt at feeding back into systems design the
conclusions that most affect human factors manpower and training
considerations. The application of reverse engineering represents a direct
recognition that human factors, manpower, personnel, and training are
critically important inputs in the weapons acquisition process.

Similarly, the Manpower and Personnel Integration (MANPRINT) initiative
makes the following considerations imperative in the materiel acquisition
process: human factors engineering; manpower/personnel/training (MPT);
systems safety, and health hazard assessments (cf., U.S. General Accounting
Office, 1985 for a bibliography of relevant studies within the three military
services). One important MANPRINT contribution to research and development
for materiel acquisition is the origination of generic analytic tools for
answering important allocation questions such as can soldiers operate
equipment effectively, how do complex man-machine systems work, and how much
and what kind of training is needed? A generic analytic tool, Hardware versus
Manpower (HARDMAN) (Mannele, Guptill, & Risser, 1985) provides a baseline
comparison methodology and uses operational concepts to predict MPT needs.
This type of analysis provides information about required sustainment costs,
training costs, and projects how many people will be needed to service and
operate systems in the field. Another generic analytic tool uses simulated
equipment to develop operational concepts in laboratories before any money is
spent to build weapon systems.

Despite MANPRINT and other attempts to use human factors engineering and
systems analysis to help man-machine systems reach maximum performance within
specified constraints, we believe that inadequate attention appears to be paid
to Individual Differences and Practice effects as related to human factors
engineering design. Moreover, neither of these are well incorporated into
military standards in any formal way. Therefore, they are largely ignored in
the design of equipment. An exception we know of is the leverage that can be
applied by modelling anthropometric differences between members of a user
population (cf., Bittner and Moroney, 1984, 1985, for a description of this
approach). A full documentation of the individual influence of Individual
Differences and Practice and how they may impact on suitable design of systems
goes beyond the scope of the present review, but some examples of each
follow.

Individual Differences

These differences include all of the many identifiable variations in people from sensory sensitivities and anthropometric variances to mental capabilities. For example, the distance at which one pilot customarily detects opponent aircraft is sometimes 50-70% better than another, resulting in 2-3 mile advantages in early detection (Jones, 1981, personal communication). This finding has obvious implications for winning in air combat (Ault, 1969, Campbell, 1970). Moreover, some pilots who are better at visual detection can even "outsee" the poorer ones when the latter use telescopes (Jones, 1981, personal communication). In this example, if Equipment factors were evaluated to determine effects on performance in terms of the amount of variance accounted for, one could not adequately assess the question without taking into account the differing performances of the individual pilots.

Cognitive and other mental capabilities also show wide variation (cf. Schoenfeldt for a review, 1982). There are also substantial Individual Differences in basic information processing capacities (Rose, 1978). For example, the speed of mental rotation varies considerably across individuals. A recent study (Hunt, 1984) found that the fastest subject could perform a mental rotation at approximately 2.5 degrees per msec compared to 18.5 degrees for the slowest subject. Men are generally faster at rotation than women, and young adults are generally faster than people in their 30s and beyond (Berg, Hertzog, & Hunt, 1982). Even among good readers by general population standards, there are substantial variations in the speed of lexical identification. In one study, there was approximately a 25% variation in speed (560 to 700 msec) between the faster and the slower lexical decision makers (Hunt, Davidson, & Larsman, 1981; Palmer, McLeod, Hunt, & Davidson, 1983). People also vary markedly in the number of sentences that they can process while still being able to recall the words. College students show differences of from 2 to 5 sentences, and people who show more "verbal aptitude" seem to have markedly longer spans (Daneman, 1983).

While mental competence is apparently bounded by a person's information processing capabilities, there are very large variations in performance within these bounds which may be attributable to differences in problem solving strategy and by knowledge of a content area. For instance, one study explored models of strategy and strategy shifting on a spatial visualization task using high school and adult subjects (Kyllonen, Woltz, & Lohman, 1981). For each of three successive task steps (encoding, construction, and comparison), different models apply for different subjects suggesting that different subjects used different strategies for solving the same items. Numerous other studies (e.g., Yalow, 1980) provide evidence that neither aptitude nor instructional treatment alone can fully describe learning and performance outcomes. Interactions between them exist and are consistently demonstrated. Instructional supplements can effectively "fill-in" for student weaknesses and reduce differences between high and low ability students. However, such supplements must be used with caution because reducing the difficulty of instructional materials may enhance immediate learning but fail to display any long-term advantages.

At the physical end of the human performance spectrum, muscular strength (Alluisi, 1978, p. 354) also shows sufficiently wide variances such that, in tasks which require upper body lifting, one would find that the 95th percentile female could not perform as well as the average male. At the more global end of human performance team performance (tanks) is largely a function of the intelligence of the tank commmander (Wallace, 1982). Individual Differences such as these have obvious implications for human factors engineering design because they can overshadow the effect of Equipment modifications themselves. Yet there is no formal mechanism to incorporate them into military standards, nor do any of the manpower management systems deal with them effectively.

Practice Effects

Recently a large review of the learning literature has been completed for DOD (Lane, 1986) where the sheer magnitude of the information defies simple report. Yet many lawful relationships exist. It is well known that the shape of the learning function is such that the most rapid amount of learning occurs initially and the best description of the overall relationship is that log trials (or Practice) is a linear function of log performance (Newell & Rosenbloom, 1981). What this means is that ranges of improvement in performance during military training in formal schools can be an order of magnitude of improvement for each epoch of time spent in training (cf., Schendel et al., 1978; Hagman & Rose, 1983; Lane, 1986, for reviews). Therefore, improvements of as much as 500% are not unusual. It follows that tasks which can only be performed with great difficulty and extreme concentration initially, may be performed with far less mental attention after modest amounts of Practice. Moreover, the advantages of display aiding (e.g., Smith and Kennedy, 1976) or artificial intelligence may be largely during these intial stages and of far less utility when the learning curve has slowed down. Such a range of improvements can temper any expected change due to Equipment factors.

Although some of these findings have been used for decisionmaking in industrial settings they appear not to have found their way into the existing manpower management models, like HARDMAN and MANPRINT. Furthermore, these improvements with Practice can be compounded by the fact that there are also large Individual Differences in Practice effects. For example, Kennedy, Bittner, Harbeson, and Jones (1982) found that performance improvement on a video game task proceeded at very different rates, and some of those who learned slowly at first eventually outperformed the fast learners if sufficient trials were given. Because of large Individual Differences in rates of learning, accuracy of prediction suffers when performance data are collected too early. A large literature (reviewed in Harbeson, Bittner, Kennedy, Carter, & Krause, 1983) is available showing representative ranges of these relationships. Further these aptitude by treatment interactions (ATI) Snow, 1980) have shown that the relation of general ability to learning tends to increase as instruction places increased information processing burdens on learners, and to decrease as instruction is designed to reduce the information processing demands on learners.

The problem we outline above is not one which will lessen with time, but rather the converse. We believe that the problem of function allocation

becomes more critical with the growing complexity of man-machine systems. Since the publication of a landmark article by Fitts in 1951, little progress has been made toward the solution of this problem. Fitts proposed what is now informally called the "Fitts list." This two-column list compares one column headed by the word "man" and another column headed by the word "machine." Fitts' recommendation was to compare the functions for which man is superior to machine to the functions for which the machine is superior to man. While rational, this formulation has yielded little progress in our understanding of man-machine systems interactions and tells us little about how to determine trade-off allocations of function (Jordan, 1963). The twentyfive year old comment by Swain and Wohl (1961) is still current that: "There is no adequate systematic methodology in existence for allocating functions between man and machine. In our view this lack is the central problem in human factors engineering today."

Opportunity

To achieve such a goal, we have proposed a technique toward improving human factors engineering performance measurement which embraces and uses as a theme the notion of "trade-off technology" . This approach deals with total or systems performance and points out that differing combinations of Individual Differences, Training, and Equipment variables can lead to the same desired outcome.

The specific opportunity in this area emerged from our experience with the experimental conduct of flight simulation studies, and the use of multivariate analyses of the data (Simon, 1976). The studies from the Navy's Visual Technology Research Simulation program, with which we have been involved for eight years, contained encouraging results for such a model (Lintern, Nelson, Sheppard, Westra, & Kennedy, 1981). In experimental studies of the effects of performance and Equipment, including Individual Variations, one emerges from the analysis with a breakdown of the total variance attributable to each of the main effects; "Equipment," "Practice," "Aptitude," and some interactions of these (cf. Kennedy, Berbaum, Collyer, May, & Dunlap, 1983).

Our general finding in analyses of studies of this sort is that the Individual Differences or Aptitude variables account for a very substantial proportion of the total explained variance, and more than either Practice or Equipment Variations (Lintern & Kennedy, 1984; Westra & Lintern, in press; Westra, Simon, Collyer, & Chambers, 1982). Furthermore, as a rule, Practice accounts for more than Equipment (Lintern et al., 1981). This finding permitted us to make an inference that could be useful: it allowed us to say something about the importance of the three major components in the determination of performance at the end of appreciable lengths of Practice. What it did not do is give us any explicit understanding of the trade-offs among the three major components relative to producing a given level of performance.

The Isoperformance Model

We apply the term isoperformance to describe this specified level of performance. For example, supposing we were to fix on a given level of performance, and were to ask ourselves how Aptitude and Practice varied. What

combinations of Aptitude and Practice would produce that particular level of performance? Now the answer to this latter question would result in an equation such that we could take a very high Aptitude person and with relatively little Practice arrive at this same level of performance, or with a lot more Practice, we could take a low Aptitude person and arrive at the same level of performance. Similarly, investment in Equipment features may elevate the performance of low ability persons, perhaps with modest Practice. Such trade-off statements about human performance have a great deal of value for applied systems engineering purposes because we can then attach dollar values to the Practice and the Equipment, and possibly even to selection (Aptitude) and classification and obtain some notion of to what extent we can exhange one major component for another in the contribution of operational performance.

Just as important, it may be that no amount of Practice will compensate for certain deficiencies in Equipment or in Aptitude. Furthermore, natural Aptitude or previous experience would be equally important because it would mean that if we wish to train people to certain high levels of performance we may have to take them only from a relatively high Aptitude range; in other words, there may need to be cut-off scores. Moreover, if we are asked to admit larger percentages of the applicant population, perhaps because the available pool is becoming smaller (cf. e.g., Merriman & Chatelier, 1981), we need to know whether Training or Equipment can be substituted and at what cost.

In the three dimensional schematic below we have shown the three dimensions (Individual Differences, Equipment and Practice) in x, y and z space and have depicted two surfaces which correspond to two different levels of the same (Iso) performance. In this example all points on the same surface indicate that the same level of performance is obtained by different mixtures of amounts of the x, y and z ingredients. So, descriptively, it may be that 100% of the students could obtain the same level of performance with three weeks of training and an aided display as 70% of the better students with less training (say two weeks) and no display aiding. So, too, the best performance may only be attainable with the best equipment and maximum practice, but not by all the students.

students   equipment

training
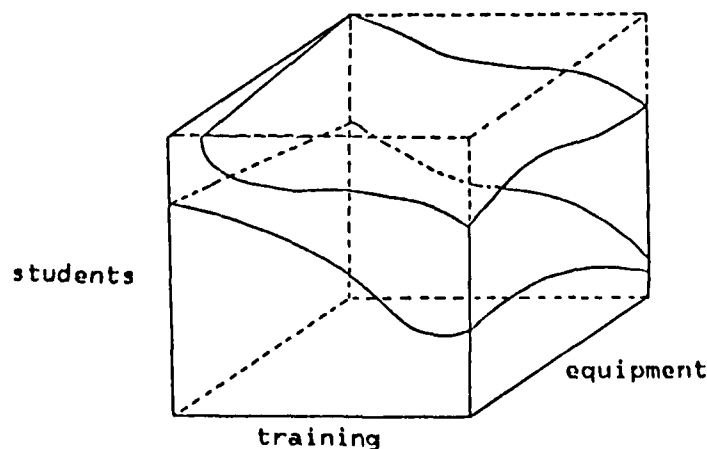
These latter considerations suggest a regression approach that would give explicit and numerical substance to the type of trade-off or compensatory sorts of relationships to which we refer. There is no technical obstacle to translating the results of an analysis of variance into a regression model along the lines that we have just suggested. The analysis of variance is a special case of the general linear model and, in fact, there are statistical packages which permit us to read out the results of what could be an analysis of variance in regression form. We could then obtain standardized regression weights for each of the major components for the dependent variable which would permit the problem to be in the form needed in order to carry out trade-off decisions.

Significance

Clearly what is needed is a methodology to incorporate various combinations of Individual Differences and Practice variables. To improve human factors engineering performance measurement, we also need to include differing combinations of Equipment variables. The present proposal constitutes advocacy of "trade-off technology" as a technique toward improving the integration of human factors into the systems acquisition and management process. The technology is called Performance Reckoning and is different from, but very much in keeping with, the general approaches presently under way within DOD to better integrate the human into weapons systems, (cf., Promisel, Hartel, Kaplan, Marcus, & Wittenburg, 1985, and U.S. Government Printing Office, 1985 for a list of studies). The approach deals with total or systems performance as an outcome and suggests different ways to achieve this outcome by differing combinations of human and Equipment variables.

History

In a review of human factors engineering experiments, Simon (1976) concluded that the methods most commonly used were often misapplied or inadequate for obtaining the desired information. In Simon's analysis, a quantitative evaluation of the quality of the data produced in human factors engineering experiments and the methods employed to obtain these data were presented. The data were reported as distribution and "proportions-of-variance-accounted-for" by experimental factors in 239 experiments. His discovery was that Equipment factors accounted for less variance than Subject and other factors like Practice, at least when Subject and Practice factors were seriously interpreted. But as the number of factors in an experiment was increased, increasing proportions of variances became attributable to Equipment features.

Experiments at the Navy's Visual Technology Research Simulator have followed Simon's holistic methodologies and have generally supported this projection. In these studies, although the amount-of-variance-accounted-for by Equipment features is not a large proportion of total experimental variance, they would be higher if the worst combination of Equipment features ever resulted in an "unflyable" simulation. Similarly, even though the Subject variables and Practice variables are also restricted in range, they appear to account for larger proportions of variance. In fact, in one experiment in which ten simulator Equipment factors, including major cost variables, motion, and field of view were tested, all of the Equipment factors

combined accounted for less variance than the reliable pilot differences of highly experienced fleet pilots (Westra et al., 1982). A literature review was therefore undertaken in order to determine whether such findings are generalizable beyond our simulator work and beyond the time-frame used in Simon's review.

Purpose

It was our intent in Phase I to perform a literature review and subsequent quantitative analysis to determine whether there were sufficient lawful relationships available from the literature so that an Isoperformance model could be formulated. Specifically, we proposed to determine the relative contributions of Practice effects, Individual Differences, and Equipment Variations on performance as were available in the human factors scientific literature. These three elements are directly related to issues in improved human factors engineering design.

SECTION III

PHASE I FINAL REPORT

Literature Review

Approach

Green and Hall (1984) report on the rapidly growing field of quantitative methods for literature reviews in the behavioral sciences. They enumerate several methods which take into account approaches to identification of dependent variables and then later independent variables. They refer to these approaches as meta-analyses after Glass (1976), but other terms (e.g., research integration, quantitative assessment of research domains) have also been employed. Examples of these methods include: simple, descriptive (e.g., box score, tally of the direction of effect) more sophisticated, descriptive (e.g., size of the effect or d prime [Swets, Tanner, & Birdsall, 1961]), and more inferential (e.g., eta squared, omega squared [Hays, 1977]). Green and Hall (1984) point out that independent variables may be used to rate the quality of individual studies. For example, in some studies, the age of the study, the nature of the sample, the type of analysis done, the quality of the study, and the refereed journal it appears in, may all be used to weight the studies that are examined.

We decided to follow the Green and Hall approach and identify studies in the human factors engineering literature which examined at least two of the following variables together: Practice, Individual Differences, and Equipment features. The review included a computerized search at the University of Central Florida through the NASA-Southern Technology Applications Center (STAC) data base. The National Technical Information Service (NTIS), NASA, and human factors literature were reviewed. A list of key words to be used in the computer literature search was generated. Venn diagrams were used to structure the search and otherwise filter out the literature that was not of interest. For example, over 11,000 articles were catalogued under the subject heading "Human Factors Engineering." However, the combination of "Human Factors Engineering" and "Training/Learning" yielded 153 articles (30 of which were classified). Combining terms in this manner made the number of citations to review a much more manageable figure. The search was divided into the following subject heading classifications:

Training Devices
or
Training Simulators
or
(Human Factors Engineering) & Training Evaluation & (1980 to Present)
or
Training Analysis
or
Learning
or
Achievement
or
Education

We elected to review carefully the literature from 1980 to present because there was previous coverage of earlier material in a related review by Simon (1976). We also believed our selection would produce a large and sufficiently representative sample of the most recent literature, and the reference sections from this recent literature would provide us with relevant studies which had been published prior to 1980. The criteria used for including articles in our pool of relevant studies were:

(a) must be an empirical study with statistical description of results,

(b) must include an Equipment variable and either an Individual Difference or Training variable, or both, and

(c) must report results as an ANOVA table.

Results

We surfaced about 10,000 titles, distributed as shown in Table 1. A total of 240 abstracts were printed from the STAC search. It should be noted that many (approximately 50) abstracts identified in the search could not be printed because they were classified; too few to have otherwise influenced our conclusions. Although many of the citations came from the open literature, and some were symposia proceedings, another large category were abstracts of technical reports produced by contracting firms. The abstracts were reviewed for relevance and a list was made of articles to be acquired.

In our preliminary survey of the literature, any citation that examined Equipment features and either Individual Differences or Training Methods or both was obtained. Included in this assortment were both empirical and non-empirical or theoretical papers. This collection appears under the "Purportedly Relevant Citations" column of Table 1. It was decided that only empirical studies (i.e., those reporting experiments) would be summarized. Reviews of these articles were prepared for analysis regardless of the statistical method(s) incorporated. This group appears in the "Relevant Citations" column of Table 1. Later, our analysis was narrowed to examine only those studies that reported ANOVA tables in order that the effect size of the variables could be compared as Green and Hall (1984) require. These studies appear in the "# With Usable Info" column of Table 1.

**TABLE 1. RESULTS OF LITERATURE REVIEW IN
TERMS OF RELEVANCE TO THE PRESENT STUDY**

Source	Years Searched	Total # Citations (Approx.)	Purportedly Relevant Citations	%	Relevant Citations	# With Usable Info.
COMPUTER SEARCH						
NASA STAC	1980-1985	240	57	(.24)	4	0
JOURNALS						
Human Factors	1958-1985	3000	100	(.03)	46	26
J. Applied Psychology	1980-1985	720	3	(.00)	0	0
Ergonomics	1975-1985	860	20	(.02)	3	1
J. Experimental Psych.	1960-1970	2000	15	(.01)	9	1
SYMPOSIA PROCEEDINGS						
Symp. Aviation Psychology	1981-1985	210	5	(.02)	0	0
Human Factors Society	1980-1985	1600	40	(.03)	0	0
OPEN LITERATURE						
Browsing	1975-1985	1500	15	(.01)	6	2

We originally decided to review the human factors literature back only to 1972 and this was done through reviewing the open literature. As we proceeded with our search it appeared that starting with 1972, fewer appropriate studies surfaced, rather than the converse. Reasons for this are that around that time, journal editors began limiting the size of ANOVA tables, particularly in the Human Factors Journal (Carter, 1986, personal communication). Therefore, for the period of our originally intended search (1972-1985) we only found 20 articles dealing with Equipment variables that reported data in some fashion. However, only eight of those reported complete ANOVA tables. To further examine this issue we extended our search back to 1958 for the Human Factors Journal. Within the 10-year period from 1958 to 1968, 26 articles were found pertaining to Equipment variables, 18 of which contained complete ANOVA tables but some of these were surfaced too late to be included in our final analysis. If one were to look at the percentages of relevant articles with usable information (complete ANOVA tables) for the 10-year period from 1958 to 1968, 69% of the 26 relevant articles contained complete ANOVAs, while for the 13-year period of 1972 to 1985, of the 20 relevant articles only 40% contained complete ANOVA tables which confirms our observation of a trend in the literature to drop ANOVA tables.

The search of those years prior to 1972 implied that similar changes were made in other journal policies regarding the way the results were being reported which should limit the usefulness of meta-analyses attempted in other domains. Indeed, it is possible that other quantitative reviews should include a greater proportion of earlier years in their analysis. For example, in reviewing the years 1960 to 1970 of the Journal of Experimental Psychology it was found that there was a difference in ANOVA table inclusion across these years. In the early 1960s, the literature was abundant with ANOVA tables. The common procedure was to report source, degrees of freedom (df), mean square, F ratio, and p values, although it was also popular only to report df and F. Other methods of reporting data included means, standard deviations, variances, proportions, and graphs. These latter methods became the rule rather than the exception at the end of the decade, and in cases where analysis of variance was reported, df and F, or F, and p values became the standard way of reporting ANOVA results.

Because so few studies manipulated more than one of the three components of interest (Individual Differences, Practice, or Equipment Features), we sought to determine whether the literature of disciplines related to human factors engineering showed the same tendency. We performed a cursory frequency count for the Perception, Learning, and Personnel/Selection literature. Our expectation was that in the visual perception literature we would find more examples of studies dealing with Equipment Features, and sometimes Practice and sometimes Individual Differences; that the learning literature would predominantly cover Practice and sometimes Individual Differences, and that the personnel literature would chiefly study Individual Differences, and sometimes Equipment (loosely translated as environmental) features.

Table 2 presents a breakdown of the proportion of a small sampling of studies reported in representative journals. The studies are categorized as those which are narrative, review, or theoretical articles (non-empirical); those covering three factors, Individual Differences, Practice, and Equipment (ID/PR/EQ); those covering two factors (ID + PR; ID + EQ; or PR + EQ) and research which studied only one factor (Other).

Table 2. PROPORTION OF NON-HUMAN FACTORS ENGINEERING
ARTICLES WHICH STUDY COMBINATIONS OF INDIVIDUAL
DIFFERENCE, PRACTICE, AND EQUIPMENT FACTORS

Source of Article	Non-Empir-ical	ID/PR/EQ	ID PR	ID EQ	PR EQ	Other	Total #
Quarterly Journal of Experimental Psychology: Human Experimental, 1984	0	.03	.06	.22	.22	.47	32
Perception & Psychophysics, 1984	.07	0	0	.17	.03	.73	30
Aviation, Space & Environmental Medicine, 1981	.21	0	0	.19	.06	.55	33
Ergonomics, 1981	.13	0	0	.17	.03	.08	78
Multivariate Behavior Research, 1981	.48	0	0	0	0	.52	21
Personnel Psychology, 1981	.10	0	0	0	0	.90	10
Organizational Psychology & Human Performance, 1981	.24	0	0	.03	.03	.09	29
							233

As expected, the major variables studied in the perception literature were stimulus variations (which we categorized as Equipment Features), such as: adapting verbal test stimuli; rates of speech; dimensionality, connectedness, and structural relevance of figures; odor mixtures, luminance levels, and other contextual conditions. Individual difference variables included, for example, pain threshold tolerances and differences in visual fixation abilities. Practice or training manipulations included successive vs. individual presentation formats, exposure to differing psychophysical methods and inspection procedures, and trial-by-trial changes in attention.

The learning literature (Quarterly Journal of Experimental Psychology) also primarily covered stimulus variations (categorized as Equipment) such as: phonological similarity, position of stimulus, serial position in a list, type of script, stimulus-onset asynchrony, word frequency, stimulus quality, and other contextual conditions. Individual difference variables included between-hand differences, exposure to intravenous drugs, brain-damaged patients, dyslexia, and bilinguals. Practice or training variations included backward counting tasks, instructional warnings, interstimulus interval variation, task experience, and concurrent verbalization.

Human factors related literature (<u>Aviation, Space and Environmental Medicine</u> and <u>Ergonomics</u>) studied stimulus and Equipment variations in factors such as perceptually mislocated visual and nonvisual targets, luminance levels, contrast levels, display sizes, whole body vibrations, beta blockers, ozone exposure, work-rest schedules, time of day, level of difficulty of loading tasks, and other workload conditions. Individual differences studied included drivers' steering behaviors, age, peripheral visual acuity, physical activity levels, eye color, amount of smoking, etc. Practice or training manipulations covered long-term habituation to treadmill walking, speed-accuracy stress instructions, course instructions, and instructor pilot teaching behavior (positive or negative).

The applied psychology personnel/selection literature mostly studied Individual Differences and environmental or contextual components (classified here as Equipment or stimulus variations). Individual differences included cognitive abilities and personal characteristics, I-E locus of control, moral development, anxiety, job satisfaction level, performance level, age, racial differences, biorhythm status, attitudes, lateness behavior, and salary level. Environmental variables included reinforcement parameters, leaders' influence tactics, job situations, fear messages, performance information and task feedback, task variety, and employee participation. Practice or training variations included differing test instructions, guided memory procedures, repeated exposure to salient environments, goal level and type of incentive manipulations, and rater training and participation.

It is evident from inspection of Table 2 that there are few studies in the open literature, outside of the human factors engineering literature that we surveyed, which simultaneously vary the three variations necessary for the Isoperformance model. Table 2 shows only one study from a total of 233 (proportion = .4%) which manipulated Individual Difference, Practice, and Equipment factors simultaneously. This proportion is comparable to our finding of few triple-factor experiments in the human factors literature. From approximately 7400 citations, 10 were identified that met all the necessary criteria for analysis.

In summary, from all sources over 10,000 citations were reviewed. As shown in Table 1 from this total 30 were identified that met all the criteria. We used this set to test whether, in fact, we would be able to extract size of effect estimates from these experimental results. The next section of this report describes how we applied Omega Squared calculations to the 10 representative studies located in our literature review in order to determine the relative contributions to performance variance of the three major factors of interest (Subjects, Training, and Equipment).

Formal Analysis of Studies

Approach

Analysis of variance ordinarily results in mean squares, F ratios, and tests of significance. As Hays pointed out more than 20 years ago, and repeated in his newer edition (1977), this usage of the analysis fails to assess the magnitude of the effects under study. Significance is always a joint function of both magnitude and sample size. A larger effect may be

significant in a small sample, whereas a much smaller effect may reach exactly the same level of significance but in a much larger sample. Differential psychologists, who work constantly with the correlation coefficient, are familiar with this distinction. The correlation coefficient itself is a pure measure of magnitude. Thus, a correlation coefficient of .50 always accounts for 25% of the variance in either variable being correlated no matter what the sample size. A correlation coefficient of .50 is not significant in a sample of 10 paired scores, nor 20 paired scores. In a sample of 30 paired scores, an r of .50 is significant at the .05 level.

In recent years, experimental psychologists have been sensitized to the importance of effect magnitudes by the need to specify them in analyses of statistical power (Cohen 1977); nevertheless, experimental psychologists still do not report effect magnitudes in their studies. The last serious attempt to survey the human factors literature with respect to effect sizes was Simons' (1976). Simon wrote more than 10 years after Hays first called these questions to the attention of psychologists. Yet, virtually no one calculated effect magnitudes in their original report. It was necessary for Simon to calculate effect magnitudes from the published data. However, most of the time, insufficient information was given in the original report to allow Simon to make the calculation. Today the situation is essentially the same. With rare exceptions, effect magnitudes are not reported. Hence, they must be calculated, but usually the data reported are insufficient to permit such a calculation. Perforce, therefore, any generalizations about effect sizes in the published human factors literature are based on a small fraction of the published studies.

Several different procedures for estimating effect magnitudes are available (Fleiss, 1969). The two most important are eta squared (η)² and omega squared (ω ²). Simon (1976) used eta squared and we will use omega squared. It will, therefore, be necessary to explain both calculations and to give our reasons for using the latter.

Eta squared is the easier to explain. In a standard ANOVA table the sums of squares for the various effects are simply additive. Together they always equal the total sum of squares calculated by differencing each data point from the grand mean, squaring, and summing. In a simple repeated measures design (Subjects x trials of Practice), for example, there are three independent sources of variation: Subjects, trials, and the interaction between Subjects and trials. The sum of squares for these three sources exactly equals the total sums of squares calculated as indicated above. Eta squared for any given source is obtained by dividing the sum of squares for that source by the total sum of squares. Thus, eta squared for trials, that is, "the proportion of sample variance attributable to trials," equals the sum of squares for trials divided by the total sum of squares. Eta squared for any effect always varies between zero and unity. Like a Pearson product moment correlation, the closer to unity the greater the magnitude of the effect in the sample is said to be.

Omega squared depends essentially on the expected mean squares and population variance components. The model for a simple repeated-measures design may be taken as:

$$X_{ij} = \mu + \pi_i + \tau_j + \varepsilon_{ij} \qquad (i = 1, \ldots, n; \ j = 1, \ldots \kappa),$$

where the π_i (representing persons) are independent and distributed normally with mean equal to zero and variance equal to $\sigma^2\pi$, the τ_j are constants associated with trials, and the ε_{ij} (error terms) are distributed independently of the π_i and normally, mean equal to zero and variance equal to $\sigma^2\varepsilon$ (Winer, 1971, pp. 276-281). Less formally, Subjects are interpreted as a random effect and trials as a fixed effect; the interaction between them is taken as error. The degrees of freedom and expected mean squares for the three sources of variation under this model are given in Table 3.

TABLE 3. SOURCES OF VARIATION, DEGREES OF FREEDOM, AND
EXPECTED MEAN SQUARES, E(MS), FOR A SIMPLE REPEATED-
MEASURES DESIGN

Source of Variation	df	E(MS)
Subject	$n - 1$	$\kappa\sigma^2\pi + \sigma^2\varepsilon$
Trials	$\kappa - 1$	$n\sigma^2\tau + \sigma^2\varepsilon$
Subjects x Trials	$(n - 1)(\kappa - 1)$	$\sigma^2\varepsilon$
Total	$n\kappa - 1$	—

The quantity

$$\sigma^2\tau = \sum_{j=1}^{\kappa} \tau_j^2 \ / \ \kappa - 1$$

by definition. A table similar to this one can be constructed for any analysis of variance. The expected mean squares always equal some function of population components ($\sigma^2\tau, \sigma^2\pi, \sigma^2\varepsilon$, etc.) and sample sizes (κ, n, etc.).

Omega squared for any given source of variation, for example, σ^2_τ in the above design is <u>approximately</u> equal to

$$\omega^2 \simeq \frac{\sigma^2_\tau}{\sigma^2_\tau + \sigma^2_\pi + \sigma^2_\varepsilon}$$

The denominator of ω^2 always equals a sum of variance components, not necessarily all of them, and the numerator always equals some subset of the components appearing in the denominator. The interpretation is that the components appearing in the numerator account for or explain at the population

level the calculated proportion (ω^2) of the components appearing in the denominator.

Two points remain to be explained before turning to the reasons for preferring omega squared. The first point is what is meant by "at the population level." The word "population" here means what it usually means; that is, it refers to single observations rather than samples of two or more observations. Omega squared refers to proportions of variance in single observations drawn from whatever population is specified in the model. If we treat all trials as "equally likely" and then draw data points from the total population of subject-trial combinations, the variance of these single observations will be composed of all the components in the model. In this case,

$$\sigma_\tau^2 + \sigma_\pi^2 + \sigma_\varepsilon^2$$

and the proportion attributable to any subset of these components is calculable as omega squared.

The second point to be explained is the nature of the approximation used until now for omega squared. This point is closely connected to the one just discussed concerning the meaning of "at the population level." Variance components for fixed effects are defined in the manner indicated above for σ_τ^2 , that is,

$$\sigma_{2\tau} = \sum_{j=1}^{\kappa} \tau_j^2 \,/\, \kappa{-}1$$

One could, of course, define these components differently but, if one did, the expected mean squares would not take the simple forms they do in Table 3 and similar tables. Specifically, the issue at stake concerns the denominator, in this case (k-1). If omega squared is to sustain the interpretation just given, that is, as a proportion of variance components in distributions of single observations, the denominator should be k alone (Vaughan & Corballis, 1969). If the k trials are "equally likely," then the variance due to trials in distributions of single observations is

$$\sigma_{2\tau} = \sum_{j=1}^{\kappa} \tau_j^2 \,/\, \kappa$$

For random effects, such as σ_π^2 , for example, no approximation is needed; that is, one can set θ_π^2 equal to σ_π^2 directly. Omega squared then takes a slightly different but exact form, namely, for trials

$$\omega^2 = \frac{\sigma_\tau^2}{\sigma_\tau^2 + \sigma_\pi^2 + \sigma_\varepsilon^2}$$

This is exactly the same as before for random effects ($\sigma_{\hat{\pi}}$ and $\sigma_{\hat{t}}$) but slightly different for σ^2. We have now substituted

$$\sigma_{\hat{t}}^{} = \frac{\kappa - 1}{\kappa} \; \sigma_{\hat{t}}^{}$$

for $\sigma_{\hat{t}}$ in the approximate formulation (see Winer, 1971, pp. 428-430).

The reasons for preferring omega squared over eta squared are now clear. The value of omega squared does not depend on how many Subjects or trials, to take an illustrative case, a particular study happens to have, whereas eta squared does. It is arbitrary, however, whether one uses 5, 10, or 15 Subjects, 1, 2, or 10 trials. Yet eta squared depends on these arbitrary variations and omega squared does not.

Suppose, for example, that in an illustrative, subject x trials design

$$\sigma_{\hat{\epsilon}}^2 = 0$$

and

$$\sigma_{\hat{\pi}}^{} = \sigma_{\hat{t}}^{}$$

Then

$$\omega_{\hat{t}}^{} = \omega_{\hat{\pi}}^{} = 1/2,$$

no matter what n and k happen to be; not so for eta squared. Under the same assumptions,

$$\eta_{\hat{t}}^{} = \frac{n}{2n - 1}$$

and

$$\eta_{\hat{\pi}}^2 = \frac{n}{2n - 1}$$

With only one subject, all the sample variance is attributable to trials and none to Subjects; with two Subjects, two thirds is attributable to trials and one third to Subjects; with three Subjects three fifths to trials and two fifths to Subjects. In the limit, one half of the sample variance is attributable to each source and eta squared equals omega squared.

Plainly, however, it makes no sense in comparing different human factors studies to have the results depend on how many subjects the studies happened to use. If our purpose is to compare different studies or to generalize over them (which implies comparing them), then omega squared is unavoidably the index of choice.

Omega squared is defined for all sources of variance, interaction and error terms as well as main effects. In the present survey we are interested

in main effects only. We will, therefore, ignore interaction and error components and deal instead with main-effect components only. Our question, therefore, becomes: what proportions of the main-effect variance at the population level are attributable to Subjects, Practice, and Equipment? The denominator of omega squared in our analyses will consist of main-effect components only and the numerators, of course, a subset of the components in the denominators. The calculation of omega squared for a subtotal of population variance is discussed by Vaughan and Corballis (1969). The principal requirement is that the components which form the subtotal and for which ω^2 is calculated be orthogonal. The present procedure satisfies this requirement.

Results

The 10 Analyzable Studies. Table 4 presents omega squared values for Subjects, Equipment, and Training in the 10 studies where sufficient information was available to permit the omega squared calculation to be made. As mentioned above we only found 10 articles that contained complete ANOVA tables and twenty others contained AVOVA results, but had incomplete tables or were available too late in the contract period to be included in the formal analysis. Several points are immediately clear from these 10 however. First, there were no results obtained which are consistent enough for generalizations to have any meaning. The average value for Subjects, for example, is .53, but individual values in different studies range from .01 to .99 and only one study lies between .22 and .74! Five studies have values above .74 and four have values under .22. The mean value (.53), in fact, represents only one out of ten studies. There are, moreover, no obvious commonalities among studies with high (or low) omega squared for Subjects. Three of the five studies with high values involved no Equipment variation, but two did -- so the absence of an Equipment variation does not explain the high value of omega squared. A similar situation obtains among the studies with low values of omega squared. Two of the four studies involved several important Equipment Variations but the other two did not.

TABLE 4. OMEGA SQUARED VALUES FOR 10 ANALYZABLE STUDIES

Study	Omega Squared		
	Subjects	Equipment	Training
Kasprzyk et al., 1979	.82	--	.18
Shannon et al., 1982	.96	--	.04
Loo, 1978	.11	.89	--
Whitehurst, 1982	.04	.96	--
Simon, 1965	.01	.99	--
Rouse, 1979	.44	.12	.44
Goodwin, 1975	.21	.78	.01
Barsam & Simutis, 1984	.98	--	.02
Westra et al., 1982	.75	.25	--
Kellogg et al., 1984	.97	.01	negligible

It should be noted, parenthetically, that where a study involved more than one Equipment variation, the value quoted is for all of the Equipment Variations together. Similarly, where a study involves grouping subjects on an Aptitude variable, the value given is for Aptitude plus individual subject variation (within aptitudinal groups).

Second, what is true of Subjects is equally true of Equipment. The overall, the contribution of Equipment (counting only those studies, seven of them, where Equipment Variations were involved) comes to .57. Again, however, the range is from .01 to .99 and this time there are no studies at all with values between .26 and .77. The importance of Equipment relative to Subjects varies enormously.

Third, the values of omega squared for Training vary less than they do for Subjects and Equipment but still a great deal, from zero to .44. The average value, .14, represents only one study. With the exception of this one study, there are no values between .03 and .43. The value of .44, the largest by far for Training, was obtained in a study centering on a Training variation, as distinct from Practice. This fact may be related to the observation by Lane and Dunlap (1978) that papers reporting positive results are more likely to be published or to be submitted for publication than papers reporting negative results.

This result carries the clear implication that meta-analysis of the existing literature will not suffice to illustrate the isoperformance approach. If extrapolations from the existing literature to real-world situations are to be made, they are going to have to be exemplified by experiments carried out for the purpose and implemented under an innovative technical framework. Such a framework is described in summary form below, and in greater detail with experimental exemplification of the framework and application to a real-world situation is our Phase II proposal. Before turning to these matters a bit more detail concerning the literature search and its results will be provided.

D. Implications for Phase II

Our conclusions are based on a full analysis of 10 studies as well as a review of our work with flight simulator studies in one laboratory following a single paradigm plus a review of a few studies from the literature prior to 1972. We believe that the level of abstraction involved in calculations of omega squared from these published papers and studies is much too high to permit useful results. To develop and apply an Isoperformance model, that model must be stated precisely so that what constitutes a "relevant study" in the literature is much more lightly circumscribed. We have attempted to spell out how this can be done in our Phase II proposal in Appendix A.

In summary, our analysis of the literature has shown by the lack of comparability among studies that there is an enormous need to develop a formal trade-off model where variables of interest to military and scientific communities alike (Individual Capabilities, Training Methods, Equipment Variations, or variants of each) can be systematically manipulated in multifactor experimental designs. This would allow us to evaluate and

trade-off the magnitude of such effects and combinatorial effects under specific constraining conditions so that we can more accurately predict the performance level on various pieces of Equipment or tasks for any person for whom we have adequate data on personal or training history. The isoperformance model which we will test in Phase II, particularly as it is supplemented by evaluations by expert judges in the field, will greatly assist extrapolation to the military operational environment in Phase III.

SECTION IV

PERFORMANCE RECKONING - A TUTORIAL

Summary

Performance Reckoning is a way of bridging from an existing technical literature to a real-world performance system. The literature with respect to Personnel, Training, and Equipment is frequently relevant but rarely directly applicable to a real-world system. Almost always there is a gap, usually a rather large gap, between what has been studied and the real-world situation to which one wants to apply what is known. Typically, too, one cannot experiment with the real-world system itself; sometimes a high-fidelity simulator is available but more often not. Unavoidably, therefore, one must extrapolate from the scientific literature to an applied situation.

Performance Reckoning is a comprehensive approach and isoperformance is one application of the more general model having cost effectiveness as the item of interest. Performance Reckoning takes or can take all of the usual human-factors considerations into account. Psychologists, even applied psychologists, tend to be divided into non-communicating groups. Some study Personnel characteristics, others Training Methods, and still others Equipment Variations. Integrating the three kinds of results is a task typically left to the managers of military resources without much technical guidance from the human factors community. What is needed is a systematic procedure for pulling these diverse results and expert opinion together and bringing them to bear on an applied situation. In Performance Reckoning this "pulling together" is accomplished theoretically by the design of an "ideal experiment" which, while impossible to conduct, serves to specify the parameters that need to be estimated in order to arrive at a conclusion. A more applied and practical approach to performance reckoning is described below and in more detail in Appendix A. What matters here is that this approach can be applied over a very broad range of real-world problems; it can integrate diverse considerations (Personnel, Training, Equipment) and can also focus on specific problems, for example, skill retention or transfer.

Synopsis of Performance Reckoning

1.   Performance Reckoning shares several features in common with other contemporary approaches in applied psychology like HARDMAN (Mannele et al, 1985). Since the technique involves estimating in advance the likely consequences for performance of particular combinations of Personnel, Training, and Equipment, it is like a front-end analysis. Since it takes all three major components of human-factors research into account (Personnel, Training, and Equipment), it bears a resemblance to other comprehensive approaches, for example, MANPRINT (Anon, 1985). The approach makes use of expert judgment, something that is being done by other workers in DOD For example, Wing (personal communication, 1985), has utilized the judgments of personnel psychologists in predicting job performance in connection with the Army Research Institute's Project A. Most of all, however, Performance Reckoning is a special case of cost-effectiveness studies.

Any cost-effectiveness approach may be implemented in either of two ways. One way is to begin by determining possible programs or alternatives that cost the same amount (typically whatever one can afford) and then to implement the most effective of these equally costly options. The other way is to begin by determining equally effective programs or alternatives and then implementing the least expensive. Where performance is the measure of effectiveness, one starts by determining combinations of Personnel, Training, and Equipment that produce the same levels of performance (isoperformance). To date, Performance Reckoning has been applied only to this second approach, though it could be applied to the first also.

In explaining where we are at present in the development of Performance Reckoning, it is necessary to discuss the following topics:

- the use and function of an ideal experiment,

- blocking out an ideal experiment to produce an isoperformance model,

- testing the adequacy of an isoperformance model,

- the use of best evidence from the scientific literature to constrain expert judgment,

- validating individual experts,

- the use of expert judgment to finish the isoperformance model, and

- the idea of isoperformance curves and how one derives them from a finished isoperformance model.

Once this last step has been taken, it only remains to cost out those equally effective combinations of Personnel, Equipment, and Training and determine which one is least costly.

2. We take it as granted that no experimentation can be carried out in the real-world situation itself. But suppose it could. How would we design an experiment to answer the questions at issue? In the case at hand, the obvious design is for two groups of subjects, each nested within one or the other of the two equipment variations, to be given extended Practice on the real-world task and to use measured relevant Aptitudes as a covariate.

The advantage of conceiving and designing such an experiment, although it cannot be carried out, is that doing so indicates clearly what we need to know to answer our questions. For example, the needed items of information are performance as a function of Practice for each Equipment variation and performance as a function of Aptitude at each level of Practice. This statement is not complete, yet even so it is too general and admits of too many complexities to be useful in Performance Reckoning. Before expert judgment can be profitably used, the ideal experiment must be simplified or "blocked out". The number of parameters necessary to describe performance as a function of Aptitude, Practice, and Equipment must be drastically reduced.

3. Blocking out consists of imposing on the ideal experiment a series of constraints. In Appendix A, we impose three constraints:

a. Practice is divided into three segments: early, middle, and late;

b. All relations within segments must be linear;

c. No interactions are admitted within segments except Aptitude X Equipment.

In effect, this third constraint means that not only is Practice segmented into linear constraints, but so are its interactions with aptitude and equipment.

A blocked-out ideal experiment is called an isoperformance model. It consists exclusively of straight lines and not very many of them. The isoperformance model in Appendix A will consist of 12 lines, for example, performance as a function of Aptitude under either Equipment variation early in Practice or performance as a function of Practice under either Equipment variation late in Practice. These last two lines, it should be noted, will be nearly flat.

4. An isoperformance model need not, of course, capture all or even the bulk of the systematic (nonerror) variance in the behavior of a military performance system. It is our hypothesis, however, that it does. The total variance in performance can be divided into three mutually exclusive and collectively exhaustive parts:

- systematic (nonerror) variance accounted for by the isoperformance model;

- systematic (nonerror) variance not accounted for by the isoperformance model; and

- error variance, that is, interactions with individual subjects not accounted for by aptitude.

By the adequacy of an isoperformance model we mean the proportion of the systematic variance in performance it accounts for. To be acceptable, adequacy must be equal to or greater than .90.

To test this requirement one carries out a laboratory experiment having exactly the same design as the ideal experiment. The task should also bear as much resemblance as possible to the real-world performance system. A demonstration that an isoperformance model captures 90% or more of the systematic variance in a laboratory experiment does not mean that it would do so in the real-world. It does constitute a check, however; that is, isoperformance models that do not capture 90% or more of the systematic variance in laboratory situations are not used in Performance Reckoning.

5. Expert judgment is used to estimate the parameters in an isoperformance model, for example, correlations, regression coefficients, intercepts, and the like. These judgments are heavily constrained, however.

First, they must conform to the requirements of the isoperformance model. Second, they must conform to certain additional requirements derived from the scientific literature.

For example, the prediction of operational performance from Aptitude measures obtained prior to the start of Practice rarely exceeds $r = .50$ and, if at all well done, usually exceeds $r = .20$ (cf., Kennedy, Dunlap, Reschke & Calkins, 1986, for a review). In estimating such a correlation, therefore, the experts might well be required to make their estimates within .20 and .50.

Other plausible constraints may be derived from experimental studies. It may be known, for example, from simulator studies that favorable Equipment variations produce larger gains among high-Aptitude than low-Aptitude personnel. If so, the overall regression of performance on Aptitude might first be obtained by estimating correlations and then presented to the experts, who would be asked to modify those lines according to Equipment variations -- subject to the constraint that the Equipment lines diverge with increasing Aptitude.

Many possible constraints of these general sorts are possible. In applying an isoperformance model to a real-world performance system, the literature relevant to that system is searched. The outcome of such a search is not necessarily positive. One might conclude that too little was known about likely performance in the real-world situation adequately to constrain expert judgment and, therefore, that additional experiments should be performed. One might conclude that so little was known that Performance Reckoning ought not to be attempted. Much of the time, however, the use of expert judgment to finish the isoperformance model will be indicated.

6. Individual experts are credentialed in the first place by experience and subject-matter knowledge but here, just as with adequacy of the isoperformance model, it is desirable to have a check. Two such checks are possible. First, experts can be asked to estimate key parameters in the laboratory experiment used to test adequacy. Since the results of this experiment are known, the accuracy of the experts' judgments can be determined. Second, the experts can be asked to make estimations without being told about the constraints that seem warranted by the literature. The judgments of some experts will conform to those constraints, while those of others may not. The former would be better candidates for use in Performance Reckoning than the latter.

Once the experts have been selected, the isoperformance model can be finished. Each individual expert is asked to estimate relevant slopes, intercepts, correlations, etc., and their estimates averaged.

7. The final step in the process is to derive isoperformance curves from the finished isoperformance model. In the design proposed·in Appendix A, these curves take the following form:
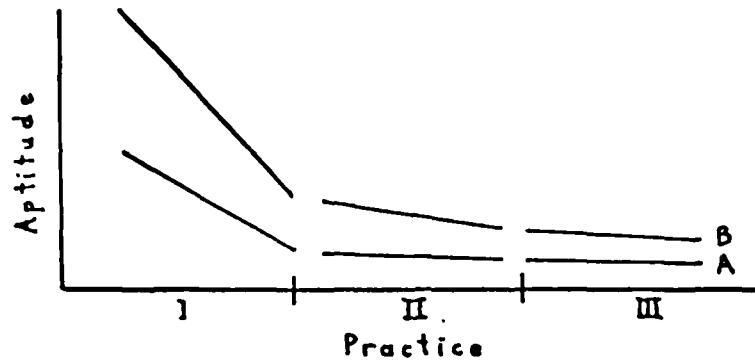
Any two points on these curves result in the same level of performance. In the case illustrated, the same performance level can be achieved with high-Aptitude people, little Practice, and the more effective Equipment variation (B) as can only be achieved with low-Aptitude people after long Practice using the less effective Equipment variation (A).

8. At this point, the analysis is complete. Combinations of Personnel, Training, and Equipment have been determined that are equally effective performance-wise; and this has been done for various levels of performance (high, medium, and low). This reckoning is not certain, of course, but it does make optimal use of expert opinion and what is known from the scientific literature.

SECTION V

REFERENCES

Alluisi, E. A. (1978). Performance measurement technology: Issues and answers. Norfolk, VA: Performance Assessment Laboratory.

Armstrong, H. G. (1943). Principles and practice of aviation medicine. Baltimore, MD: The Williams & Wilkins Company.

Army Ballistic Missile Agency. (1959). PERSHING development specification for weapon system human factor design criteria. ABMA-XPD-844, Redstone Arsenal, AL.

Army Ballistic Missile Agency (1961). Weapon systems human factors engineering criteria. ABMA-STD-434, Redstone Arsenal, AL.

Ault Committee Report. (1969). Report of the air-to-air missile system capability review. Washington, DC: COMNAVAIRSYSCOM.

Barsam, H. F., & Simutus, Z. M. (1984). Computer-based graphics for terrain visualization training. Human Factors, 26, 656-665.

Berg, C., Hertzog, C., & Hunt, E. B. (1982). Age differences in the speed of mental rotation. Developmental Psychology, 18, 95-107.

Bittner, A. C., Jr., & Moroney, W. F. (1984). Estimation of comprehensive anthropometric data from limited information (176-178). Proceedings of the 28th Annual Meeting of the Human Factors Society, San Antonio, TX.

Bittner, A. C., Jr., & Moroney, W. F. (1985). The accommodated proportion of a potential user population: Compilation and comparisons of methods for estimation. (376-380). Proceedings of the 29th Annual Meeting of the Human Factors Society, Baltimore, MD.

Boff, K. R. (Ed.). (1984). Handbook of perception and human performance. Wright Patterson AFB, OH: Air Force Aerospace Medical Research Laboratory.

Campbell, A. E. (1970). Red Baron I summary: Visual detection in air-to-air combat (Technical Note 4011-2: Doc. No. 40-1327). China Lake, CA: Naval Weapons Center.

Carter, R. C., & Biersner, R. J. (1985). Validation of job attribute requirements derived from the Position Analysis Questionnaire using military aptitude test scores. Unpublished manuscript. New Orleans, LA: Naval Biodynamics Laboratory.

Chaikin, G. (1978). Human engineering design criteria - the value of obsolete standards and guides (Technical Note 13-78). Aberdeen Proving Ground, MD: U.S. Army Human Engineering Laboratory.

Chaikin, G., & Chaillet, R. F. (1965). Engineering practice study-human engineering. Department of Defense Standardization Program Project 1410-0016, U.S. Army Missile Command Reports No. RC-S-65-1, Redstone Arsenal, AL.

Chapanis, A., Garner, W. R., & Morgan, C. T. (1949). Applied Experimental Psychology: Human Factors in Engineering Design. New York: Wiley.

Chapanis, A., Garner, W. R., Morgan, C. T., & Sanford, F. H. (1947). Lectures on men and machines: An introduction to human engineering (Report No. 166-I-19). Baltimore, MD: The Johns Hopkins University, Department of Psychology.

Cohen, J., (1977) Statistical power analysis for the behavioral sciences (revised edition). New York: Harcort Brace Jovanovich.

Cohen, J., & Cohen, P. (1975). Applied multiple regressional correlation analysis for the behavioral sciences. Hillsdale, NJ: Erlbaum Press.

Committee on Undersea Warfare. (1949). A survey report on human factors in undersea warfare (Report No. N7-ORR-291). Washington, DC: National Research Council.

Daneman, M. (1983). The measurement of reading comprehension: How not to trade construct validity for predictive power. Intelligence, 6, 331-345.

Department of Defense. (1959). Military Standard H-22174 AER. Human factors data for aircraft and missile systems (MIL-H-22174 AER). Washington, DC.

Department of Defense. (1966). Military Standard H-81444(AS). Military specifications for system analysis data (MIL-H-8144 AS). Washington, DC.

Department of Defense. (1970). Military Standard H-46855. Human engineering requirements for military systems, equipment and facilities (MIL-H-46855). Washington, DC.

Department of Defense. (1981). Military Standard 1472C. Human engineering design criteria for military systems, equipment, and facilities (MIL-STD-1472C). Washington, DC.

Department of the Air Force (1959). Human engineering criteria for aircraft, missile, and space systems ground support equipment. MIL-STD-803, Wright-Patterson Air Force Base, OH.

Department of the Air Force (1964). Human engineering design criteria for aerospace systems and equipment: Part 1. Aerospace system ground equipment. MIL-STD-803A1, Washington, DC.

Department of the Air Force (1967). Human engineering design criteria for aerospace systems and equipment: Part 3. Aerospace vehicles and vehicle equipment. MIL-STD-803A3, Washington, DC.

Fitts, P. M. (Ed.). (1951). Human engineering for an effective air investigation and traffic control system. Washington, DC: National Research Council.

Fitts, P. M. (1963). Engineering psychology. In S. Koch (Ed.), Psychology: A study of a science (Vol. 5, pp. 908-933). New York, NY: McGraw-Hill Book Company.

Fleiss, J. L. (1969). Estimating the magnitude of experimental effects. Psychological Bulletin, 72, 273-276.

Glass, G. V. (1976). Primary, secondary, and meta-analysis of research. Educational Research, 5, 3-8.

Goodwin, N. C. (1975). Cursor positioning on an electronic display using lightpen, lightgun, and keyboard for three basic tasks. Human Factors, 17, 289-295.

Green, B. F., & Hall, J. A. (1984). Quantitative methods for literature reviews. Annual Review of Psychology, 35, 37-53.

Hagman, J. D., & Rose, A. M. (1983). Retention of military tasks: A review. Human Factors, 25, 199-213.

Harbeson, M. M., Bittner, A. C., Jr., Kennedy, R. S., Carter, R. C., & Krause, M. (1983). Performance evaluation tests for environmental research (PETER): Bibliography. Perceptual and Motor Skills, 57, 283-293.

Hays, W. (1977). Statistics for psychologists. New York, NY: Holt, Rinehart, & Winston.

Hunt, E. (1984). Intelligence and mental competence. Naval Research Reviews, 37-42.

Hunt E., Davidson, J. E., & Larsman, M. (1981). Individual differences in long-term memory access. Memory and Cognition, 9(6), 599-608.

Jones, M. B. (1966). Individual differences. In E. A. Bilodeau (Ed.), Acquisition of skill. New York: Academic Press.

Jones, M. B., Kennedy, R. S., & Bittner, A. C., Jr. (1981). A video game for performance testing. American Journal of Psychology, 94, 143-152.

Jordan, N. (1963). Allocation of functions between man and machines in automated systems. Journal of Applied Psychology, 47(3), 161-165.

Kasprzyk, D. M., Drury, C. G., & Bialas, W. F. (1979). Human behaviour and performance in calculator use with Algebraic and REverse Polis Notation. Ergonomics, 22(9), 1011-1019.

Kearns, J. H. (1982). A systems approach for crew station design and evaluation (AFWAL-TR-81-3175). Wright-Patterson AFB, OH: Flight Dynamics Laboratory. (NTIS No. AD A115663)

Kellogg, R. S., Kennedy, R. S., & Woodruff, R. R. (1984). Comparison of color versus black-and-white visual displays as indicated by bombing and landing performance in the 2B35 TA-4J flight simulator (Final Report No. AFHRL-TR-84-22). Williams Air Force Base, AZ: Operations Training Division.

Kennedy, R. S., Bittner, A. C., Jr., Harbeson, M. M., & Jones, M. B. (1982). Television computer games: A "new look" in performance testing. Aviation, Space and Environmental Medicine, 53, 49-53.

Kennedy, R. S., Berbaum, K. S., Collyer, S. C., May, J. G., & Dunlap, W. P. (1983). Visual simulation requirements for aircraft aspect recognition at real world distances (NAVTRAEQUIPCEN 81-C-0105-5). Orlando, FL: Naval Training Equipment Center.

Kennedy, R. S., Dunlap, W. P., Reschke, M., & Calkins, R. (in prep.). Predictors of space adaptation syndrome. Houston, TX: NASA Johnson Spacecraft Center.

Kennedy, R.S., Wilkes, R.L., Lane, N.E., & Homick, J.L. (1985). Preliminary evaluation of a micro-based repeated measures testing system. Final report. National Aeronautics and Space Administration. Report No. 85-1. Orlando, FL: Essex Corporation.

Kyllonen, P. C., Woltz, D. J., & Lohman, D. F. (1981). Models of strategy and strategy shifting in spatial visualization performance (Technical Report No. 17, Aptitude Research Project), Stanford, CA: School of Education, Stanford University

Lane, D. M., & Dunlap, W. P. (1978). Estimating effect size: Bias resulting from the significance criterion in editorial decisions. British Journal of Mathematical and Statistical Psychology, 92, 59-70.

Lane, N. E. (1986). Skill acquisition curves: An analysis of applicability to the time course and sequencing of military training (Final Report, Contract No. 5540, Institute for Defense Analysis). Orlando, FL: Essex Corporation.

Lane, N.E., Streib, M., Glenn, F. & Wherry, R. J., (1981) The human operator simulator: An overview. In J. Moraal and K. F Kraiss (Eds.), Manned Systems Design: New Methods and Equipment. New York: Plenum Press.

Lintern, G., & Kennedy, R. S. (1984). A video game as a covariate for carrier landing research. Perceptual and Motor Skills, 58, 167-172.

Lintern, G., Nelson, B. E., Sheppard, D. J., Westra, D. P., & Kennedy, R. S. (1981). Visual Technology Research Simulator (VTRS) human performance research. Phase III (NAVTRAEQUIPCEN 78-0060-11). Westlake Village, CA: Canyon Research Group, Inc.

Loo, R. (1978). Individual differences and the perception of traffic signs. Human Factors, 20(1), 65-74.

Malone, T. B., Shenk, S. W., & Moroney, W. F. (1976). Human factors test and evaluation manual (HFTEMan, Vols. 1-3 (Technical Report No. N00123-76-C-1364). Point Mugu, CA: Pacific Missile Test Center.

Mannele, T. E., Guptill, R. V., & Risser, D. T. (1985). HARDMAN comparability analysis methodology guide (Vols. 1-5). Wilmington, MA: Dynamics Research Corporation.

Marcus, A., & Kaplan, J. D. (1984). Reverse engineering of the MI fault detection and isolation subsystem: Human factors, manpower, personnel, and training in the weapons system acquisition process (ARI Research Note 84-101). Alexandria, VA: U.S. Army Research Institute. (NTIS No. AD A142850)

McFarland, R. A. (1953). Human factors in air transportation: Occupational health and safety. New York, NY: McGraw-Hill Book Company, Inc.

Merriman, S. C., & Chatelier, P. R. (1981). Impact of the Department of Defense Science and Technology Program on training and personnel systems technology. Washington, DC: Office of Environmental and Life Sciences. (NTIS No. AD A103382)

Miller, N. E., (1947). Psychological research on pilot training (AAF AAP Report No. 8). Washington, DC: Government Printing Office.

Morgan, L. T., Cook, J. S., III, Chapanis, A., & Lund, M. W. (Eds.). (1963). Human engineering guide to equipment design. New York: McGraw-Hill Book Company, Inc.

Newell, A., & Rosenbloom, P. S. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), Cognitive skills and their acquisition. Hillsdale, NJ: Erlbaum.

Palmer, J., McLeod, C., Hunt, E., & Davidson, J. (1983). Information processing correlates of reading: (Unpublished paper), University of Washington Psychology Department.

Perkins, J. C., Binel, D. C. R., & Avery, L. W. (1983). U.S. Army Test and Evaluation Command test operations procedures. Human Factors Engineering (Part I-Test Procedures, Part II-HEDGE), TOP 1-2-610. Aberdeen, MD: U.S. Army Test and Evaluation Command.

Pew, R. W., Baron, S., Feehrer, C. E., & Miller, D. C. (1977). Critical review and analysis of performance models applicable to man-machine systems evaluation (BBN Report No. 3446). Bolling AFB, DC: Air Force Office of Scientific Research.

Promisel, D. M., Hartel, C. R., Kaplan, J. D., Marcus, A., & Whittenburg, J. A. (1985). Reverse engineering: Human factors, manpower, personnel, and training in the weapon system acquisition process (Technical Report 659). Alexandria, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.

Reid, G. B. (1982). Subjective workload assessment: A conjoint scaling approach. _Proceedings of 1982 Annual Scientific Meeting, Aerospace Medical Association_, Washington, DC: Aerospace Medical Association, 153-154.

Rose, A. M. (1978). _An information processing approach to performance assessment_ (Report No. AIR 58500-11/78-FR). Washington, DC: American Institutes for Research.

Rouse, W. B. (1979). Problem solving performance of maintenance trainees in a fault diagnosis task. _Human Factors, 21_, 195-203.

Schendel, J. D., Shields, J. L., & Katz, M. S. (1978). _Retention of motor skills: Review_. Technical Paper 313. Alexandria, VA: Army Research Institute.

Schoenfeldt, L. F. (1982). Intra-individual variation and human performance. In M. D. Dunnette and E. A. Fleishman (Eds.), _Human performance and productivity: Human capability assessment_. (Vol. 1, pp. 107-135). Hillsdale, NJ: Lawrence Erlbaum Associates, Publishers.

Shannon, R. H., Krause, M., & Irons, R. C. (1982). _Attribute requirements for a simulated flight scenario microcomputer test_ (NBDL-82R0004). New Orleans, LA: Naval Biodynamics Laboratory.

Shingledecker, C. A. (1982). Performance evaluation of the embedded secondary task technique. _Proceedings of 1982 Annual Scientific Meeting, Aero- space Medical Association_, Washington, DC: Aerospace Medical Association, 151-152.

Simon, C. W. (1965). Rapid acquisition of radar targets from moving and static displays. _Human Factors, 7_, 185-205.

Simon, C. W. (1976). _Analysis of human factors research engineering experiments_ (TR CWS-02-76). Westlake Village, CA: Canyon Research Group.

Smith, R. S., & Kennedy, R. S. (1976). _Predictor displays: A human engineering technology in search of a manual control program_ (PMTC 76-05). Point Mugu, CA: Pacific Missile Test Center.

Snow, R. E. (1980). _Aptitudes and instructional methods: Research on individual differences in learning-related processes_ (Final Report 1975-1979, Aptitude Research Project), Stanford, CA: School of Education, Stanford University.

Swain, A. D., & Wohl, J. G. (1961). _Factors affecting degree of automation in test and checkout equipment_. Stamford, CT: Dunlap & Associates.

Swets, J. A., Tanner, W. P., Jr., & Birdsall, T. G. (1961). Decision processes in perception. _Psychological Review, 68_, 301-340.

Taylor, F. V. (1963). Human engineering and psychology. In S. Koch (Ed.), _Psychology: A study of science_ (Vol. 5, pp. 831-907). New York, NY: McGraw-Hill Book Company.

Tice, J. (1986, January 27). Army program integrates personnel training and weapons design. Defense News, p. 6.

U.S. Air Force (1958). Human engineering design standards for missile system equipment. AFBM Exhibit 57-3A. Air Force Ballistic Missile Division, Inglewood, CA.

U.S. General Accounting Office. (1985). The Army can better integrate manpower, personnel, and training into the weapon systems acquisition process. Report to the Secretary of the Army, GAO/NSIAD-85-154.

Vaughan, G. M., & Corballis, M. C. (1969). Beyond tests of significance: Estimating strength of effects in scheduled ANOVA designs. Psychological Bulletin, 72, 204-213.

Wallace, J. R. (1982). The Gideon criterion: The effects of selection criteria on soldier capabilities and battle results (USAREC Research Memorandum 82-1). Fort Sheridan, Il: U.S. Army Recruiting Command.

Westra, D. P., & Lintern, G. (in press). Simulator design features for helicopter landing on small ships. I. Performance Study I. Orlando, FL: Naval Training Equipment Center.

Westra, D. P., Simon, C. W., Collyer, S. C., & Chambers, W. S. (1982). Simulator design features for carrier landing: I. Performance experiments (NAVTRAEQUIPCEN 78-C-0060-7). Orlando, FL: Naval Training Equipment Center.

Wherry, R. J., Jr. (1984). Random sampling of domain variance: A new experimental methodology (Final Task Report 1800.16). Warminster, PA: Naval Air Development Center.

Whitehurst, H. O. (1982). Screening designs used to estimate the relative effects of display factors on dial reading. Human Factors, 24(3), 301-310.

Wickens, C. D. (1984). Engineering psychology and human performance. Columbus, OH: Charles E. Merrill Publishing Company.

Williges, R. C., & Clark, C. (1973). Response surface methodology central-composite design modifications for human performance research. Human Factors, 15(4), 295-310.

Williges, R. C., & Simon, C. W. (1971). Applying response surface methodology to problems of target acquisition. Human Factors, 13, 511-519.

Williges, R. C., & Topmiller, D. A. (1980). Task III: Technology assessment of human factors engineering in the Air Force. Air Force Systems Command.

Winer, B. J. (1971). Statistical principles in experimental design (2nd ed.). New York: McGraw-Hill.

Woodson, W. (1955). <u>100 rules for improving man-machine effectiveness.</u> San Diego, CA: TeleTech Electronics Industries.

Yalow, E. (1980). <u>Individual differences in learning from verbal and figural materials</u> (Technical Report No. 12, Aptitude Research Project). Stanford, CA: Stanford University.

END

DTIC

7 — 86