HUMAN FACTORS, 1992, 34(4), 387–398

Selecting Measures for Human Factors Research

BARRY H. KANTOWITZ,[1] *Battelle Seattle Research Center, Seattle, Washington*

Selecting measures is a necessary component of human factors research. Proper selection must take into account the representation problem (how is the assignment of numbers to objects or phenomena justified?) and the uniqueness problem (to what degree is this assignment unique?). Other key human factors measurement issues include subject representativeness, variable representativeness, and setting representativeness. It is difficult to create a single measure that captures essential characteristics of complex systems. Several examples illustrate how theory can guide measurement selection in such diverse human factors research as vigilance, turning off warning alarms, information requirements for military command centers, subjective workload, heart-rate signal analysis, and heat stress in nuclear power plants.

INTRODUCTION

Measurement in human factors is the antithesis of weather in daily life. As has been often noted, everyone talks about the weather, but nobody does anything about it. In human factors everyone engages in measurement, but few discuss it. Measures are the gears and cogs that make empirical research efforts run. No empirical study can be better than its selected measures. The utility of a human factors research project is linked intimately to the choice of measures.

MEASUREMENT FUNDAMENTALS

In all science, measurement is the process of assigning numbers to objects in a systematic manner. This presents two fundamental problems to the scientist (Suppes and Zinnes, 1963): (1) How is the assignment of numbers to objects or phenomena justified? and (2) To what degree is this assignment unique?

The first problem is called the *representation problem* by measurement theorists. It concerns an isomorphism between an empirical relational system and a numerical relational system. The second problem is called the *uniqueness problem*. It concerns the kind of scale transformation that is admissible. For example, only linear transformations are admissible for interval scales, whereas any strictly monotonic transformation can be applied to ordinal scales. Both problems need be considered in selecting measures appropriate for human factors research.

The constructs of reliability and validity (Elmes, Kantowitz, and Roediger, 1989) of measures have been developed to solve these

[1] Requests for reprints should be sent to Barry H. Kantowitz, Battelle Seattle Research Center, Human Affairs Research Centers, P.O. Box 5395, 4000 N.E. 41st St., Seattle, WA 98105-5428.

two problems. *Reliability* is an index of the consistency of a measure and addresses the representation problem. *Validity* is an index of the truth of a measure and is related to the uniqueness problem.

An excellent human factors example of solving the representation problem can be found in the study by Bittner, Carter, Kennedy, Harbeson, and Krause (1986) of performance evaluation tests for environmental research (PETER). They examined 114 measures for consistency in repeated sequential testing. These included measures of most popular experimental tasks that had been studied in repeated-measures applications such as aiming, choice reaction time, grammatical reasoning, letter classification, and so on. The research goal was to identify a set of measures for use in the study of environmental and other time-course effects. For example, environmental stressors aboard ship produce gross effects such as motion sickness that are well known, but a sensitive measuring device is required to assess more subtle effects on human performance. Tasks were deemed suitable for the PETER test battery if and when they showed appropriate stability over practice for means, variances, and intertrial correlations. The Spearman-Brown equation was used to calculate a temporal "reliability efficiency." Bittner et al. identified 30 measures that met their criterion for suitability for repeated-measures applications in human factors. However, 37 measures were identified as less than suitable.

The Bittner et al. study was less successful in solving the uniqueness problem and establishing construct validity for these measures. All the measures were classified into three domains: cognitive, perceptual, or motor. The basis for this classification was never explained, so one must assume that informal expert judgment was used, along with results of previous factor analytic studies. It is not clear what theory, if any, guided the classification of measures.

Even if researchers are successful in meeting high standards of reliability and validity, this does not guarantee that the human factors research will be useful. Human factors research must meet high standards of generalizability, much higher than does basic research. Especially for laboratory and simulator research, the practitioner needs some assurance that results can be extended to real-world systems. This vital issue has three major components (Elmes et al., 1989): subject representativeness, variable representativeness, and setting representativeness (often called *ecological validity*).

Subject Representativeness

The extent to which the subjects tested in research represent the population to which the results need to be applied is always a key human factors issue. Some practitioners believe that only subjects from the actual population in the system are appropriate. If, for example, one is studying safety issues in landing airplanes on aircraft carriers, then only certified naval aviators with a history of N landings are appropriate subjects. I question this and instead suggest that there are occasions when other subjects may provide useful human factors results.

For example, there appears to be little similarity between humans and garden peas, so selecting garden peas as subjects for research aimed at improving the human condition would seem inappropriate. However, Mendel was able to determine the basis of genetics—today an active field of study using the human organism—by observing garden peas. Similarly, establishing the dangers of cigarette smoking to human health can be traced back to research involving beagles. Although

early skeptics claimed that this research demonstrated only that dogs should not smoke, it is now believed that such behavior is harmful to people. If animal research has been shown to be useful and generalizable to humans, perhaps the human factors profession should be more tolerant of research using some humans to make inferences about other human populations.

This is not intended to imply that research using college sophomores is always generalizable to other populations. First, there may be differences in the age, abilities, and anthropometric qualities of two populations that make one unsuitable as a test bed for the other. For example, no sensible human factors designer would select crawlway and access hatch dimensions for middle-age navy enlisted personnel based on anthropometric data from a younger and slimmer college population. Second, there may be substantial training differences between populations that limit generalizability. However, these clear-cut differences should not blind us to cases for which populations need not be identical. For example, the visual psychophysics of college students who do not need corrective lenses is probably easily generalizable to navy pilots. Similarly, the motor control exhibited by college students in an aiming task probably can be successfully generalized to the motor control of a nuclear power plant technician who must match a wrench and a bolt. Thus fundamental human information-processing capabilities often can be usefully generalized from one population to another. However, strategies based on years of training and experience should probably be identified in the target population first. Once this has been accomplished, it seems reasonable to study them in other, more convenient populations if short periods of training are sufficient to induce similar strategies in the naive population.

Variable Representativeness

Human factors research is conducted to answer pragmatic questions. For example, we might be concerned with safety issues in glass cockpits versus traditional cockpits. One might conduct a study to evaluate this using reports in the Battelle-NASA Aviation Safety Reporting System (ASRS) data base (e.g., Kantowitz and Bittner, 1992). Would this be representative?

To answer this question, we must first distinguish between information in the data base itself and the variables used in such a research study. Even if the data base were representative, it is certainly possible by maladroit selection of variables to design a research study that is not representative of issues in aviation safety. Thus two sources of representativeness error must be considered: the data base itself and variables used in research studies that draw on the data base.

Because ASRS reports are submitted voluntarily, they are not necessarily representative of all incidents in aviation. Submitted reports are not a random sample from a population of incidents; instead, they represent a lower bound. No one really knows what percentage of actual incidents resides in the data base. Indeed, a standard warning sheet included with all ASRS data base searches cautions against interpreting these data as absolute incident frequencies that would support conclusions such as that large, four-engine, wide-bodied aircraft are safer than small, two-engine, narrow-bodied aircraft.

Variables used by Kantowitz and Bittner (1992) were drawn from two taxonomies. The first was based on previous work in aviation and covered practical, flight-specific topics such as manual control, navigation errors, and pilot situational awareness (Wiener, 1989). Although these variables are indeed representative, based on a large corpus of

prior aviation safety research, one can never demonstrate that they are 100% representative. There is always the chance that some important variable has been omitted. The second taxonomy was based on theoretical constructs in experimental psychology and included variables related to stages of mental processing such as perception, cognition, motor control, and attention (Triggs, Kantowitz, Terrill, Bittner, and Fleming, 1990).

Therefore this study, like any human factors research study, cannot guarantee 100% that its variables are truly representative. Its authors believe that the variables used are at least as representative as those used in prior research, but it would be foolish to state an exact percentage for how representative the variables are in any particular study. In the field of human factors we have two ways to establish variable representativeness. First, we depend on the practical experience of the researchers (or subject matter experts chosen by the researchers) to select representative variables. Second, we can use theory to help select representative variables because one criterion for a good theory is testability (Elmes et al., 1989). My preference is to do both in the same study, thus increasing the chances of obtaining representative variables.

Setting Representativeness

Setting representativeness, often called *ecological validity* in experimental psychology and *operational fidelity* in human factors, refers to the coherence between the test situation in which research is performed and the target situation in which research must be applied. In considering this issue, it is important to distinguish between realism and generalizability (Berkowitz and Donnerstein, 1982; Sidman, 1971). Realism is the extent to which the test situation mirrors the physical target situation. Its importance has been vastly overrated. As Sidman pointed out years ago, the realism (admittedly quite low) of a rat in a Skinner box could be improved by training the rat to stuff coins underneath the litter in its cage. We would then have the realism of a miserly rat that could be compared with a miserly human. This comparison would fail, however, because the contingencies that govern the rat's behavior are different from those controlling the human miser. In the human factors arena, great physical fidelity in a training simulator does not necessarily improve learning (Kantowitz, 1988b), though it does increase realism. Generalizability emerges from comparability of psychological processes in test and target environments, not from improvements in realism per se.

Thus the solution of studying behavior in the field rather than the laboratory will not always improve generalizability. Dipboye and Flanagan (1979) observed that field studies in industrial psychology tended to involve only a narrow range of subjects, variables, and settings. It is certainly conceivable that a laboratory study that efficiently investigated a broad range of variables could be more generalizable than a less efficient field study that had sufficient resources only to examine a narrow range of variables. Again, my preferred solution is to do both. For example, a study in progress in my laboratory uses an unrealistic low-fidelity simulator to develop a test workload protocol for heavy truck driving. Simulator results will then be evaluated in a real truck on the road. The combination of efficient laboratory simulation to narrow the large universe of test alternatives and a real truck offers the best possibility for generalizable results. The laboratory results taken alone may not be practical on the road. The road tests done without preceding laboratory investigation would allow evaluation of only a narrow set of variables. Both together should meet the pragmatic goals of human factors as well as the scientific goals

that require maximizing experimental control.

Another form of setting representativeness occurs because subjects know they are participating in an experiment and so may not behave as they would in an operational system. Experimental instructions can strongly influence human performance. For example, subjects might cheerfully follow instructions that stress speed in a simulator, whereas in an operational setting, safety considerations might require emphasis on accuracy, despite instructions from higher authority to stress speed. This conflict between nominal and actual instructions (imposed by the operational setting itself) would severely limit the generalizability of simulated results. Similarly, demand characteristics and experimenter bias (Kantowitz, Roediger, and Elmes, 1991) limit the generalizability of experimental results.

ASSOCIATED SYSTEM PERFORMANCE MEASURES

The human factors specialist is, by definition, interested in the performance of human operators as system components and wishes to know if the human improves or degrades overall system performance (Kantowitz and Sorkin, 1983). This requires the ability to measure not only the performance of the individual operator but also the performance of teams of operators and, furthermore, the performance of the total system.

It is seldom easy to obtain a measure that reflects overall system performance. A complex system, such as a nuclear power plant, has many potential indicators of system performance. Recently some of my Battelle colleagues completed an exhaustive study of performance indicators that might be related to safety in nuclear power plants (Olson, Chockie, Geisendorfer, Vallario, and Mullen, 1988). Their goal was to provide a set of indicators that would provide consistent, systematic, and objective measures of plant safety performance. Such indicators would be of great practical use by providing mechanisms for plant monitoring, setting safety performance goals, identifying where utilities should allocate resources, and assessing safety programs. Forty-six possible indicators were selected for initial study. Table 1 shows the seven categories into which these indicators fall. These indicators were carefully evaluated as to data quality, availability, and relation to safety. No single indicator was by itself an adequate measure of plant safety. Instead, the statistical combination of multiple indicators was required. Furthermore, existing measures were not optimal for predicting plant safety. The report concluded by suggesting that new data collection procedures be implemented in nuclear power plants.

A poorly constructed indicator system can be worse than no system at all. The Olson et al. (1988) study showed how difficult the measurement problem can be when one wishes to study a complex system. No simple solution was found, despite the large corpus of data reviewed. Although plant performance indicators can be helpful tools for certain problems, measurement of overall plant safety is elusive at best.

Considered as a stand-alone system, the human operator is even more complex than a nuclear power plant. Even worse from a measurement perspective is the researcher's relative inability to tap measurement points within the human system. Psychophysiological measures, though offering great potential (Kantowitz, 1987), have not yet been developed sufficiently to provide adequate guidance for detailed system specification. Most measures of the performance of individuals used in human factors are indicators of throughput that treat the operator as the traditional black box. They do not by themselves indicate what is happening within the black box. For this goal, theory is necessary to interpret the measures.

TABLE 1

Candidate Safety Measures

Management/ administration	Turnover rate % vacancies Number of administrative licensee event reports (LERs) Number of repeat violations Number of repeat human errors and equipment failures Amount of overtime worked by functional area Ratio of contractor to plant personnel Supervisory ratio
Operations	Operator exam pass/fail rate Time in limiting condition of operation (also relevant to maintenance) Operator error events (LERs, forced outages, violations) Control room instrument inoperability % continuously alarming annunciators Number of temporary procedures
Maintenance	Equipment out of service (or degraded) Safety system rework Maintenance work request status (backlog) Maintenance-related events (LERs) Preventive maintenance requests completed on safety-related equipment Number of maintenance requests issued on safety-related equipment Realignment errors during maintenance Wrong unit/wrong train events
Training/experience	Operator exam pass/fail rate Number of personnel errors Average years of licensed operator experience
Quality programs	Corrective action request backlog Quality assurance audit deficiencies
Health physics/ radiation control	Number of skin contaminations Water chemistry out of specification Work areas (or % of work areas contaminated) Collective exposure (person rems per site) Ratio of individual doses greater than 1.5 rems to the total collective dose
Configuration management	Backlog of design change requests Backlog of drawing updates

Source: Olson et al. (1988).

USING THEORY TO SELECT MEASURES

All too often in human factors the selection of a measure depends primarily on the individual experience of the human factors practitioner. Measures are not derived from formal selection algorithms; instead, they are chosen in a less systematic manner based on the wisdom of the selector and the practical constraints of the situation at hand, in which some measures are more available than others. My criticism is not that this selection technique necessarily results in poor measures; some excellent, experienced human factors practitioners do an outstanding job of selecting measures. However, until it is possible to clone such experts (perhaps with expert systems) the discipline as a whole will not progress very rapidly as long as we are dependent on the wisdom of individual practi-

tioners to make good choices. The human factors practitioner selects a measure much as the eighteenth-century Italian artisan selected wood and varnish to create a violin. Some marvelous violins were created, but all too often the secrets behind them died with the artisan. Today the science of materials analysis is helping researchers to understand what makes an excellent violin. Perhaps in the near future the secrets of Stradivarius may be available to all instrument makers. Human factors as a discipline needs the same kinds of algorithmic tools used in other sectors of science and engineering. The systematic use of theory in human factors could be such a tool for guiding the selection of measures.

Benefits of Theory

There is some misunderstanding about the role of theory in human factors. Even those who advocate theory may do so for reasons that are only partially correct. For example, Meister (1989) has stated:

> Theory is particularly necessary when the system is complex. However, the purpose of research is not to develop theories. Ideally, if one had all the data that bore on a problem, and if all the data were completely meaningful, there would be no need for theory—one would know. But hardly anyone expects to live to see this happy situation in human factors. (P. 174)

Although I agree that theory is even more useful when the system is complex, I strongly disagree that theory would be unnecessary if only we had enough data. Perhaps an analogy will make my objection clear. Data are like building blocks—bricks, nails, studs, and so on—for a house. Without these components we cannot build a house, but we also need a blueprint to assemble the components. The blueprint is analogous to theory. The world is a complicated place filled with all manner of data. An infinite supply of data would take an infinite time to figure out; it would be like trying to assemble a house by randomly putting components together until eventually a complete house of sorts emerged. Theory tells us where to look in a complex environment; it helps us to focus our resources on elements that make an important contribution to system performance.

Theory offers five substantial benefits to the practitioner faced with a real-world problem. First, as already discussed, it fills in where data are lacking. There will never be sufficient empirical data to solve all human factors problems. Theory is needed for accurate and sensible interpolation. Theory allows the human factors practitioner to predict what will happen when a variable is changed in a new way. Second, theory can yield the precise predictions required by designers and engineers before a system is built. Third, theory prevents us from reinventing the wheel. It allows us to recognize similarities across a range of practical problems. Fourth, theory can offer a normative basis for human behavior as well as system performance. It tells us how close we are to some upper limit and whether additional effort is likely to produce meaningful improvement. Fifth, theory is the best practical tool. Once an appropriate theory is available, it can be used cheaply and efficiently to aid measurement and system design.

There is a deplorable tendency for many practitioners to avoid theory because it may not seem relevant to the immediate problem at hand. Each problem is seen as an isolated issue, and practitioners who avoid theory run the risk of reinventing the wheel time and time again without realizing it. One theory, however, goes a long way. It can be applied to many different practical scenarios. Theories offer generality in that a separate theory is not needed for each and every problem. We may not even need a complex theory to set us in the right direction for starting to solve a problem (Kantowitz, 1988a; Moray, 1990). A modest theory can be a filter that narrows a large set of possible solutions to a few. The following exposition gives several examples

that show how theory can focus attention on a few measures that are most likely to yield satisfactory solutions.

Signal Detection

The theory of signal detection is well known to human factors specialists (e.g., Kantowitz and Sorkin, 1983). It has been applied to a wide variety of practical problems including medical imaging, materials testing, information retrieval, weather forecasting, and human audition (Swets, 1988). The theory distinguishes between a parameter (d') related to the sensitivity of a system and another parameter (beta) related to a decision criterion. A normative component of the theory, called the *theory of ideal observers*, specifies optimal levels of performance for particular signals.

The vigilance task (Jerison and Pickett, 1963) has long been studied in human factors because of its practical implications. A standard way to improve vigilance performance is to give knowledge of results. Recently Becker, Warm, Dember, and Hancock (1991) replicated earlier results showing that knowledge of results feedback for hits (correct detections) and false alarms (errors of commission) were far more effective than feedback for misses (errors of omission). The theory of signal detection instructed the researchers to use d' and beta as their measures. Appropriate selection of measures allowed the researchers to discover trends over time that could be easily interpreted. Furthermore, similar effects of knowledge of results were also obtained for subjective workload, an intriguing result that runs counter to the expectations of many human factors specialists. If there had been no theory of signal detection, it is unlikely that this experiment would have been performed or that its results could be so clearly interpreted.

Another important practical problem occurs when operators intentionally disable critical alarms, increasing the probability of accidents. Such incidents have occurred in railroad locomotives, airplanes, and nuclear power plants. Again the theory of signal detection can help to explain this effect (Sorkin, 1989). Using a theoretical analysis based on d' and beta, Sorkin and Woods (1985) analyzed the human-plus-alarm system as a cascaded detection system. They showed how high alarm rates produced serious system decrements. By using the appropriate measures, plus associated theory, they explained a perplexing practical problem. Furthermore, the study also provided useful solutions such as graded alarm systems that can help an overloaded operator allocate attention successfully (Sorkin, Kantowitz, and Kantowitz, 1988).

User Information Requirements

Theory can also help in a field study, where measurement is problematic. McCallum, Bittner, and Badalamente (1990) observed the frequency of communications in a military command and control brigade tactical operations center. Their goal was to understand what kind and flow of information was required for commanders to manage their forces. Using a formula suggested by optimal control theory (Sheridan and Ferrell, 1974), they were able to successfully predict communication frequency from direct estimations of rated importance and rated perishability of messages. This provided a means for establishing criticality of messaging that is useful for evaluating decisions. This example shows that using theory to select measures can be helpful in complex settings and can have practical implications.

Subjective Workload

A valuable critique of research in the area of subjective workload (Nygren, 1991) has

revealed serious measurement problems. Workload measures have been developed in such a pragmatic manner that only face validity has been considered, and vital measurement issues such as those discussed early in this article have been virtually ignored. Nygren (1991) correctly indicated that "the lack of concern for a theoretical measurement basis" has created a conflicting morass of results because different subjective workload measurement techniques often have:

> (a) been psychometrically misrepresented or inappropriate, (b) used correlations with other unvalidated workload measures as a basis for claims of predictive capability, (c) exhibited little evidence of predictive or concurrent validity despite this being a stated goal, and (d) explained virtually nothing about the hypothetical construct of mental workload. (P. 18)

By "theoretical measurement" Nygren meant psychometric theory, rather than human information-processing theory. A complete account of workload will need to take into account both types of theory. Substantive information-processing theory helps one to select what to measure, whereas psychometric theory aids in knowing how to measure.

Nygren (1991) explained how fundamental psychometric properties of the general linear model determine empirical findings in ways not generally appreciated by human factors specialists who use subjective measures of workload. Many inappropriate conclusions have been drawn because of ignorance of fundamental measurement properties and of psychometric principles. In short, this research area is a good example of how relying on artisans driven almost entirely by pragmatic goals uncontaminated by scientific restraints can create grave flaws in an entire corpus of research. Fortunately, the human factors discipline has the requisite scientific tools and theories to repair this measurement fiasco.

Vagal Tone

The need for physiological measures in human factors is indisputable, and many psychophysiological measurement techniques offer great promise for the future (Kantowitz, 1987). Heart rate is an especially promising area because the signal is more robust and hence easier to obtain in a wide variety of settings than are lower-voltage signals, such as those associated with the brain.

However, studies that measure only mean heart rate can be quite difficult to interpret. Heart rate is heavily influenced by physical activity, and even the modest requirements of pressing keys can alter it. Heart rate variance is a better measurement choice, especially for making inferences about nonphysical workload. A more sophisticated measurement technique allows computation of vagal tone (Porges, 1986), which is derived from a Fourier transform of the heart-rate signal. For example, a recent study (Backs, Ryan, and Wilson, 1991) showed that spectral analysis of the heart rate signal was related to disturbance gain in a manual tracking task for both low-frequency (0.067–0.144 Hz: Traube-Hering-Mayer) and high-frequency (0.145–0.500 Hz: vagal tone) bands.

This kind of sophisticated measurement will prove increasingly useful to the human factors profession. It is important to realize that data transformations can occur even in the simplest measurement scale. For example, the McCallum et al. (1990) military field study discussed earlier used logarithmic transformations of frequency data. Thus knowing what kind of theoretical data transformations will be useful is a large step toward selecting particular human factors measures. Such a measurement procedure offers obvious advantages over recording some variable that is convenient and then wondering later how to use or interpret it.

Case Study in Selection: Heat Stress in Nuclear Power Plants

I will take the risky option of concluding this section by discussing research in progress. By anticipating research not yet under way, I have the opportunity to allow the reader to follow my thoughts about selection and free myself of the temptation to edit these thoughts based on actual results. I have little doubt that there may be flaws in the following analysis, but these will be corrected later. For now, my emphasis is on selecting measures to solve a pragmatic problem.

The Nuclear Regulatory Commission (NRC), which is funding this research, is concerned with the effects of environmental stressors on the performance of nuclear power plant personnel. My colleagues and I have been given the task of obtaining empirical results that can illuminate this problem in ways that will help the NRC fulfill its regulatory mission. Thus this is a pragmatic—not a theoretical—problem. However, we believe that theory will help.

Our first planned step is to conduct a field survey at four nuclear power plants to measure the exposure of workers to heat and noise during an eight-hour shift. We intend also to use this field survey to classify tasks in a general way according to the human information-processing requirements they require. Our information-processing general model (Figure 1) has been used previously to classify psychological demands on the job (Kantowitz and Bittner, 1992; Kantowitz et al., 1989), as well as in the laboratory (Triggs et al., 1990).

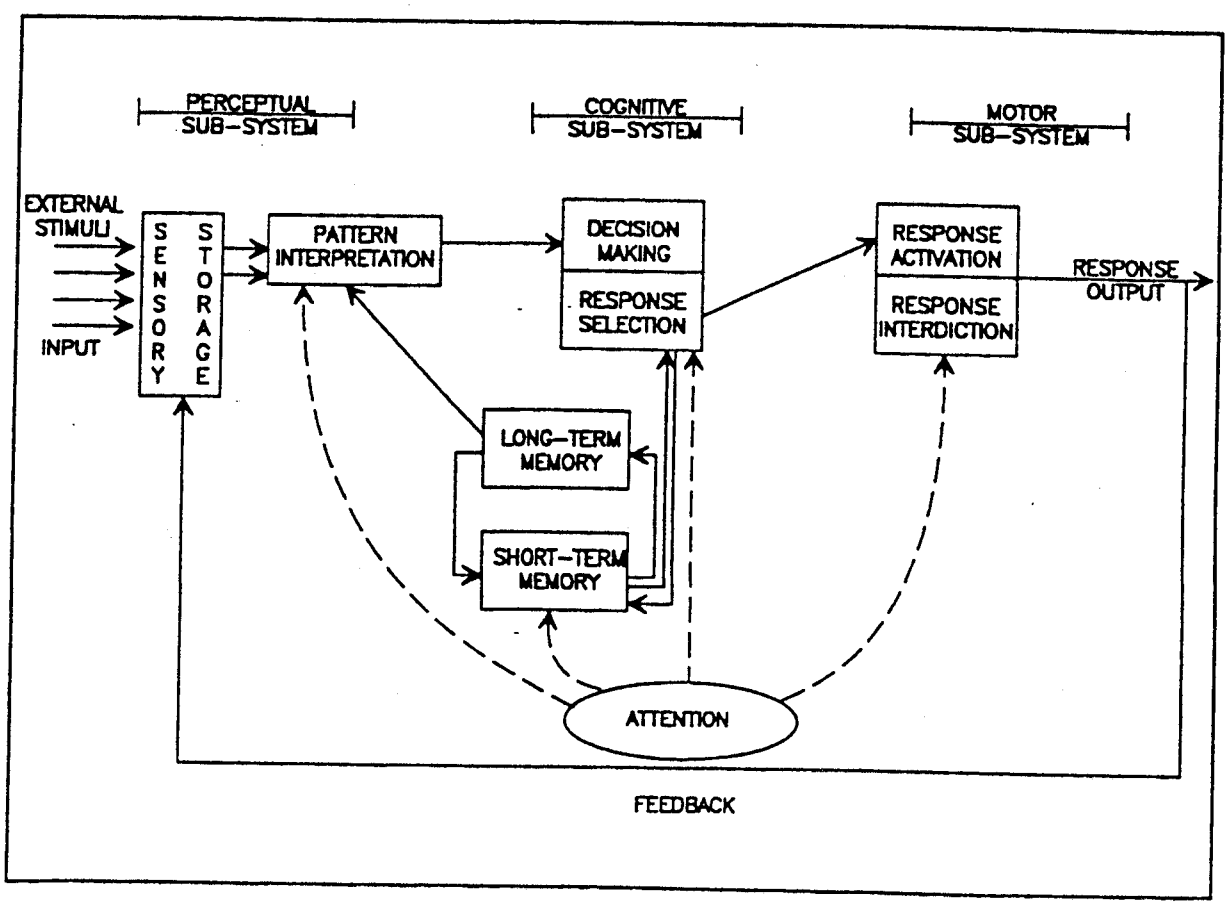

Figure 1. *Human information-processing general model.*

The second phase of research will be conducted in an environmental chamber at the Battelle Human Performance Laboratory. Reviews of prior research on heat stress (Echeverria, Barnes, and Bittner, 1991) revealed only a few studies that tested subjects for a full eight-hour shift or that evaluated cognition well. So we will measure performance for a full shift to enhance the generalizability of our findings and use a model to examine cognition. What should we measure during these eight hours of testing?

We will use our field results to decide which information-processing stages in Figure 1 are most vulnerable to heat or noise in a nuclear power plant. We will then select laboratory tasks that have a theoretical information-processing basis for being related to those stages. For example, we already know that maintenance tasks require use of the motor control stage. Data on movement control goes back to the previous century (e.g., Woodworth, 1899). Therefore, we will choose a movement control task that is generally understood, such as an aiming task (e.g., Kantowitz, 1991), and then measure reaction time, movement time, and movement accuracy. A similar procedure will identify appropriate tasks and measures that have a theoretical basis for perception, cognition, and attention stages of information processing as required. Because this research will occur in a laboratory, we have considerable freedom to select measures that meet our criteria. Our goal is to discover empirically how exposure to heat and noise during an eight-hour shift affects the stages of mental processing in Figure 1.

Another, more traditional, approach might extract part-tasks from the nuclear power plant and test in the laboratory without regard to theory. We believe, however, that this might lead to stopgap and possibly incorrect results arising from unknown measurement problems with real tasks (Lane, Kennedy, and Jones, 1986). Instead, our approach couples theory with empirical tasks likely to be affected by heat and noise. For example, once we know how heat stress affects attention, we no longer need to extract every nuclear power plant task into the laboratory. We can make useful predictions based on our knowledge of the attention stage. Of course, once the laboratory work is completed, the validity of our results must be tested against the real world, where we will meet the problems expressed in Table 1 about selecting appropriate measures as safety criteria. It is likely that multiple measures will be required. However, if we have done our job properly in the laboratory and used theory well, our task should be much simplified.

CONCLUSION

Selecting measures for human factors research should not be done by rote or because some measures are easy to obtain and appear to satisfy pragmatic goals. Instead, one must ask how well the candidate measure solves both the representation problem and the uniqueness problem. Subject representativeness, variable representativeness, and setting representativeness all are important and must be evaluated, preferably before human factors research begins.

Theory is the best practical tool. A useful theory need not be unduly complex. Selection of measures for human factors research should be guided by theory whenever an adequate theory can be brought to bear on the practical problem that engenders the research.

ACKNOWLEDGMENTS

I am grateful for the sage comments of my colleagues Alvah Bittner, Robert Sorkin, and Tom Triggs on a more primitive antecedent of the present article.

REFERENCES

Backs, R. W., Ryan, A. M., and Wilson, G. F. (1991). Cardiorespiratory measures of workload during continuous

manual performance. In *Proceedings of the Human Factors Society 35th Annual Meeting* (pp. 1495–1499). Santa Monica, CA: Human Factors Society.

Becker, A. B., Warm, J. S., Dember, W. N., and Hancock, P. A. (1991). Effects of feedback on perceived workload in vigilance performance. In *Proceedings of the Human Factors Society 35th Annual Meeting* (pp. 1491–1494). Santa Monica, CA: Human Factors Society.

Berkowitz, L., and Donnerstein, E. (1982). External validity is more than skin deep. *American Psychologist, 37,* 245–257.

Bittner, A. C., Carter, R. C., Kennedy, R. S., Harbeson, M. M., and Krause, M. (1986). Performance evaluation tests for environmental research (PETER): Evaluation of 114 measures. *Perceptual and Motor Skills, 63,* 683–708.

Dipboye, R. L., and Flanagan, M. F. (1979). Research settings in industrial and organizational psychology. *American Psychologist, 34,* 141–150.

Echeverria, D., Barnes, V. E., and Bittner, A. C. (1991). The impact of environmental exposures on industrial performance of tasks. In W. Karwowski and J. W. Yates (Eds.), *Advances in industrial ergonomics and safety III* (pp. 629–636). Bristol, PA: Taylor & Francis.

Elmes, D. G., Kantowitz, B. H., and Roediger, H. L. (1989). *Research methods in psychology* (3rd ed.). St. Paul, MN: West Publishing.

Jerison, H. J., and Pickett, R. M. (1963). Vigilance: A review and reevaluation. *Human Factors, 5,* 211–238.

Kantowitz, B. H. (1987). Premises and promises of psychophysiology. *Contemporary Psychology, 32,* 1002–1004.

Kantowitz, B. H. (1988a). Defining and measuring pilot mental workload. In J. R. Comstock, Jr. (Ed.), *Mental-state estimation 1987* (pp. 179–188). Hampton, VA: National Aeronautics and Space Administration, Scientific and Technical Information Division.

Kantowitz, B. H. (1988b). Laboratory simulation of maintenance activity. In *Proceedings of the 1988 IEEE 4th Conference on Human Factors and Nuclear Power Plants* (pp. 403–409). New York: IEEE.

Kantowitz, B. H. (1991). Effects of response symmetry upon bimanual rapid aiming. In *Proceedings of the Human Factors Society 35th Annual Meeting* (pp. 1541–1545). Santa Monica, CA: Human Factors Society.

Kantowitz, B. H., and Bittner, A. C. (1992). Using the aviation safety reporting system database as a human factors research tool. In *Proceedings of the 15th Annual Aerospace & Defense Division Conference* (pp. 31–39). Norcross, GA: Institute of Industrial Engineers.

Kantowitz, B. H., Martin, R. L., Triggs, T., Geisendorfer, C., Terrill, B., Morgenstern, M., and Bittner, A. C. (1989). *A total model of human cognition and action for the nuclear power industry in Japan* (Prepared for Institute of Human Factors Nuclear Power Engineering Test Center, BHARC-700/89/004). Seattle, WA: Battelle Seattle Research Center/HARC.

Kantowitz, B. H., Roediger, H. L., and Elmes, D. (1991). *Experimental psychology* (4th ed.). St. Paul, MN: West Publishing.

Kantowitz, B. H., and Sorkin, R. D. (1983). *Human factors: Understanding people-system relationships.* New York: Wiley.

Lane, N. E., Kennedy, R. S., and Jones, M. B. (1986). Overcoming the unreliability of operational measures: The use of surrogate measure systems. In *Proceedings of the Human Factors Society 30th Annual Meeting* (pp. 1398–1402). Santa Monica, CA: Human Factors Society.

McCallum, M. C., Bittner, A. C., and Badalamente, R. V. (1990). Empirical identification of user information requirements in command and control evaluation. In *Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 1200–1203). Santa Monica, CA: Human Factors Society.

Meister, D. (1989). *Conceptual aspects of human factors.* Baltimore, MD: Johns Hopkins University Press.

Moray, N. (1990). Designing for transportation safety in the light of perception, attention, and mental models. *Ergonomics, 33,* 1201–1213.

Nygren, T. E. (1991). Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload. *Human Factors, 33,* 17–34.

Olson, J., Chockie, A. D., Geisendorfer, C. L., Vallario, R. W., and Mullen, M. F. (1988). *Development of programmatic performance indicators* (NUREG/CR-5241, PNL-6680, BHARC-700/88/022). Washington, DC: U.S. Nuclear Regulatory Commission.

Porges, S. W. (1986). Respiratory sinus arrythmia: Physiological basis, quantitative methods, and clinical implications. In P. Grossman, K. H. L. Janssen, and D. Vaitl (Eds.), *Cardiorespiratory and cardiosomatic psychophysiology.* New York: Plenum.

Sheridan, T. B., and Ferrell, W. R. (1974). *Man-machine systems: Information, control, and decision models of human performance.* Cambridge, MA: MIT Press.

Sidman, M. (1971). *Tactics of scientific research.* New York: Basic Books.

Sorkin, R. D. (1989). Why are people turning off our alarms? *Human Factors Society Bulletin, 32*(4), 3–4.

Sorkin, R. D., Kantowitz, B. H., and Kantowitz, S. C. (1988). Likelihood alarm displays. *Human Factors, 30,* 445–459.

Sorkin, R. D., and Woods, D. D. (1985). Systems with human monitors: A signal detection analysis. *Human-Computer Interaction, 1,* 49–75.

Suppes, P., and Zinnes, J. L. (1963). Basic measurement theory. *Handbook of Mathematical Psychology, 1,* 3–76.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science, 240,* 1285–1293.

Triggs, T. J., Kantowitz, B. H., Terrill, B. S., Bittner, A. C., and Fleming, T. F. (1990). The playback method of protocol analysis applied to a rapid aiming task. In *Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 1275–1279). Santa Monica, CA: Human Factors Society.

Wiener, E. L. (1989). *Human factors of advanced technology ("glass cockpit") transport aircraft* (NASA Contract NCC2-377). Moffett Field, CA: National Aeronautics Space Administration.

Woodworth, R. S. (1899). The accuracy of voluntary movement. *Psychological Review Monograph, 3,* 1–114.