MASTER

SC-R-67-1045

February 1967

Reprint

FIELD CALIBRATED SIMULATION

Alan D. Swain

RELEASED FOR ANNOUNCEMENT
IN NUCLEAR SCIENCE ABSTRACTS

SANDIA CORPORATION

PRIME CONTRACTOR TO THE U.S. ATOMIC ENERGY COMMISSION | ALBUQUERQUE, NEW MEXICO; LIVERMORE, CALIFORNIA; TONOPAH, NEVADA

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency Thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Issued by
Sandia Corporation,
a prime contractor to the
United States Atomic Energy Commission

LEGAL NOTICE

This report was prepared as an account of Government sponsored work. Neither the United States, nor the Commission, nor any person acting on behalf of the Commission:

A. Makes any warranty or representation, expressed or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or

B. Assumes any liabilities with respect to the use of, or for damages resulting from the use of any information, apparatus, method, or process disclosed in this report.

As used in the above, "person acting on behalf of the Commission" includes any employee or contractor of the Commission, or employee of such contractor, to the extent that such employee or contractor of the Commission, or employee of such contractor prepares, disseminates, or provides access to, any information pursuant to his employment or contract with the Commission, or his employment with such contractor.

GESTI PRICES

SANDIA CORPORATION REPRINT

HQ \$ 3.00, MN 65

FIELD CALIBRATED SIMULATION

Alan D. Swain
Sandia Corporation

RELEASED FOR ANNOUNCEMENT
IN NUCLEAR SCIENCE ABSTRACTS

February 1967

LEGAL NOTICE

This report was prepared as an account of Government sponsored work. Neither the United States, nor the Commission, nor any person acting on behalf of the Commission:

A. Makes any warranty or representation, expressed or implied, with respect to the accuracy, completeness, or usefulness of the information contained in this report, or that the use of any information, apparatus, method, or process disclosed in this report may not infringe privately owned rights; or

B. Assumes any liabilities with respect to the use of, or for damages resulting from the use of any information, apparatus, method, or process disclosed in this report.

As used in the above, "person acting on behalf of the Commission" includes any employee or contractor of the Commission, or employee of such contractor, to the extent that such employee or contractor of the Commission, or employee of such contractor prepares, disseminates, or provides access to, any information pursuant to his employment or contract with the Commission, or his employment with such contractor.

Abstract

A method called field calibrated simulation is suggested as one way of economically and expeditiously extending the results of large-scale field experimentation while maintaining necessary experimental control. In essence, the method consists of using field experimentation to calibrate the results of real-time and fast-time simulation. The method is described, with emphasis on real-time simulation, and some recent proposed applications are discussed.

PRESENTED AT THE HUMAN PERFORMANCE QUANTIFICATION IN SYSTEMS EFFECTIVENESS SYMPOSIUM, NAVAL MATERIAL COMMAND AND THE NATIONAL ACADEMY OF ENGINEERING, WASHINGTON, D.C., JANUARY 1967.

FIELD CALIBRATED SIMULATION

Introduction

In assessing the quantitative effects of human performance in operational man-machine systems, one of the major difficulties is systematically manipulating the many independent variables related to any criterion measures chosen while maintaining a satisfactory level of experimental control. If one turns to the experimental literature on human performance, the usual great disparities between the experimental conditions and the operational conditions of interest ordinarily permit only gross generalizations from this literature to the operational man-machine system. If one considers a field experiment, it soon becomes apparent that it is infeasible--probably impossible--to manipulate and maintain experimental control of all the relevant and important independent variables in a field setting. Yet if all the important variables are not investigated, then the experimental results are at best incomplete. Moreover, without appropriate experimental control, it is difficult to analyze the experimental results and to identify underlying parameters--steps necessary for generalization of those results to operational conditions.

One can think of all the independent variables related to a field criterion measure as constituting an n -dimensional data space. The problem is how to determine economically and expeditiously the qualitative and quantitative relationships between this complex data space and the criterion measures of interest.

One way of attacking this problem is to use both real-time and fast-time simulation and to calibrate the simulation results in large-scale experimental field studies; hence the term "field calibrated simulation." The complete set of the numerical values for the variables in an n -dimensional data space is generated by simulation. An adequate number of these values are also derived from field studies so that the entire data space may be calibrated. This field calibrated data space is then used to make predictions of operational performance of man-machine systems which are already operational, in various design stages, or even in conceptual stages only. Additionally, the field calibrated simulation technique can be used to help design more efficient field experiments with existing systems.

This paper describes how field calibrated simulation can be accomplished, with emphasis on real-time simulation. Some recent proposed applications are also described.

Some Definitions and Restrictions

The basic idea behind field calibrated simulation is, of course, not new. The general purpose of any simulation is to enable decisions to be made about some operational system to be used in a field (operational) setting. The term "field calibrated simulation" was coined to refer to a systematic way of generalizing from simulation trials to field trials. The term itself was used by Sandia Corporation in 1964 to propose to Joint Task Force Two a Department of Defense task force, a way of systematically handling the problem of studying the many variables involved in low-altitude flight. This particular application of field calibrated simulation was initiated late in 1966 and will be described later.

As already noted, simulation may be of two types: real-time or fast-time. Real-time simulation involves reproducing key aspects of the operational environment and the man-machine interface and then testing the performance of man or men under these simulated conditions. The result is quantitative performance data associated with each of a set of combinations of equipment, tasks, and environmental conditions. For purposes of this report, real-time simulation is considered to take place in static or dynamic devices specifically designed to simulate operational equipment and operating conditions with various degrees of fidelity. The fact should not be overlooked, however, that field studies used to calibrate such simulators also could be considered to represent real-time simulation to the extent that the field studies are not exact replicas of operational equipment, personnel, and contexts. It is merely convenient (and a standard practice) to restrict the term "real-time simulation" to simulator type devices.

With fast-time simulation, a mathematical or logical model is constructed from either field test or simulation data. The model is programmed for a digital computer, and parameters of the model are systematically varied as an input. The result is quantitative performance data associated with each parameter of the model. Fast-time simulation is most accurate when used to interpolate between data points which have been taken directly from field studies or from field calibrated simulator data. Fast-time simulation is most subject to gross inaccuracies when used to extrapolate beyond such data points.

Computers may or may not be used to control real-time simulation--depending upon the complexity of its data processing, reduction, and control requirements. Computers are essential for fast-time simulation.*

*There is another definition of fast-time simulation wherein a human subject is exposed (normally in a simulator) to stimuli at a greater rate or frequency than would be encountered in a real life situation. However, in this report, the term "fast-time simulation" is restricted to the use of computers as described above.

With either type of simulation it is not sufficient to test performance under selected simulation conditions without validating that performance against real field performance measured under comparable conditions. It is not always possible to be certain that all of the elements that affect human or system performance have been simulated.* For this reason, once data have been obtained in a simulation, the same conditions should be set up in the field and a selected number of measurements taken (or, the field data may be taken prior to the simulation). If there is a high correlation or agreement between the field and simulation data, one can be reasonably certain that the simulation is a valid one. If not, appropriate modifications must be made to those elements of the simulation that reduced the correlation between simulation and field performance. In short, all simulations should be field calibrated.

Summary of the Method

The basic approach to field calibrated simulation incorporates the usual steps in a systematic, analytical study whose purpose is to develop via experimentation a predictive model of behavior. One first lists the relevant variables and their likely influences on each other and the criterion measures of interest. Next one determines which variables should be held constant or left uncontrolled (e.g., weather conditions which can be treated statistically after the experiment) and which should be subject to experimental manipulation. Concurrently decisions are made on which variables (or levels of variables) should be manipulated in a field study and which in a simulator. After the field and simulator studies have been run, the simulator is then calibrated. The data space which has been generated by the calibrated real-time simulation is then used to generate the mathematical model, which is the set of equations that expresses the functional relations between the independent and dependent variables. This model in turn can be used for fast-time simulation. Finally, this model is used for prediction, and the necessary validation studies are planned and conducted to give the model true predictive (as contrasted with "postdictive") power.

At early stages in the design of a man-machine system, the above procedure will not likely be feasible. Therefore, one may consider using expert judgments to derive the interval scaling of estimates of human performance data, and then using existing field data from similar systems to calibrate the judgmental data. This technique can serve as a gross, stopgap approximation of field calibrated real-time simulation.

*This problem brings to mind a recent facetious definition of operations analysis as "a very exact mathematical modeling technique which uses large-scale computer facilities to predict that the Russians will not emplace missiles in Cuba."

The following sections of the report describe in more detail each of the above steps in the field calibrated simulation method.

List the Relevant Variables and Likely Influences

Before the relevant variables and their likely influences can be listed, one must, of course, have an adequate degree of familiarity with the man-machine system to be studied. As is usually the case in human factors studies of complex systems, the beginning is a man-machine system and task analysis. The extent to which this analysis has to be formalized and written down depends upon the degree of familiarity the analysts already have with the system.

The goal of the analysis is to set down the relevant independent and dependent variables and to get an idea of the degree to which the independent variables are correlated and their influence (singly or collectively) on the dependent variables of interest. If equipment is not a variable, then one lists the factors that determine human performance reliability (see Figure 1 and Table 1 for a representative arrangement and listing of performance shaping factors*). If equipment is a variable, then one lists the equipment factors that determine man-machine system performance reliability. These would include the human factors listed in Table 1 and an appropriate listing of equipment factors (e.g., type of aircraft).

Decide How to Treat Variables

There are a number of practical considerations in deciding which variables to hold constant, which to leave uncontrolled (and attempt statistically to sort out their influences after the experimental results are in), which to vary, and how to vary them in the field and in the simulator. Economics, timeliness, and experimental control versus operational realism are major considerations.

Using the simulator to generate the set of the numerical values for the criterion measures allows one to be highly selective in choosing those independent variables to manipulate in the field study to be used to calibrate the real-time simulation data. There is a real limit to the number of variables which can be manipulated and controlled in a field study.

*Figure 1 and Table 1 are taken from an unpublished Dunlap and Associates, Inc., report by A. D. Swain and A. P. Verdi. This report describes a human performance analysis program for a manned space system.

Figure 1. Schematic illustration of the totality of factors that determines human performance reliability. A representative listing of factors in each category is presented in Table 1. The numbers in the boxes relate the various groups of factors to the categories having the same numbers in Table 1.

Table 1. (p. 1 of 2). Representative Listing of Factors that Determine Human Performance Reliability

Performance Shaping Factors			
Extra-Individual		Intra-Individual	
(1) Instructions	(4) Psychological Stresses		
<ul style="list-style-type: none"> • Procedures required • Verbal or written communications 	<ul style="list-style-type: none"> • Task speed • Task load • High jeopardy risk • Threat of failure • Monotonous work • Long, uneventful vigilance periods • Reinforcement absent or negative • Sensory deprivation • Noise/distracting signals 		
(2) Task Characteristics	(5) Physiological Stresses	(6) Individual (Organismic) Factors	
<ul style="list-style-type: none"> • Perceptual discrimination required • Motor discrimination required • Complexity (information load) • Repetitiveness • Continuity (discrete vs. continuous) • Feedback (knowledge of results) • Human engineering factors Re: Man-machine interface design 	<ul style="list-style-type: none"> • Fatigue • Pain or discomfort • Hunger or thirst • Temperature extremes • G force extremes • Weightlessness • High G • Atmospheric pressure extremes • Oxygen Insufficiency • Vibration • Movement constriction • Lack of physical exercise 	<ul style="list-style-type: none"> • Previous training/experience • State of current practice or skill • Personality and intelligence variables • Motivation • Knowledge of required performance standards • Physical condition 	
(3) Situational Characteristics			
<ul style="list-style-type: none"> • Equipment reliability status • Weather factors visibility/turbulence • Enemy actions • Friendly actions Commander & Team Partner Activities • System organization • System social structure 			

Table 1. (p. 2 of 2). Representative Listing of Factors that Determine Human Performance Reliability

Human Performance Attributes	Human Performance Evaluation Standards	Human Performance Adequacy	Human Performance Reliability
<p>(7) Performance Dimensions</p> <ul style="list-style-type: none"> • Accuracy • Quantity • Quality • Frequency response (bandwidth) • Amplitude response • Latency • Information transmission rate 	<p>(8) Performance Standards</p> <ul style="list-style-type: none"> • Time allowed • Accuracy tolerance required • Quantity required • Quality required <ul style="list-style-type: none"> Color Switch position Target selected • Bandwidth tolerance • Information transmission rate tolerance 	<p>(9) Success</p> <ul style="list-style-type: none"> • Within tolerance • Correct quantity/quality <p>(10) Failure</p> <ul style="list-style-type: none"> • Outside tolerance • Incorrect quantity/quality 	<p>(11) Failure Rate</p> <ul style="list-style-type: none"> • Relative frequency of failures

Therefore, such variables will ordinarily be those for which operational realism is judged to be very important. For example, in studying low level target acquisition by pilots, representation of the lowest altitude of interest in a simulator only would obviously not be appropriate; there is a "gut factor" at very low altitudes which cannot be adequately incorporated into simulators. On the other hand, manipulation of the variable of level of premission target intelligence might well be adequately handled in a simulator.

However, the trade-off between experimental control and operational realism can become an issue when dealing with those who have had no formal training in the design of experiments and whose professional background and interests are primarily in operational missions. The issue can be manifested in both the simulator design and the field study design. In the former, the issue often takes shape in arguments about how dynamic and realistic the simulator need be--if indeed agreement is reached that a simulator should be a part of the experimental program. In the latter, the designer of a field experiment can be faced with a demand for realism that would effectively preclude any detailed, quantitative analysis of results and/or any generalization from the field study to other conditions of system use.

Besides the obvious greater economy of simulator trials compared with field study trials, there is another aspect of economy which increases the importance of simulator trials. In a large-scale field study there are often administrative and other difficulties which preclude any attempt to preselect subjects on such personal variables as experience level, skill level, intelligence, and so on. Quite often, the experimenter must take what is available and then try statistically to sort out the influence of certain variables after the experimental results are in. But in field studies, it is often not feasible to run sufficiently large numbers of subjects so that after-the-fact matching or other statistical adjustments of these variables can be accomplished and the appropriate statistical comparisons made. Moreover, such comparisons made after the fact often tend to lose power. Therefore, for a more efficient and powerful assessment of personal independent variables, it is often necessary to turn to simulator studies where greater control (i.e., preselection) of such variables is generally easier to accomplish.

Calibration of the Data Space Derived from Fast-Time Simulation

Calibration of Ratio Scales from Simulator Experiments

As stated already, all of the points in the n-dimensional data space are generated by calibrated simulation. Therefore, the experimental results from real-time simulation must be correlated with the results from the field studies if calibration is to be possible.

Experimental results from a simulator are used to generate a number of ratio scales* of the performance measures taken. However, neither the zero points nor the intervals between scale values established in the simulator are likely to be the same as the zero points and scale intervals established in field trials. Consider, for example, the hypothetical scale established by simulator trials in Part (a) of Figure 2. The corresponding scale as generated by field trials (but with fewer data points) might look like that shown in Part (b) of Figure 2. As shown here, there is a high correlation between the two scales. There is no crossover between the dashed lines connecting the equivalent data points on the two scales. One would feel reasonably confident in using the scale values in Figure 2(b) as key stimuli to calibrate the scale in Figure 2(a). The rescaling technique would involve generation of an equation of the form $Y = f(X)b$ (where Y is a value from the field trial scale, X is a value from the simulator scale, and b is some mathematical constant) to determine a functional relationship between common points A, E, H, K, and M from the field and simulator trials. The numerical values of B', C', D', and the other data points not present in the field trials would be derived from the equation.

If the two sets of data did not show very good correspondence, as, for example, if there were considerable crossover, one would have less confidence in the predictive power of the equation expressing the functional relationship between the two scales. But before one would give up, it would be appropriate to look very closely at the differences in conditions between the two experiments. Assume, for example, that instead of the relationship obtained in Figure 2, that in Figure 3 were obtained. Here is a crossover at the one end of the scale. What caused this crossover? How did the independent variables differ between the two experiments? Assuming that the right hand part of the scale represents lowest performance, was there some element present in the field trials that made E' more difficult than A' but which was absent or markedly reduced in the simulator trials? If so, then data point E is really not equivalent to data point E', and good correspondence should not have been expected. Of course, this kind of explanation after the fact (or, if you will, wishful thinking) maximizes experimenter error and therefore should be reverified in subsequent experimentation. Finally, if the crossovers are so numerous and inconsistent that they cannot be resolved in any behaviorally meaningful manner, the experimenter probably will have no recourse other than starting all over again.

The above hypothetical examples point to a few of the practical problems in deriving equations to express the functional relationship between field and simulator data and also suggest that these relationships should be graphically depicted (as in scatter diagrams) before the computer is used to help derive

*As used here, a ratio scale is an interval scale with a true zero point where the zero indicates an absence of the behavior trait or characteristic being considered.

Part a. Hypothetical Scale from Simulator Trials

Part b. Hypothetical Scale from Field Trials

Figure 2. Good Correspondence Between Two Hypothetical Scales Obtained from (a) Simulator Trials and (b) Field Trials.
(The capital letters stand for numerical values.)

Part a. Hypothetical Scale from Simulator Trials

Part b. Hypothetical Scale from Field Trials

Figure 3. Poor Correspondence Between Two Hypothetical Scales
Obtained from (a) Simulator Trials and (b) Field Trials.
(The capital letters stand for numerical values.)

the predictive equation (as in a regression analysis). Such graphic material will assist in the judgments required to avoid winding up a study with a set of presumed predictive equations that make no behavioral sense.

Calibration by a General Correction Factor

If it is not possible to obtain a sufficient number of data points in field trials to effect the kind of field calibrated simulation described so far, it may be necessary to derive a single correction factor (or a very small number of correction factors) to calibrate simulator results. For example, consider the problem of trying to use performance data from Keplerian trajectory experiments (such as those being conducted at Wright-Patterson Air Force Base) to get estimates of error rates for various kinds of tasks to be accomplished in a manned space system. The number of trials for a particular task in such an experiment may be too small to determine a meaningful average error rate for that task. In such a case, it may be necessary to group the results from all the tasks attempted in the zero-g environment, calculate the average error rate for the entire task group, and use that rate as the estimated average error rate for each member of that task group. Obviously, to use this figure as the basis for a calibration technique is more subject to inaccuracy than the aforementioned individual key stimuli technique; but during early research and development (R&D) stages, it may be the only technique possible.

The calibration technique in such a case would consist of the following procedure. First, determine the average task group error rate in a zero-g environment (E'_{tg}) and the average task group error rate in a 1-g environment (E_{tg}) where large numbers of trials are possible. (Instead of a 1-g environment, it might be desirable to use suspended harness studies or water studies to effect a more realistic simulation of the zero-g environment.) Next determine individual task error rates in the 1-g environment ($E_a, E_b, \dots E_n$) and multiply each of these error rates by the correction factor (bE'_{tg}/E_{tg}) to estimate individual error rates for the zero-g tasks ($E'_a, E'_b, \dots E'_n$). The ratio E'_{tg}/E_{tg} is the $f(X)$ alluded to in the previous section. The value of b would have to be determined on the basis of knowledge about human behavior. Even though one would know that $b = 1$ is not likely to be true, it might be expedient to assume so during early stages in design where information is scarce. As more information became available, however, it would be appropriate for b to take on other values depending on the type and levels of performance shaping factors (see Table 1) pertinent to any given space task. Thus, this method represents a coarse way of arriving at $Y = f(X)b$.

It would be hoped, of course, that as more and more information about task performance in space is obtained, the field calibrated simulation technique described in the preceding section could be used.

Generation and Calibration of an Interval Scale from Expert Judgments

Field calibrated simulation may start with a situation where an experiment is not yet feasible. In early R&D programs, it is often necessary to compare different design concepts before any significant body of empirical knowledge is available. Comparisons may be made on an unsystematic judgmental basis, or some system can be introduced to make the judgments more likely to lead to the best decisions. One way of approaching the problem is to obtain ratings (or rankings) from expert judges to derive an interval scale using standard psychological scaling techniques, and then derive a ratio scale via the key stimuli method. This judgmental technique serves as a gross approximation of real-time simulation using simulators.

Suppose, for example, that two design concepts for a space system must be compared with respect to the probability of reaching some criterion state. Assume (as is often the case in new systems which are not highly similar to existing systems) that there is very little error rate data that could be used to estimate the error rates associated with the various tasks subsumed by the two alternative designs. Also assume (as is the usual case) that there is no time to develop, build, and use a simulator to obtain fairly realistic error rate data on the two alternative designs.

One approach which has been used in the absence of any better alternative is to gather together a group of experts (i.e., persons knowledgeable about the tasks in question), and have them estimate absolute error rates for the pertinent human tasks and task elements. The median of these estimates is then used as the best approximation to the true mean performances for these tasks and task elements. The resulting data can then be used to conduct a quantitative human reliability analysis such as that described by Swain, 1963. Unfortunately, various investigators (e.g., Rook, 1964) who have tried this method have found that the agreement between judges is generally low.

However, Rook and other investigators have found that "Persons who will not guess at the probability of a particular event will readily rank-order several events in terms of increasing probability; and with great confidence" (Rook, 1964). Rook further noted in an informal check (ranking aircraft emergencies in terms of probability of death to pilot) that "the across-person reliability of the ranking procedure reveals a surprisingly high agreement." Irwin et al., 1964, reported a reliability coefficient of 0.92, which demonstrates high agreement among 33 raters in judging the likelihood of error of maintenance task elements. Of course, the more information the judges have about the tasks being ranked, the higher the agreement.

Blanchard et al., 1966, have verified this finding for a large number of ship-board operator and maintenance tasks. Their use and analysis of a paired comparison technique enables the investigator to derive an interval scale of

measurement of task or task element error-likeliness.* If the number of tasks or task elements is so large as to make conventional paired comparison too cumbersome (since, for example, 10 tasks require 90 comparisons**), there are shortcut methods that can be used. (See Guilford, 1954, p. 168ff., and Torgerson, 1958, p. 191ff.)

There are other techniques besides the paired comparison technique which can be used to derive an interval scale. However, as noted by Blanchard et al., 1966, "The paired comparison technique has the advantage of not forcing transitivity*** on the data--unlike ranking, rating, and sorting procedures--and permits that property of the data to be tested empirically." Thus, using the Kendall's (1948) coefficient of consistence, ζ , the number of circular triads (i.e., incidents of intransitive relations among three judged events) can be tallied for each judge. A Chi-square significance test for the coefficient of consistence for each judge can be computed to test the probability of obtaining a given value of ζ under the hypothesis that a judge's paired comparisons were made at random. (See Edwards, 1957, p. 66ff., for an easy-to-follow presentation of the technique involved.) Thus, the results from any judge showing an unacceptable degree of internal consistency can be discarded. However, if several judges demonstrate a lack of internal consistency, one might well question whether or not the events to be judged fall along the single dimension on which the judges are trying to scale them.

*This study used Case V of Thurstone's law of comparative judgment to convert the paired comparison rankings by 8 judges into mean z-scores for 11 man-machine activities. These mean z-scores are the points on the interval scale thus derived. See Edwards, 1957, p. 31ff., for an explanation of how to do the computations. The Blanchard et al. report contains perhaps the most recent thinking of several experts in paired comparison techniques. Blanchard and his co-workers contacted several of these experts, and their report constitutes a recent survey of the field.

**The number of pairs to be compared is found by $n(n-1)$ where n is the number of tasks or things to be compared. If the order of presentation is unimportant or can be balanced out by using a large number of judges, then the number of pairs equals $n(n-1)/2$. However, even dividing by 2 may still result in a cumbersome procedure.

***Transitivity in a rating or ranking method exists when the following relationship holds: If A has more of a quality than B and B has more of this quality than C; then C has less of this quality than A. Whereas this relationship makes good logical sense in the abstract, human behavior is very complex. It can make good logical sense behaviorally for you to prefer the company of A to B and B to C, but prefer the company of C to A.

The use of any rating/ranking scheme to generate an interval scale is also fraught with other methodological pitfalls. For example, it is imperative that the judges be homogeneous with respect to their "expertness" about the tasks to be ranked or rated. However, even with careful selection and indoctrination of the judges, there may be a very few who contribute most of the variability which reduces any index of rater agreement to an unacceptable level.* If so, rather than rejecting the results, one may wish to discard the results from judges who, from a statistical viewpoint, appear to be drawn from a different population of judges. The risk of discarding the results from someone who doesn't fit in with the other judges is, of course, obvious. One would want to question any such errant judge to see why he was markedly different in his judgments. If, for example, it is discovered that a judge has a different interpretation of the instructions than the other judges, one might feel reasonably safe in discarding the results from this single judge whose judgments differ markedly from the collective distribution of judgments.**

*Edwards, 1957, p. 76ff., describes a statistic developed by Kendall, 1948, which Kendall designates as u , the coefficient of agreement, to provide a means of determining the extent to which a group of judges agree in their comparative judgments. There is a Chi-square test for this coefficient of agreement which gives the probability of values of u as great as the obtained u if the comparative judgments of all of the judges were made at random.

**A search of the references cited plus Coombs, 1952, Thurstone, 1959, Remmers, 1954, and Gulliksen and Messick, 1960, and various more elementary texts failed to turn up any statistical method for discarding a judge who appears to be from a different population of judges. Let us assume that a Chi-square test has been performed for a coefficient of agreement of 10 judges and that the p value fails to reach the 0.01 level of significance. (In human factors work it is often difficult to find more than 10 experts who can be considered homogeneous with respect to the events to be pair compared.) However, in looking at the data it appears that one judge may be in large part responsible for the unacceptable significance level. (One way of detecting an errant judge would be to construct an intercorrelation matrix of each judge with the other judges showing their rank order r 's on the stimuli.) One might feel justified in eliminating the data from that judge and recalculating and testing the significance level of the coefficient of agreement of the remaining 9 judges. If the p value then reached the 0.01 level, there would be a basis for using the paired comparison data from those 9 judges. However, any such after-the-fact manipulation of the data should be reported and, as stated above, efforts should be made to interview the errant judge to see if any additional rationale can be provided for rejecting his results. As a rule of thumb, perhaps one should not attempt to eliminate on any such after-the-fact basis more than 10% of the judges. While the judge elimination procedure described above is not statistically pure, there are many instances in the human factors field when considerations of cost and timeliness force one to improvise.

If the interval scale derived from paired comparison data contained several stimulus events (known as key stimuli) whose error rates are known or which can be estimated with great confidence, these can be used to calibrate the entire scale (i.e., convert it to a ratio scale) in the manner described earlier.

As the system being developed becomes more firm, and empirical studies are possible using mockups and, later, a simulator, then the earlier calibrated scale values will be replaced with data in which one has greater confidence. Eventually, it is hoped one would arrive at the point at which a well-designed simulator was being used to generate the data points and the calibration of these points was being accomplished in realistic field experiments.

Generate Mathematical Model for Fast-Time Simulation

Even with the relatively large number of data points which can be generated by real-time simulation, there would still be a lot of empty space in the usually complex n-dimensional data space associated with a man-machine system. Therefore, equations are derived which, in effect, postdict the results from the real-time simulation. The functional relationships thus defined between the X and Y variables can be used to fill in the missing data in the n-dimensional data space.

It is beyond the scope of this paper to discuss fast-time simulation in any detail. Some precautions have already been expressed earlier. Two further observations seem pertinent.

It is imperative that specialists in behavior technology be intimately associated with the fast-time simulation studies. Such specialists can often save valuable time, especially in the required validation studies, by pinpointing empirically derived functional relationships which don't make psychological sense. It seems that some mathematical modelers lose sight of the fact that computer derived functional relationships between variables from a completed experimental study are postdictive rather than predictive, and that chance is a significant factor in these functional relationships.

A final observation is that although the primary goal of a predictive study is to predict individual system reliability (e.g., one man-machine combination) on an absolute (ratio) scale, lack of control of all relevant variables may force one to turn to a predictive model having less power. For many predictive purposes, gross prediction accuracy of the mean (average) man-machine combination may be sufficient. Even less predictive power may often be useful, as for example, being able to rank-order various man-machine combinations on a variety of, or even a single composite of, criterion variables.

Some Applications

Two applications of field calibrated simulation are described below. There is a third application dealing with a manned space system, but this information is not releasable. Suffice it to say that during the R&D and later phases of this space system, it is planned to have a complete human performance analysis program. This program emphasizes the quantitative prediction of human performance effects in the man-machine systems associated with the overall space system. Included will be all aspects of field calibrated simulation discussed in this report.

USAEPG Study

An example of field calibrated simulation was given in an as yet unpublished study of methods of testing and evaluating communications and related equipment at the U. S. Army Electronic Proving Ground (USAEPG). The emphasis in the USAEPG study was on the human factors aspects of test and evaluation, and it was recommended that this type of testing and evaluation should be expanded to place primary reliance on dynamic or near-operational performance testing of man-materiel items under controlled conditions. It was felt that testing of this type should be a routine part of the engineering tests conducted at the proving ground.

At the time of the study, much of the testing, especially human factors testing, was of the static variety, depending primarily on the use of human engineering check lists, opinion polls, and the like. As valuable as this type of information can be, it can not provide sufficient indication of how the equipment and systems under test would perform under conditions of operational use. Follow up service testing, which was intended to perform the latter function, was often performed under such poorly controlled conditions that conclusions based on the service tests often involved, in the opinion of the investigators,* unfounded generalizations from the data.

To provide the Army with a test and evaluation system with increased utility and validity, and to get around practical and administrative difficulties in using the service test as the sole vehicle for performance testing, it was recommended that the engineering tests include the necessary performance testing. Performance testing was defined as tests conducted under controlled conditions where the output performance of the total man-machine-materiel item and its human component could be quantitatively measured, analyzed statistically, and evaluated against quantitative statements of criteria.

It was recognized that it would not be feasible to conduct all of this performance testing in the field, as for example, to use soldiers in various

*A. D. Swain, H. P. Van Cott, and R. G. Pfefferkorn as consultants to the U. S. Army.

kinds of terrain and environmental conditions to test all the important independent variables related to a portable surveillance radar set. Therefore, it was recommended that the major portion of the performance testing be conducted in simulators and that carefully designed field testing be used to calibrate the simulators so that the test results obtained in them could with specified degrees of confidence be generalized to operational settings. This testing concept, called SOMATIC (SOLDIER-MATERIAL Test Instrumentation Concept), would therefore be based on field testing because such testing emphasizes operational realism. But for reasons already expressed, the concept would involve the use of real- and fast-time simulation as adjuncts to the field testing.

To implement the field calibrated simulation concept for the Army, the investigators analyzed items of materiel to be tested by USAEPG in the 1966-71 time period to determine the behavioral functions of the operators and maintainers of the materiel. These "behavioral demands" were classified into the following categories: sensing, data processing, action, and variations of each. Each behavior category was then analyzed to determine (1) how it could be measured, (2) the instrumentation required for that purpose, and (3) the computations needed to process the resulting measurement data. These requirements were regrouped into 28 Test Instrumentation Methodology Packages (TIMPs) which corresponded roughly to classes of materiel items or components of items which interface with Army personnel. For example, one of the TIMPs consisted primarily of a loudspeaker and tape playback representing "battlefield" noises. This TIMP could be used to pipe in battlefield noises as might be appropriate, for example, to evaluate a portable surveillance radar set. Other TIMPs when appropriately combined would enable the operator of such a radar set to respond to realistically sounding (over his earphones) and appearing (on the radar screen) canned targets under various environmental conditions (temperature, humidity, wind, illumination, and so on). Calibration of these simulator trials with the appropriate number and kind of field trials would then round out the test and evaluation of the portable radar surveillance set.

The Army is currently evaluating the recommendations made in the study described above. Therefore, the concept of field calibrated simulation has yet to be implemented at USAEPG.

JTF-2 Study

As mentioned at the beginning of this paper, an application of field calibrated simulation to problems of low altitude flight got underway late in 1966. This work is being done by Joint Task Force Two for the Joint Chiefs of Staff. Sandia Corporation is the technical agency of JTF-2 in its program to gather and analyze data relating to low altitude penetration by aircraft and the defense against such penetrations. Within Sandia Corporation, the technical direction of the simulator studies being used to implement the field calibrated simulation concept has been assigned to the human factors group.

The major activity in the JTF-2 effort is the conduct of large scale field experiments in which many variables can be manipulated in an environment which has high operational realism, i.e., face validity. But it was apparent that

there are so many variables affecting low altitude performance of aircraft systems and ground defensive systems that it would be impossible to manipulate all of them in a field setting. Therefore, even before the first large scale field test was conducted in 1965, plans were laid for development and fabrication of a simulator to implement the field tests.

A simulator facility was constructed which consists of a one-man stationary cockpit, a wide angle projection screen with an arc of 160° horizontal and 60° vertically to the observer, a 70 mm motion picture projector, response printout equipment, cinematic imagery, reference imagery, and cartographic materials. The subject's experience is analogous to that of an observer in the front cockpit of a two-place aircraft. The illusion of flight is very compelling; some non-pilot observers have become nauseated when experiencing the simulated treetop flights.

The first study was initiated in December 1966 and is to validate the simulation technique for the basic variables of speed and altitude. Each of 160 pilots will make two sorties (one of which is about 120 miles in length and the other about 140 miles) and attempt to acquire four well-briefed targets on each sortie. The targets, background, and target acquisition procedure are those used in the major field test during the Summer of 1966. The targets consist of two simulated AC&W Sites, two simulated SA-2 Sites, a simulated POL Dump, and two highway bridges. The initial points (IP's) (called FRP's for final reference points) for these targets consist of six forest service lookout towers, an OMNI range station, and a dam control tower. As the pilot visually acquires each FRP and target, and also when he flies over them, he pulls the trigger on the joystick. Thus, a measure of his performance is the range of acquisition. This criterion measure is the same as was used in the field test. The pilot briefing and debriefing procedures are also the same as the field test.

The major difference (hopefully the only significant difference) between the field test and this initial simulator test is that of pilot task loading. The results of the simulator test will be correlated with the results obtained in the field test. If the correlation is high for the basic variables of speed and altitude, a logical argument can be established, using the same field calibrated simulation technique, for testing more complex variables believed to affect target acquisition performance at low altitudes. Assuming that the validation test is satisfactory, three more tests using the same field targetry are planned. These tests are intended to:

- (1) Evaluate the influence of specified intelligence briefing levels on visual target acquisition performance.
- (2) Evaluate the effectiveness of two types of avionics displays (including digital and pictorial aspects) as they relate to target recognition performance.
- (3) Evaluate the effects of navigational errors and pilot orientation on acquisition performance.

In the meantime, another study has been initiated which demonstrates another advantage of the simulator--namely, the capability of the simulator to help design more efficient field experiments with existing aircraft systems. In Summer 1967 JTF-2 plans to conduct a series of field tests related to fleet-ing, mobile type targets of opportunity. In order to determine the target density and conspicuity and to evaluate an in-flight target location and re-
porting procedure appropriate for this field test, some pretests are being run in the simulator. These simulator pretests should enable JTF-2 and Sandia Corporation to have sufficient lead time to lay out and construct the appropriate field courses and targets.

As a matter of interest, it can be noted that although the field targets planned for next Summer's test are of the fleeting, mobile type, they will not be actually moving during the test. One can imagine the near impossibility of achieving identical target/background characteristics for all experimental subjects if target movement were included in the field test. Therefore, this aspect of the overall testing program is being reserved for the simulator where the same motion picture films of moving targets can be used for all subjects. However, a few field trials with moving targets will be completed to provide the necessary field calibration of the simulator data.

Summary and Conclusions

Field calibrated simulation appears to offer a useful combination of the empirical findings conducted under the greater operational realism inherent in field tests and the greater experimental control and flexibility that is afforded by real-time simulators. For complex man-machine systems, it appears to be the only practical way of achieving desired levels of understanding of the n-dimensional data space generated by the dependent and independent variables in the system. When data from field calibrated simulation is reduced to mathematical functional relationships, these relationships become the basis for fast-time simulation if necessary for the highest degrees of generalization. The concept of field calibrated simulation appears to offer great promise in studies of man-machine systems, and an extensive program of validation is underway to evaluate its predictive power as distinguished from mere postdiction.

References

Blanchard, R. E., Mitchell, M. B., and Smith, R. L., Likelihood of Accomplishment Scale for a Sample of Man-Machine Activities, Dunlap & Assoc., Inc., Western Division, Santa Monica, California, June 1966.

Coombs, C. H., A Theory of Psychological Scaling, University of Michigan Engineering Bulletin No. 34, Ann Arbor, Michigan, May 1952.

Edwards, A. L., Techniques of Attitude Scale Construction, Appleton-Century-Crofts, Inc., New York, 1957.

Guliford, J. P., Psychometric Methods, McGraw-Hill Book Co., New York, 1954.

Gulliksen, H., and Messick, S. (eds), Psychological Scaling: Theory and Applications, John Wiley & Sons, Inc., New York, 1960.

Irwin, I. A., Levitz, J. J., and Freed, A. M., Human Reliability in the Performance of Maintenance, Aerojet-General Corp., Sacramento, California, Report LRP 317/TDR-63-218, May 1964.

Kendall, M. G., Rank Correlations Methods, Charles Griffin & Co., London, 1948.

Remmers, H. H., Introduction to Opinion and Attitude Measurement, Harper & Bros., New York, 1954.

Rook, L. W., "Evaluation of System Performance from Rank-Order Data," Human Factors, 6, 1964, pp. 533-536.

Swain, A. D., A Method for Performing a Human Factors Reliability Analysis, Sandia Corporation, Albuquerque, New Mexico, Monograph SCR-685, August 1963.

Thurstone, L. L., The Measurement of Values, University of Chicago Press, Chicago, 1959.

Torgerson, W. S., Theory and Methods of Scaling, John Wiley & Sons, Inc., New York, 1958.