Flach, J. M, and Bennett, K. B. (1996). A theoretical framework for representational design. In R. Parasuraman and M. Mouloua (Eds.), Automation and Human Performance: Theory and Application (pp. 65-87). Mahwah, N.J.: Lawrence Erlbaum Associates.

Methodological Issues for Evaluation of Interfaces: A Case for Representative Designs

John M. Flach and Kevin B. Bennett

Wright State University
Dayton, OH

In complex systems, problems often arise that were not anticipated by the designers. These problems go beyond the rule based capacity of automated systems. These problems require creative, "productive" thinking ---- an achievement that has yet to be captured fully in automated systems. A human, armed with an appropriate representation, is often uniquely capable of this achievement. The catch, and the challenge for interface design, is to provide the appropriate representation. How do we evaluate the appropriateness of a representation?

The importance of representation has been a recurrent theme in the problem solving literature. The following quote from Wertheimer (1959) summarizes the Gestalt position on the requirement that a representation (i.e., envisioning) reflect the "structural truths" of a problem.

> Thinking consists in envisaging, realizing structural features and structural requirements; proceeding in accordance with, and determined by, these requirements; . . . that operations be viewed and treated in their structural place, role, dynamic meaning, including realization of the changes which this involves; realizing structural transposability, structural hierarchy, and separating structurally peripheral from fundamental features . . . looking for structural rather than piecemeal truth.
>
> from "Productive Thinking", (p. 235 -236).

In this chapter, we will address methodological issues for evaluating interfaces as representations to support productive problem solving. The chapter begins with a discussion of our theoretical framework. This theoretical framework will provide a context for understanding the need for representative designs to evaluate interfaces and productive problem solving.

A Theoretical Framework

In this section we will present our theoretical framework for evaluating interfaces. Of course, it is not possible for us to do this without bias. The best we can do is to state clearly that this is our position and to caution the reader to evaluate these ideas critically. Also, it is important to note that in explicating constructs we are dealing with first principles. These are intuitive commitments that are not open to empirical falsification. Lachman, Lachman, and Butterfield (1979) refer to the shared intuitive commitments

of a group of scientists within a common paradigm as the conventional rules of science. No matter how rigorous a scientific approach, it is impossible to escape from the requirement to bootstrap the empirical process on a set of intuitive commitments or conventional rules (e.g., commitments about the fundamental nature of causality, of time, of information, of meaning, the dimensionality of space, etc.).

Perhaps the logical place to start our discussion of the theoretical basis for evaluating interfaces is with a definition of interface. The *interface is the medium between the human and a work or task environment.* In human-machine systems, the medium typically includes displays and controls. However, it is almost never the case that the medium is limited to displays and controls.[1] In addition to the artifactual displays and controls, the interface includes the natural sources of information (e.g., optical flow fields) and natural constraints on action (e.g., biodynamic constraints). It also includes the operators' memory and knowledge base (i.e., mental model or internal representation). The role of operator expertise in determining the functionality of the medium was emphasized by Hutchins, Hollan, and Norman (1986). They note that the the facility and flexibility of a tool such as a text editor depends critically on the skill of the user (e.g., even a non-direct interface can appear to be direct when the user has extensive experience). Human operators have adapted and learned to perform fluently using some very poorly designed interfaces.

The medium reflects the *interactions* among the physical constraints on action, the information constraints on perception, and the value constraints (e.g., goals and cost functionals) that explicitly or implicitly define a task or work domain. These interacting constraints are illustrated in Figure 1 in a way that emphasizes the belief that these interacting constraints are central to the problem of coordinated or situated action.

Although most of the current diagrams of the human information processing system include a feedback loop, these diagrams typically show stimulus and response as distinct entities --- the stimulus entering from the left and the response leaving from the right. This creates the impression that the relation between stimulus and response is remote or arbitrary. Thus, in the laboratory the stimulus is manipulated as an independent variable and the response is measured as a dependent variable. The feedback loop is generally cut so that the experimenter can maintain strict control over the independent variable, unconfounded by actions of the experimental subject.

Dewey (1896/1972) challenged this approach. He felt that this approach resulted from a failure to appreciate the implication of the closed-loop structure of the system. He argued that "the reflex arc idea, as commonly employed, is defective in that it assumes sensory stimulus and motor response as distinct psychical existences, while in reality they are always inside a co-ordination and have their significance purely from the part played in maintaining and reconstituting the co-ordination" (p. 99). In Figure 1 we have attempted to illustrate the closed-loop system in a way that emphasiz-

---

1. It is not uncommon that the terms display and interface are used interchangeable. We will use the term display to refer to an artifactual representation that is often one local component within the interface.

es the intimate coupling between stimulus and response within the co-ordination. It should be noted that there is no explicit distinction between human and environment in Figure 1. Value and goal constraints may be internalized by the operator or may be explicit design constraints for the system. Action constraints may arise from the biodynamics of an operator's body, from the controls, or from the plant or vehicle being controlled. Perception constraints reflect the perceptual systems of the operator, the physical sensors of the system and the associated displays, and natural displays such as flow fields. The medium too, represents both the physical representation as well as the internal representation. Within this diagram there is no distinction between human and environment. The human and environment are integrated throughout the diagram.

The word *interactions*, as used in our definition of interface, is highlighted to emphasize an assumption that will be fundamental to our approach to the evaluation of interfaces in the context of functional systems. This is the assumption that the constraints on action, information, and value are not independent. In fact, we will make a much stronger claim --- *meaning arises out of the interactions among these various sources of constraint.* Still stronger, we claim that *the measure of an interface is the ability of the human agent to make contact with meaning* , where contact with meaning is reflected in 'structural understanding' and 'coordinated action' appropriate to the value system for a specific work space.

This focus on meaning as a relational property within the medium is in contrast to a more traditional view in which meaning is considered to be a product of information processing. To paraphrase Mace (1977), whereas the traditional information processing view places meaning inside the head, our position considers meaning to result from what the head is inside of.

Figure 1. The closed-loop human-machine-environment system. This diagram is intended to emphasize the medium as the center of a coordinated perception-action system.

The traditional information processing view is linked to Shannon and Weaver's (1963) statistical notion of information. Information statistics (e.g., bits/sec) are excellent for addressing issues of channel capacity but fail to address the issue of correspondence with an external reality. That is, information statistics cannot distinguish between being precisely right (always saying yes when yes is the correct answer) and being precisely wrong (always saying yes when no is the correct answer). Thus, information statistics do not provide a very effective framework for addressing issues of semantics. There is no basis for addressing the meaning of a message. While the basic research community has recognized some of the limitations of the information statistic as a performance parameter (e.g., Lachman, Lachman, & Butterfield, 1979), vestiges of this approach remain in terms of the communications channel metaphor. Research has focused on properties of the channel (e.g., capacity, parallel vs. serial processing, locating the bottleneck) and has neglected the semantics of the message that is being communicated over this channel. The result is a tendency to focus on channel capacity or bandwidth as the critical issue for display design and a failure to address the problem of meaning. Our view, in contrast, emphasizes meaning or correspondence as a fundamental issue and channel capacity is only a secondary consideration. Therefore, *our theoretical position might be characterized as a meaning processing approach, rather than an information processing approach*. In this approach meaning is not the product of processing, but rather, it is the raw material from which coordinated, productive, or adaptive behaviors are molded. Haken (1988) presents a somewhat similar position:

> The concept of information is a rather subtle one . . . . information is linked not only with channel capacity or with orders given from a central controller to individual parts of a system --- it can acquire also the role of "medium" to whose existence the individual parts of a system contribute and from which they obtain specific information on how to behave in a coherent, cooperative fashion. At this level, semantics may come in.
>
> from "Information and Self-Organization" (p. 23)

It is important to note that there is ample evidence that issues of meaning and bandwidth are not independent. Miller's (1956) famous article summarizes the position that the capacity of working memory depends on the ability to organize or chunk information into meaningful units. A logical implication of Miller's analysis is that the dimensions of meaning which determine the organization of information into chunks cannot be ignored when predicting processing capacity. In fact, we argue that issues of meaning must take precedence in evaluation of the functional bandwidth of the operator when interacting with a work domain through an interface. In particular, the bandwidth of an interface will be greatly effected by the experience and knowledge of the operator and by the organization (i.e., in Gestalt terms the deep structure) of the interface.

The statistical notion of information together with the concept of independent processing stages fitted well within the zeitgeist of a linear world view that has until recently been a dominant perspective of science. However, science is gaining a new respect for the power and importance of nonlin-

ear systems. In nonlinear systems, interaction is the rule and independence is rare. We believe that human-machine systems are nonlinear and that this fact must be reflected in our research programs.

The claim that the human-machine system is a nonlinear system, where interactions are the rule, has important implications for the distinction between syntax and semantics. The term syntax will be used to refer to the form, appearance, or surface structure of an interface (e.g., alphanumeric versus graphical, object versus bar graph, integral versus separable). In the terms of Hutchins, Hollan, and Norman (1986) syntax relates to the construct of articulatory distance. That is, it has to do with the physical form or structure of the interface vocabulary.

The term semantics will be used to refer to the meaning, or deep structure of an interface. That is, how it maps on to distinctions within the problem or work space. This relates to what Hutchins et al. refer to as semantic distance. However, there may be subtle differences between our notion of semantics and the ideas of Hutchins et al. In discussing, semantic distance Hutchins et al. tend to ask questions such as --- "does the language support the user's conception of the task domain? does it encode the concepts and distinctions in the domain in the same way that the user thinks about them?" Such questions imply that the benchmark for semantic distance is the user's "mental model." For us, the benchmark should reflect a normative analysis of the work domain constraints --- what could or what should the user be thinking about?

These contrasting relations between the mental model (or internal representation) and display (external representation) have been discussed by Wilson and Rutherford (1989):

> When applied within systems design -- process control, HCI, or other -- two positions seem to be taken. One is that the displays of a process or system must be compatible with operators' internal representations of the system; the other is that the displays themselves ought to determine that certain mental models be built up ... (p. 628).

Thus, one perspective is that representation aids should be designed to *match* the mental models that expert users have developed. In this design approach experts' understandings and conceptualizations are studied and then translated into representation aids. The second design approach maintains that representation aids should be designed to *determine* the user's mental model. In this approach, the domain itself is analyzed and the results are used to guide the development of representation aids. We advocate this second approach. However, in very complex domains, knowledge elicitation will often be a critical aspect of the domain analysis and the mental models of domain experts may provide the best available window to the real domain constraints. However, the ultimate design should be framed in terms of an understanding of the ecological constraints. The goal is to provide a representation that allows the cognitive agent to fully utilize the opportunities that the ecology affords. Performance should not be constrained by an under specified mental model. This is a lofty, and perhaps, unattainable goal. It may be very difficult or impossible to fully understand all the implications for the many interacting variables within a complex work domain. However, focusing on mental models does not necessarily reduce the complexity (the expert's mental model will reflect the requisite variety, i.e. full complexity,

of the work domain). Further, as scientists, we generally will have greater control and will be capable of more direct analysis of the work domain, than of the mental model. The challenge of representational design is to open up and broaden the perspective of the cognitive agent so that all the possibilities within the workspace are accessible.

In sum, semantics raises the question of what information to present - what distinctions are important for the operator. Syntax raises the question of how to present that information. Syntax and semantics are conceptually distinct degrees of freedom for interface design. The same information mapping (semantics) can be accomplished in many different forms (syntax). However, in operation, the two dimensions will generally become intimately coupled. The medium literally becomes the message.

This blending of semantics and syntax is implicit in Rasmussen's (1986) observation that operators interact with processes at multiple levels --- knowledge-based, rule-based, and skilled-based interactions. The distinction between syntax and semantics holds only for knowledged-based interactions in which the interface functions as a symbol representing features of the external world that the operator must interpret based on a conceptual understanding or "mental model." However, in the course of interaction within a work domain, the relation between operator and interface often evolves so that rule- and signal-based interactions emerge. At this stage of skill, surface structures of the display will directly trigger productions and actions. In an important sense, the interface becomes the world. For skill- and rule- based interactions the operator does not explicitly attend beyond the interface. The mapping to the domain beyond the interface is implicit in the actions and rules that have evolved and survived due to past successes. The capacity for skill- and rule-based interactions emerge with experience and training. These modes will dominate for experienced operators under "normal" operating conditions. However, novel events or system faults will often require a shift to knowledge-based interactions. Because the interface will be required to function at all three levels, semantics and syntax become intimately bound together within any particular interface. The result is that it will be very difficult, if not impossible to unconfound syntax and semantics when evaluating displays. Prescriptively, the implication of this confounding is that the syntactical structure should be designed to reflect the semantic constraints (i.e., deep structure) of the problem space. This prescription is central to an approach that has been variously referred to as ecological interface design (Rasmussen & Vicente, 1989; Vicente, 1991; and Vicente & Rasmussen, 1990), representational aiding (Woods, 1991), or the semantic mapping principle (Bennett & Flach, 1992).

To summarize our position, the claim is that meaning is a central construct when evaluating interfaces. Meaning permeates the problem of display design. It is an emergent property of the interactions between action, information, and values. Meaning reflects constraints arising from the work space, constraints on action (in AI terms the application of operators), and the value or cost functionals by which actions and solutions are scored. Within displays, semantics (meaning) and syntax (form) are intimately coupled. A representation will be effective to the extent that form (syntax) reflects function (i.e., meaning or semantics). Finally, it is humans' capacity to pick-up meaning, to think productively, to achieve a structural understanding, that make them such valuable components in complex systems.

Representative Design

Internal validity refers to the soundness of inference, the logical consistency of a methodology. The focus on internal validity is a defining attribute that differentiates scientific reasoning from other forms of knowing and argument. The primary threats to internal validity are confounds. Concern for internal validity is a primary motive behind experimental reductionism. That is, an approach in which the phenomenon of interest is parsed so that each dimension can be examined in isolation from other confounding components. This approach can be a very successful approach to the extent that systems are linear. That is, to the extent that performance of the whole is the sum of the components. However, if the system is nonlinear, if there are emergent properties, then parsing must be done with great care.

It is difficult, if not impossible to escape reductionism in experimental approaches. In order to have control, in order to avoid confounds, in order to reduce problem complexity to a manageable level, experiments are generally reduced versions of the phenomenon of interest. The issue, with respect to evaluating interfaces, and with respect to experimental science in general, is not whether or not to use reductionist methodologies. The issue is - what are the most effective ways to parse the problem so that the phenomenon is simplified, but not broken? That is, our methodologies must simplify in a way that preserves critical emergent properties.

Hammond (1993) contrasts two approaches to the parsing problem --- Wundt's choice and Brunswik's choice. Wundt's choice was based on the assumption that the deep structure of basic causal relationships was obscured by the surface features of the environment. Thus, Wundt argued that "by experiment . . . we strip the phenomenon of all its accessory conditions, which we can change at will and measure" (Wundt quoted by Hammond, p. 206 - 207). Wundt's choice evolved into the traditional information processing approach in which the deep structure was characterized in terms of a series of information processing stages that are assumed to function relatively independently. Thus, these stages provide a natural partitioning by which cognitive processing can be reduced for experimental evaluation.

Brunswik, however, believed that understanding the environment was fundamental to the problem of cognition. As Hammond (1993) observed, Brunswik believed that the "irregular, uncertain, confusing environment is the environment of interest, not the sanitized environment of the psychophysics laboratory or the perception laboratory of illusions and other 'impoverished stimulus' conditions" (p. 208). Hammond continues, "Brunswik's choice led to a design that includes a *formal* representation of all those conditions toward which a generalization is intended; representative design thus refers to the logical requirement of representing in the experiment, or study, the conditions toward which the results are intended to generalize' (p. 208). Brunswik called his methodological approach representative design.

The important implication of representative design for experimental methodologies is that the researcher must look to structural properties of the task environment in order to make decisions about how to partition the phenomenon of interest so that controlled experimentation is possible. For

Brunswik these structural properties were modeled in terms of the Lens model. The Lens model predicts the "achievement" of the organism as a relation between "ecological validity" and "cue utilization." Ecological validity refers to the probabilistic mapping between the environment and the medium of perception and cue utilization refers to the integration over the medium necessary to make an inference about the corresponding environment. For Brunswik, ecological validity provided both a guide for experimental design and a normative limit for achievement. The normative characteristic of ecological validity is clear as Brunswik (1956) writes that "one of the most important aspects of functional theory concerns the relationship between ecological validity and utilization. *Ideally, cues should be utilized in accordance with their validity*" (p. 141, emphasis added).

Gibson (1979) also considered ecological validity to be central to the problem of perception:

> First, the environment must be described, since what there is to be perceived has to be stipulated before one can even talk about perceiving it. Second, the information available for perception in an illuminated medium must be described. This is not just light for stimulating receptors but the information in the light that can activate the system. . . . Third, (and only here do we come to what is called psychology proper), the process of perception must be described. This is not the processing of sensory inputs, however, but the extracting of invariants from the stimulus flux.
>
> Gibson (1979, p.2)

For Gibson, the mapping from the medium to the objects of the world was lawful not probabilistic. The existence of lawful or invariant structural relationships in the optic array is a central premise of the concept of direct perception.

The trend toward more representative, naturalistic, ecological approaches to cognition can also be seen in memory research. Early memory research was dominated by Wundt's choice, as typified in Ebbinghaus's research program that treated meaning and knowledge as nuisance variables to be controlled out. However, the current trend is toward more naturalistic, context-rich approaches as typified by Bartlett's (1932) and more recently Neisser's (1976; Neisser & Winograd, 1988) work, that considers meaning and knowledge to be critical variables for remembering. Bahrick and Karis (1982) note that "scientists have become more aware of their obligations to be responsive to the problems of society, and the lack of ecological relevance in most memory research has become a matter of explicit concern" (p. 427).

The importance of considering the environment when parsing the problems of cognition was also found to be important for early research in artificial intelligence. Simon (1981) illustrated this in his parable of the ant. In this parable, he pointed out that the structure of the beach (i.e., the problem space) over which the ant locomotes provides critical information for modeling the ant's trajectory. Thus, research in artificial intelligence typically begins with a formal specification of the *state space* or *problem space*. The state space shows the critical dimensions of a problem and the possible paths from the initial condition to the goal. Like the concept of ecological validity, the state space provides a framework for bringing normative considerations to research on problem solving. For example, it allows us to compare the solution paths of humans to the "shortest" path through the problem space. Such normative considerations, if possible, could also be im-

portant for evaluating interfaces. Although AI researchers have taken great care to study the task constraints, they have had the luxury to choose their tasks. Some might argue that tasks such as cryptoarithmatic, missionaries and cannibals, and the tower of hanoi are not representative of the kinds of tasks that operators face when trying to manage complex socio-technical systems such as a chemical processing plant.

Pew (1994) also makes a strong case for the need to consider the environment when evaluating human-machine systems and when designing displays to support effective situation awareness (SA) :

> The SA requirements are the essential elements of information and knowledge needed to cope with each unique situation. Since virtually all measurements in human factors are relative, we argue that measuring SA implies having a standard, a set of SA requirements, if you will, against which to compare human performance. Such a standard must encompass an **abstract ideal,** a **physically realizable ideal**, and a **practically realizable ideal**. The abstract ideal includes the full set of information and knowledge that would make a contribution to accomplishing a particular goal. This is an abstract ideal because it is unconstrained by the design of the crew station and the information that is actually available to the crew member. Definition of the physically realizable ideal introduces the constraint of a real crew station. It is the information and knowledge that a crew member could obtain, given the current information sources in the workplace, that is, the current suite of displays and controls. It places no constraints on the information processing capacities and limitations of the crew member. Finally, we think in terms of a practically realizable ideal, what any real individual might be able to achieve under the best of circumstances, taking into account typical human performance capacities and limitations. It sets the standard against which to evaluate how well an individual performed given the system he or she had to work with. The definition of the abstract ideal helps us to understand what might be accomplished with better design and implementation.
>
> Pew (1994, p. 2)

What do these calls for representative experimental design imply for research to evaluate interfaces? We believe the implication is that the interface should not be dissected from the functional work domain. The dissection of interface from its natural work domain destroys the interactions from which meaning emerges within the medium. Thus, it becomes impossible to evaluate the interface in terms of structural truth or in terms of normative models of what the operator ought to know or ought to do given the constraints of a particular problem or work space. This does not mean that we must give up reductionism. The implication is that experimenters must consider the medium in terms of the interactions among task environment, action, and perception. The problem must be parsed in light of those interactions so that functional meaning is preserved in the laboratory. The laboratory environment will always be different from the actual work environment, however, the laboratory environment should be structured so that functional meanings are represented as fully as possible.

A Few Representative Cases

In this section, we will review a few examples from the display literature that illustrate the interplay of semantics and syntax and that highlight the importance for parsing problems within the framework of the task semantics.

Sanderson, Flach, Buttigieg & Casey (1989). Earlier research by Wickens and his colleagues (Barnett & Wickens, 1988; Carswell & Wickens, 1987; Casey & Wickens, 1986; Wickens, 1986; Wickens, Kramer, Barnett, Carswell, Fracker, Goettl, & Harwood, 1985) had suggested that object displays (e.g., a triangle) resulted in more effective information integration than separated displays (e.g., bar graphs). However, Sanderson et al. questioned whether the superior performance of object displays was due to the surface feature of "objectness" or whether it might reflect differences in the deep structure of the representations. That is, differences in the mapping of the representation to the underlying task structure. The task structure was to evaluate the state of a process where two inputs combined (averaged) to produce an output. The subjects were to detect a deviation from the normal scaling of input to output. Sanderson, et al. noted that the triangle (object) display had a higher order feature (the apex angle) that mapped directly to the scaling of input to output. They noted that "in all normal transformations of the triangle for the present system, the angle of the apex varies only between about 85 and 95 deg. Any deviation outside this range immediately signals that the process is in an abnormal state" (p. 185). No such higher order property was present in the bar graph displays used by Wickens. However, it was possible to create such a higher order property by simply rearranging the bar graph so that the output was positioned between the two inputs on a common baseline. The result is that the tops of the bar graph would be collinear when the scaling was normal and would deviate from collinearly when a failure occurred. Results showed that the performance advantage for the object display was reversed and superior integration of information was achieved with the bar graph display.

The early research by Wickens et al. had confounded semantic and syntactic dimensions of the interface. The conclusion that "objectness" supports more effective information integration ignored the semantic dimension. The Sanderson et al. study is a good illustration of how semantic considerations (the mapping of display constraints to process constraints) supersede syntactic considerations (object versus separated forms). They conclude that "whether a display is best described as an object or separated display, the crucial determinant of its effectiveness will be how well the significant states (e.g., normal vs. failed) of the system it represents are mapped onto changes in its emergent, or configural, properties" (p. 197).

MacGregor & Slovic (1986). Whereas the studies discussed in the previous section evaluated an abstract process (the averaging of two inputs to create an output), MacGregor and Slovic assessed performance in a more naturalistic task of predicting the times for runners to complete a marathon (real data for actual runners were used to predict time in an actual race). Four variables were displayed in four different graphical formats. The formats included a bar graph display, a deviation bargraph display, a spoke display (quadrangle object display), and a face display. The variables displayed included the runner's age, the total number of miles run in the 2 months prior to the marathon, the fastest time in a 10K race, and the runners self-rating of motivation level. The interesting result of the MacGregor and Slovic study was that the face display resulted in either clearly superior performance or

performance equivalent to the worst of the other formats depending on the mapping of variables to features in the face. When the most diagnostic variable (fastest 10K time) was mapped to the most salient feature of the face (mouth), then performance with the face display was superior to all other displays. When the most diagnostic variable was mapped to a feature with lower salience (height of eyebrows) performance was no better than the worst of the other formats.

Again, as with the Sanderson et al. study, the MacGregor and Slovic study illustrates that semantics (mapping to the domain) supersedes syntax (surface features of the graphic) in determining performance. However, the MacGregor and Slovic study further illustrates how a form such as a face has its own "semantics." A happy face has its own meaning. This semantics affects the salience of elements in the display. Success of the display depends on the mapping of salience to the diagnosticity or importance of domain variables relative to the task or decision being supported. Highly diagnostic variables should be mapped to salient features. MacGregor and Slovic note that "as integral display designs become more complex and pictorial, the potential for incompatibility between the normative importance of information features and the psychological salience of display features becomes greater, requiring thorough understanding of how display features are perceived and the quality of attention they are given" (p. 198). In the case of face displays and other pictorial formats the interface functions as a metaphor. Thus, issues of the structural properties of the mapping from the metaphorical to target domains become critical concerns (e.g., Gentner, 1983; Gentner & Gentner, 1983). The range of performance, from best to worst, found with the face display clearly shows how critical the mapping can be. This emphasizes the danger of drawing inferences from effects due to syntax without considering interactions with domain semantics.

Vicente (1992). Semantic considerations also have important implications for the performance measures that are used to evaluate interfaces. A performance index such as fault diagnosis, in which a display is scored in terms of the operator's ability to identify changes in system state with respect to the design goals for a process, is rich in semantics. However, a performance index in terms of retrospective memory in which a display is scored in terms of the operator's ability to recall the state of all (or a subset of) system variables can be a semantically impoverished measure, particularly if the variables that are to be recalled are chosen arbitrarily.

Vicente (1992) used both diagnosis performance and retrospective memory performance to evaluate two displays in the context of a thermo-hydraulic process simulation. One display was a mimic display that showed the physical variables (pump states, flow rates, reservoir levels, heater settings, temperatures, etc.) and a second display was constructed to show additional higher order process constraints such as the mass and energy balances. The results showed that the addition of the functional information in the second display resulted in clear improvements in diagnosis. However, performance in the retrospective memory probe task was not consistent. When memory performance for all variables was evaluated there was no clear advantage for either display --- "memory was better for whichever display was experienced second" (p. 368). When the performances for physical and functional variables were evaluated independently, the results again showed no relation between diagnosis and recall of physical variables. However, there were significant correlations between diagnosis and the recall of functional vari-

ables. The functional variables reflected constraints on the process and were relevant to diagnosing the status of the process.

Moray, Jones, Rasmussen, Lee, Vicente, Brock & Djemil (1993). Moray et al. also compared diagnosis performance with retrospective memory in evaluating three display formats for the control of feedwater in a nuclear power plant. The three formats were a single-sensor-single-indicator analog display format, a similar analog format with an added animated pressure-temperature graphic, and an animated graphical display in which all the variables were integrated within a space defined by a Rankine cycle diagram (this is a temperature/entropy state diagram for a heat engine). Three performance tasks were evaluated. A quantitative retrospective memory recall task required the operators to specify the values for 35 system variables. A qualitative retrospective memory recall task required the operators to answer 21 yes/no questions about the state of the process (e.g., Did the reactor trip during the trial?; Did the hotwell level exceed 90% full?; Did the generator function normally?). Finally, a diagnosis task required the subjects to specify whether a fault had occurred and to describe the fault. Results showed a clear advantage for the Rankine cycle display in the diagnosis task, but the picture for memory task performance was less clear. There was a disassociation between performance in the diagnosis task and the memory tasks. The authors conclude that "performance on the quantitative recall tests did not correlate with diagnostic performance to a significantly useful extent. Performance on the qualitative recall test correlated more strongly, but not at a level which is sufficient to make it a reliable indirect performance indicator to evaluate displays. The evidence from this project suggests that diagnostic performance itself is the best of the three ways to rank the quality of displays" (p. 56).

The Vicente and Moray et al. studies illustrate, again, the importance of semantic considerations. In evaluating displays, the meter for performance should reflect the semantic constraints of the domain. Semantically neutral tasks, such as global tests of recall for all system variables, will not provide valid indices for scaling the merits of a display.

Roth, Bennett & Woods (1987). The previous examples have focused on graphical displays. However, there are many other types of representations that can be effective for supporting human problem solving. Roth et al. evaluated an expert system designed to assist technicians in trouble shooting a new generation electro-mechanical transport system. There are several aspects of this experiment that make it a prototypical example of semantic-based interface evaluation. First, the expert system was an actual system that was in its final stages of development; one that was deemed ready for placement in the field. Second, the experiments were conducted with the actual device and with actual technicians. Third, the six problems that were developed for the experiment varied on semantically relevant dimensions --- whether or not knowledge resulting from previous experience with the old technology was relevant to problem solution and the number of competing hypotheses that needed to be ruled out. For each problem the appropriate fault was placed in the device, the technicians were provided with a brief description (much like the trouble report they would normally receive), and were asked to interact with the expert system to diagnose and repair the problem. Each problem session was video taped, a computer log was recorded, and notes were taken by the experimenters.

The evaluation was semantic-based. The interaction of the technicians and the expert system were observed in the course of solving actual domain problems. A "canonical" solution path was defined that specified the optimal series of expert requests and corresponding technician responses required to reach a correct solution. Deviations from the canonical or most efficient solution path were analyzed for the cause of deviation and for the types of knowledge that were applied by the technicians to bring the problem solving episode back on track. The results indicated that, despite the expert system support, deviations were the rule (observed in 78% of the cases). Because of the traditional "question and answer" format for the expert system much of the knowledge was hidden within the "expert system." Thus, success depended largely on the knowledge and skills of the operator. Roth et al. conclude that "successful performance depended on the ability of the technician to apply knowledge of the structure and function of the device and sensible trouble-shooting approaches. This ability was necessary to follow the underspecified instructions, to infer machine intentions, to resolve impasses and to recover from errors (person or machine) that led the machine expert off track" (p. 491).

Summary. The four cases presented here were chosen to illustrate the intimate coupling of semantics and syntax and the dangers that result when experiments or theories treat these as independent dimensions. The first two studies (Sanderson et al. and MacGreger and Slovic) illustrate that the form of a display (i.e., object, integral, separable, configural) can not be evaluated independently from the structural properties (i.e., deep structure, ecological validity) of the problem being represented. It is the mapping of structure in the representation to the structure of the problem that ultimately determines performance. Bennett and Flach (1992) have articulated this in terms of the "semantic mapping principle" and Woods (1991) has articulated this in terms of principles for analogical integration.

The second two studies (Vicente and Moray et al.) illustrate that the ruler against which a display is measured (the performance index) must also reflect the semantic constraints of the target domain. It is important not to be seduced by the convenience and apparent generality of generic indices such as performance in a retrospective memory task or other generic task batteries that have only mundane links to the semantics of the task domain. The experimental task should be meaningful with respect to the role of the interface in the target work domain.

The last case (Roth et al.) also demonstrates the importance of semantically based norms (i.e. cannonical solutions) for performance measurement. These norms suggest the dimensions that are relevant to task semantics and provide benchmarks against which performance can be scaled.

Another vestige of the information processing approach is a tendency for research to fixate on reaction time (RT) as the critical index of performance. There is the implicit assumption that minimum RT (i.e., maximum information processing rate) is the ideal. Roth et al. illustrates the use of more semantically rich measures of performance.

With regard to time as a performance index, "timing" may be more important than absolute time. While much has been written on speed-accuracy tradeoffs, the implicit assumption has been that speed and accuracy are

independent performance dimensions (that of course may trade-off). This approach fails to acknowledge the possibility for accuracy "in time." This can be most readily seen in music. The quality does not improve if the music is speeded up (even though all the notes are still played accurately). There is an appropriate or optimal rhythm. The construct of "timing" emphasizes the importance of synchrony or coordination with a process (i.e., making the right response at the right time --- neither too slow or too fast). Synchrony is a relational construct. That is, synchrony is a measure of the match between temporal constraints in the task domain and temporal constraints on the operators' actions. Synchrony cannot be found in either the operator or the process. It is yet another factor that illustrates the need for constructs that integrate over human and environment.

The other reason to include the Roth et al. study was to emphasize that semantics are not simply a concern for graphical interfaces, but are a concern for any medium of representation intended to support the operator to adapt to the demands of a particular work domain, Graphical displays and expert systems are different in form, but as representations they serve a similar function --- to support the operator in managing complex task domains.

Another important factor that interacts with the domain semantics and that has important implications for experimental methodologies is expertise. Both Vicente and Moray et al. manipulated expertise as an independent variable. This is clearly a critical dimension. However, we will not elaborate on this dimension here, as there is ample awareness and discussion of the implications of sampling for the validity and generality of experimental research. Despite this awareness, there has been far too much reliance on college sophomores and manufactured "experts" (i.e., experts created in the lab using 3 - 20 hours of practice in a toy world domains) when evaluating displays. The Moray et al. study is noteworthy in terms of being one of the few studies that has taken the trouble to enlist real experts to evaluate displays for a complex work domain.

Conclusions

> . . . an active research science cannot be intelligently understood by reference to the rational rules of science alone. It is equally necessary to understanding the paradigm that guides the scientists to do experiments. Without understanding the paradigm, a student may find the experiments unrelated to each other; or the answers the experiments are supposed to provide may seem incomprehensible. The questions the scientists have chosen to ask may seem trivial or exotic, and their controversies may resemble tempests in teapots. However, to the student who grasps the paradigm guiding research, the relationship between theory and experiment will become clearer. The way in which experiments relate to each other will become more evident. The questions scholars in the field have chosen to ask will not seem so arbitrary, and their approach to answering the questions will look more reasonable" (p. 15).

In this chapter, we have focused more on elucidating a paradigmatic framework for evaluating displays, than on detailing methodological prescriptions for research. In fact, the detailed methodological implications of this paradigm remain to be worked out. The critical premise of the paradigmatic framework is that meaning is a central consideration when evaluating

interfaces. Every choice that a researcher makes when designing experimental programs to evaluate displays--- choice of the dimensions of displays to vary, of the task, of the performance measures, of the subject populations, of the cover story --- must be informed by considerations of meaning.

For those who disagree with this central premise, for those who consider meaning to be peripheral (a nuisance variable, an obstacle to generalizability, or an orthogonal dimension to syntactical considerations) our arguments will indeed be perceived as a tempest in a teapot. There is no rational basis to resolve this disagreement.

Others agree in principle that meaning is central, but have pragmatic concerns about the possibility for science in a world where evaluations reflect the specific constraints of particular work domains. The concern focuses on the issue of generalizability. Is it possible to discover general principles? Is it possible to generalize from solutions in one domain (e.g., aviation) to problems in another domain (e.g., process control)? Without such generalizations, is it possible to do science?

In the information processing model, generalization was guided by the boxes in our models. So, for example, the Sternberg (1966) task was used as a "dip stick" to determine whether a particular problem loaded on peripheral (encoding/response generation) or central (working memory) stages of processing. A solution that was effective for relieving memory load in one domain could be generalized to reducing memory load in a second domain.

What is the basis for generalization in a paradigm where meaning is central? Again, much work needs to be done, but we believe that it will be possible to generalize based on structural properties of task domains. Hammond (1993) has argued that generalizations are possible with naturalistic approaches to decision making. He notes that such generalization will require the development of formal models of the environment. These formal models will be reductionist, in the sense that many of the environmental details will not be critical to the generalization. So, for Hammond, social judgment theory (SJT) provides a general theory of task environments that is "independent of the substantive materials of any particular judgment task" (p. 212).

Carswell and Wickens (1987; 1990) have also identified an important structural property of work domains or tasks that has important general implications for displays. This is the dimension of task integrality. Integrality refers to the extent to which the meaningful distinctions within a task space are the measured variables themselves (separable), or whether the meaningful distinctions depend on relations across multiple measured variables (integral). In most complex socio-technical-systems, integrality is the rule. The demands for integration across measured variables in an important motivation behind the increasing interest in graphical displays. Configural graphic displays can be powerful devices for supporting perceptual integration. It is important to note, however, that despite the increased need for integrality, there are still important occasions that demand direct attention to the measured variables. For example, communication between multiple remote operators often depends on the precision of specific values for measured variables (Hansen, In press). So, often representations will have to allow operators to perceive both specific measured variables and higher order relations across those variables. Bennett and Flach (1992) discuss this dual

design goal and the implications for configural displays.

The research on manual and process control is another example where such generalizations have been successful. In these research areas, domain constraints such as the order of control and the magnitude of feedback delays provide a basis for bounding problems and generalizing solutions. Models such as McRuer's crossover model (e.g., McRuer & Jex, 1966) and the Optimal control model reflect both cognitive and domain constraints (see Flach, 1990 for an expanded discussion of these models and detailed references). In particular, McRuer's crossover model illustrates that there is no invariant human transfer function. Invariance is only apparent at the level of the human-machine system, and this invariance reflects global constraints on stability in closed-loops systems. These constraints are independent of whether that system is composed of human and machine or of purely electronic or physical components. These constraints reflect properties of the deep structure of control problems. Rather than generalizing based on common elements (common processing stages), we believe that generalizations should be guided by global properties of the human/machine/environment-organization. Thus, control theory and dynamical systems theory may help to guide generalizations and to provide the framework for discovering structural, rather than piecemeal truths.

The critical point is that generalizations must be grounded in a theory of meaning. Meaning is a relational property that requires theoretical constructs that integrate over actor and environment. Meaning is not of mind, nor is it of matter. Meaning is about what matters (Flach 1994)! A good interface must provide a representation of what matters! It must make the distinctions that matter perceivable! A sound methodological approach to display evaluation must do the same thing. It must be representative of the semantic as well as the syntactic constraints on performance. If displays are to support creative problem solving and coordinated adaptation, then evaluations of displays must be meaning-full! This is not an anti-reductionist position. Simplification and control will still be important for insuring internal validity. However, parsing must preserve the essence of the functional human-machine-environment system. The parsing must preserve meaning. If, as research scientists, we sacrifice meaning to achieve control, then we will have nothing to offer in the quest to design representations for visualizing the possibilities of tomorrow.

Acknowledgments

Our sincerest thanks to Raja Parasuraman and Mustapha Mouloua for their invitation to contribute to this volume. John Flach received support through a grant from the Air Force Office of Scientific Research during preparation of this chapter. Opinions expressed are those of the authors and do not represent an official position of the Air Force or any other organization.

References

Bahrick, H. P., & Karis, D. (1982). Long-term ecological memory. In C. R. Puff (Ed.), *Handbook of research methods in human memory and cognition* (pp. 427-465). San Diego: Academic Press.

Barnett, B. J., & Wickens, C. D. (1988). Display proximity in multicue information integration: The benefits of boxes. *Human Factors*, 30(1), 15-24.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*.

Cambridge, England: Cambridge University Press.
Bennett, K. B., & Flach, J. M. (1992). Graphical displays: Implications for divided attention, focused attention, and problem solving. *Human Factors*, 34(5), 513-533.
Brunswick, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley, CA: University of California Press.
Carswell, C. M., & Wickens, C. D. (1987). Information integration and the object display. *Ergonomics*, 30, 511-527.
Carswell, C.M. & Wickens, C.D. (1990). The perceptual interaction of graphical attributes: Configurality, stimulus homogeneity, and object integration. *Perception & Psychophysics*, 47, 157-169.
Casey, E. J., & Wickens, C. D. (1986). Visual display representation of multidimensional systems: The effect of information correlation and display integrality (Tech. Report CPL-86-2). Urbana-Champaign: Cognitive Psychophysiology Laboratory, University of Illinois.
Dewey, J. (1896). The reflex arc concept in psychology. *Psychological Review*, July, 357-370. Also in *John Dewey: The early works, 1882 - 1898, 5: 1895 - 1898*. Carbondale: Southern Illinois University Press.
Flach, J. M. (1990). Control with an eye for perception: Precursors to an active psychophysics. *Ecological Psychology*, 2(2), 83-111.
Flach, J.M. (1994). Ruminations on mind, matter, and what matters. *Proceedings for the 38th Annual Meeting Human Factors and Ergonomics Society*. Santa Monica, CA: Human Factors and Ergonomics Society.
Gentner, D. (1983). Structural mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2), 155-170.
Gentner, D., & Gentner, D. (1983). Flowing waters or teeming crowds: Mental models of electricity. In D. Gentner and A. L. Stevens (Eds.), Mental models (pp. 99-129). Hillsdale, NJ: Lawrence Erlbaum Associates.
Gibson, J. J. (1979). *The ecological approach to visual perception*. Boston, MA: Houghton Mifflin.
Haken (1988). *Information and self organization*. Berlin: Springer-Verlag.
Hansen, J.P. (In press). Representation of system invariants by optical invariants in configural displays for process control. In Hancock, P.A., Flach, J.M., Caird, J.K. & Vicente, K.J. (eds.) *Local applications in the ecology of human-machine systems*. Hillsdale, NJ: Lawrence Erlbaum Associates.
Hammond, K. R. (1993). Naturalistic decision making from a Brunswikian viewpoint: Its past, present, future. In G. A. Klein, J. Orasanu, and C. E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 205-227). Norwood, NJ: Ablex Publishing Corp.
Hutchins, E. L., Hollan, J. D., & Norman, D. A. (1986). Direct manipulation interfaces. In D. A. Norman, and S. W. Draper (Eds.), *User centered system design* (pp. 87-124). Hillsdale, NJ: Lawrence Erlbaum Associates.
Lachman, R., Lachman, J. L., & Butterfield, E. C. (1979). *Cognitive psychology and information processing: An introduction*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
Mace, W. M. (1977). James J. Gibson's strategy for perceiving: Ask not what's inside your head but what your head's inside of. In R. E. Shaw & J. Bransford (Eds.), *Perceiving, acting and knowing* (pp. 43-65). Hillsdale, NJ: Erlbaum.
MacGregor, D., & Slovic, P. (1986). Graphic representation of judgmental information. *Human-Computer Interaction*, 2, 179-200.
McRuer, D. T., & Jex, H. R. (1967). A review of quasi-linear pilot models. *IEEE Transactions on Human Factors in Electronics*, HFE-8(3), 231-249.
Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81-97.
Moray, N., Jones, B. J., Rasmussen, J., Lee, J. D., Vicente, K. J., Brock, R., & Djemil, T. (1993). A performance indicator of the effectiveness of human-machine interfaces for nuclear power plants. NUREG/CR-5977. Washington, DC: USNRC.
Neisser, U. (1976). *Cognition and reality*. San Francisco: Freeman.
Neisser, U. & Winograd, E. (1988). *Remembering reconsidered: Ecological and traditional approaches to the study of memory*. New York, NY: Cambridge University Press.

Pew (1994). Situation awareness: The buzzword of the '90s. *CSERIAC Gateway*, 5(1), 1 - 4.

Rasmussen, J. (1986). *Information processing and human-machine interaction: An approach to cognitive engineering*. New York, NY: Elsevier Publishing Co., Inc.

Rasmussen, J., & Vicente, K. (1989). Coping with human errors through system design: Implications for ecological interface design. *International Journal of Man-Machine Studies*, 31, 517-534.

Roth, E.M., Bennett, K.B. & Woods, D.D. (1987). Human interaction with an "intelligent" machine. *International Journal of Man-Machine Studies*, 27, 479-525.

Sanderson, P. M., Flach, J. M., Buttigieg, M. A., & Casey, E. J. (1989). Object displays do not always support better integrated task performance. *Human Factors*, 31(2), 183-198.

Shannon, C. E. & Weaver, W. (1963). *The mathematical theory of communication*. Urbana, IL: University of Illinois Press.

Simon, H.A. (1981). *The sciences of the artificial*. 2nd Ed. Cambridge, MA: MIT Press.

Sternberg, S. (1966). High-speed scanning in human memory. *Science*, 153, 652-654.

Vicente, K. J. (1991). Supporting knowledge-based behavior through ecological interface design. (Tech. Report EPRL-91-1). Urbana-Champaign: Engineering Psychology Research Laboratory and Aviation Research Laboratory, University of Illinois.

Vicente, K. J. (1992). Memory recall in a process control system: A measure of expertise and display effectiveness. *Memory & Cognition*, 20(4), 356-373.

Vicente, K. J., & Rasmussen, J. (1990). The ecology of human-machine systems II: Mediating "direct perception" in complex work domains. *Ecological Psychology*, 2 (3), 207-249.

Wertheimer, M. (1959). *Productive thinking*. New York: Harper & Row.

Wickens, C. D. (1986). The object display: Principles and a review of experimental findings (Tech. Report CPL-86-6). Urbana-Champaign: Cognitive Psychophysiology Laboratory, University of Illinois.

Wickens, C. D., Kramer, A., Barnett, B., Carswell, M., Fracker, L., Goettl, B., & Harwood, K. (1985). Display-cognitive interface: The effect of information integration requirements on display formatting for C3 displays (Tech. Report EPL-85-3). Urbana-Champaign: Engineering Psychology Research Laboratory and Aviation Research Laboratory, University of Illinois.

Wilson, J. R., & Rutherford, A. (1989). Mental models: Theory and application in human factors. *Human Factors*, 31 (6), 617-634.

Woods, D. D. (1991). The cognitive engineering of problem representations. In G. R. S. Weir & J. L. Alty (Eds.), *Human-computer interaction and complex systems* (pp. 169-188). London: Academic Press.