SPECIAL SECTION

A Framework for Epistemological Analysis in Empirical (Laboratory and Field) Studies

Yan Xiao, University of Maryland School of Medicine, Baltimore, Maryland, and **Kim J. Vicente,** University of Toronto, Toronto, Canada

In their search for generalizable behavioral patterns and design principles, cognitive field researchers should reflect on the epistemological limitations of empirical studies. In this paper we describe a framework for epistemological analysis that can help serve this purpose and discuss its application to two prototypical cases of cognitive engineering research: laboratory experiments and field studies. The framework examines two, often implicit, processes in empirical research: the abstraction from empirical data and the substantiation of theoretical constructs and principles. By explicitly considering these two processes in several systematic steps, we can gain appreciation for the epistemological contribution of empirical studies to cognitive engineering research. The framework and its application also provide guidance to such important issues as generalizability of results and external validity. Possible applications of this research include providing guidance to researchers and practitioners in evaluating design principles or conducting field studies.

INTRODUCTION

> Although they are a critical part of science, data are not the purpose of science. Science is about predictability, and predictability derives from models. Data are limited to the special case of what happened when the measurements were made. Models, on the other hand, subsume data. Only through models can data be used to say what will happen again, before subsequent measurements are made. Data alone predict nothing.
>
> Ahl and Allen (1996, p. 45)

Empirical studies – including both laboratory experiments and field observations – are indispensable research methods that can be used to compare alternative design proposals, to evaluate design principles, and to understand how a task is accomplished in a given setting. Cognitive field research and cognitive task analysis both rely extensively on empirical observations. Such studies provide crucial points of contact with the regularities that govern the interaction between people and their environment. They provide a means of knowing how the world works (as opposed to how one might think it works). However, such points of contact are invariably limited by the specific ways in which a particular study is conducted.

In comparison, researchers' interests often lie well beyond the exact conditions of any one study. For example, in an experiment evaluating a particular interface design, the primary object of interest is likely to be the value of the principles that were used to generate the design, not just the particular interface, tasks, or participants in that experiment. Similarly, in a field study of the strategies used by anesthesiologists in surgical operating rooms, the primary object of interest is likely to be the set of findings that generalize to a range of settings and operators, not just those observed in that

Address correspondence to Yan Xiao, Anesthesiology Research Laboratory, University of Maryland School of Medicine, 10 S. Pine St., MSTF 534, Baltimore, MD 21201; yxiao@umaryland.edu. **HUMAN FACTORS,** Vol. 42, No. 1, Spring 2000, pp. 87–101.

particular study. Thus the inevitable fact of empirical life is that one must study the specific in order to know about the generic. This is unavoidable because a generic construct cannot be directly observed, only an instantiation of a construct (Chow, 1996; Meehl, 1978).

The frequent need to generalize empirical findings beyond the specific details of the study in which data were collected raises a critical question: How does one assess the boundary conditions of the results obtained from an empirical study? This question is of obvious interest to researchers in any field of inquiry because it is fundamental to the development of scientific theory. Theory is what allows one to generalize the results of a particular study to other conditions. However, the question is particularly important for the field of cognitive engineering for another reason.

Designers and human factors engineers – the intended consumers of the results obtained in empirical studies – have a practical interest in the issue of boundary conditions. Although designers may not be interested in theory per se, the transfer of information from basic research to applied problems cannot be accomplished unless designers know whether the findings of a particular study, or set of studies, generalize to the particular problem with which they happen to be struggling. Thus the issue of boundary conditions on empirical knowledge is at the very core of research and design in human factors and cognitive engineering (Chapanis, 1967, 1988; Fisher, 1993).

Nevertheless, examples abound in the human factors literature of studies that have paid insufficient attention to the limits on generalization. Woods (1993) pointed out several manifestations of such lack of attention in field studies. We will illustrate this point by using a sanitized, anonymous example of laboratory studies from the literature on voice input technology.

From a human performance perspective, voice input technology seems attractive because it can potentially eliminate the use of keyboards, thereby freeing operators' hands for other purposes (e.g., flight control). Thus an obvious research issue is the relative effectiveness of voice versus keyboard input methods. One way of answering this research question is to conduct an experiment to compare the two input methods. Let us assume that the keyboard input method is found to be superior to the voice input method in the task or tasks used in the experiment. Some researchers might conclude, as a universal design principle, that keyboard input is superior to voice input.

How generalizable might such a design principle be? It is difficult to answer this question without critically examining the process by which the experiment was conceived and conducted and the way in which the experimental results were theoretically linked to the design principle. Should a reader ignore the design principle? Not necessarily. However, it would be prudent to keep in mind that the results may be attributable to the primitive state of voice input technology, not to voice input per se. In other words, the particular implementation tested in the experiment may be deficient, but the general concept of voice input technology may in fact be of some value, despite what the data seem to suggest. If so, then the next generation of voice input technology could very well invalidate the design principle previously described.

The general point to keep in mind is that design principles of this type can be implemented in different artifacts and in different application domains. When a particular artifact and a particular application domain are chosen, the question arises as to whether the results observed can be interpreted as supporting the generic design principle or are caused by other idiosyncratic factors (such as the primitive state of technology of voice input). Questions would also remain as to whether other artifacts designed according to the same principle but for another application domain would obey the prediction made by the design principle.

The cognitive engineering community has yet to agree on an explicit, defensible way of answering these fundamental questions, despite the fact that they are at the very heart of the research process. This paper describes a framework that can be used as a tool to obtain answers to these questions.

Despite Chapanis's (1988) claim to the contrary, the issue of generalizability has been extensively debated. This debate has been centered predominantly in the context of field versus laboratory studies in a seemingly rival manner, as if one of the two types of studies was privileged to produce generalizable results (e.g., Barker, 1969; Berkowitz & Donnerstein, 1982; Dipboye & Flanagan, 1979; Henshel, 1980; Mook, 1983; Webster & Kervin, 1971). These issues have been revisited (e.g., Banaji & Crowder, 1989; Driskell & Salas, 1992; Fisher, 1993; Hoffman & Deffenbacher, 1993; Vicente, 1997). Although it is important to examine the generalizability of a research finding (from a laboratory study or from a field study) as a final product, we believe that one useful way to do so may be to examine the *process* by which a finding is generated and validated.

In this paper, rather than viewing generalizability as a property of the end product of a research effort, we examine generalizability in a larger context and consider the process of conducting empirical research. By reflecting on this process, we may be able to address the limitations to generalization of empirical studies without pitching the two types of studies (field and laboratory) against each other. In other words, we propose that we need to analyze the process of empirical research in a structured and systematic manner to better understand the boundary conditions of any study. We refer to this critical activity as *epistemological analysis* because our focus is on the validity and limits of knowledge (see definition of "epistemology" in Webster's 9th Collegiate Dictonary).

Although the term *epistemology* might sound overly philosophical to some, each empirical study involves epistemological concerns: how the researcher went about collecting empirical data, analyzing data, reasoning from the data, and deriving potentially generalizable results. When this process is considered implicitly, the limitations to generalization are often unclear, as the voice input technology example given earlier illustrates. For a laboratory study, it could mean that unwarranted limitations are made on how the results could be used in applied settings. For a field study, the lack of an explicit epistemological analysis could mean that the results may be difficult to transfer into different domains or even across cases within the studied domain.

We first describe a systematic framework for epistemological analysis. The framework provides not a rigid recipe but, rather, a consistent structure that can be instantiated in various forms to suit the needs of particular instances. Two examples are provided to illustrate the application of the framework to typical activities in cognitive engineering research: the planning of experimentation and the abstraction of field data. The ideas in the framework are not new (Hollnagel, Pedersen, & Rasmussen, 1981; Holton, 1986), but they are not well known or widely used in the cognitive engineering community (see Patterson & Woods, 1997, and Woods, 1993, for two exceptions). Thus one contribution of this paper is to bring these ideas to the attention of the *Human Factors* audience. The first example is a laboratory investigation of interface design principles (Vicente, 1992) that is used here to demonstrate how the framework helps to identify the limitations of generalizability of experimental results. The second example is a field study of problem-solving expertise (Xiao, 1994) that used the framework to help derive findings that may generalize across other application domains. These examples illustrate how the application of the epistemological framework can help the cognitive engineering community address the critical problems centered on generalizability mentioned earlier.

TWO COMPLEMENTARY PROCESSES OF EPISTEMOLOGICAL ANALYSIS

In a classic paper, Meehl (1978) illustrated vividly the substantial distance between empirical contact with the world and a theory, the likely objective of an empirical study (see also Chow, 1996). Meehl observed that one cannot test a theory without making auxiliary assumptions and instantiating theoretical constructs. Thus in testing a theory, the researcher has to resolve a number of degrees of freedom – usually many – by making a number of decisions to create a protocol that is detailed enough to carry out an experiment. Surprisingly, we found that not a single *Human Factors* article has ever cited Meehl's article. We do not mean to suggest that human factors researchers are oblivious to the problems with empirical studies that are addressed in Meehl's article. Instead, we interpret this fact to indicate that such researchers have tended to avoid explicitly discussing issues of epistemology. The framework described next embodies Meehl's view on the distance between theories and empirical contact with the world.

A Framework for Epistemological Analysis

Hollnagel et al. (1981) took on the task of integrating data from different sources (e.g., laboratory investigations, accident reports, and simulator studies). To develop an approach to generalizable theories from concrete data obtained from different empirical studies, they proposed a multiple-stage process, illustrated in Figure 1, with each successive upward stage representing an abstraction of the previous one. This abstraction process removes specific details that are peculiar to the context of a study and replaces them with strategies and performance criteria that may be relevant to a variety of contexts. The goal of the analysis is to achieve the representation of data at different levels of abstraction. Because a "higher-level" representation of a finding contains less reference to details, such a representation enables researchers to integrate results from different studies (thus potentially allowing greater generalizability). In essence, this structure is an abstraction hierarchy (Rasmussen, 1985) for the work domain of a researcher (see Vicente, 1999).

Such a view of successive abstraction for generalizability complements Einstein's epistemological view of science, as described by Holton (1986). This view embodies two essential aspects of scientific endeavor: the testing of theories by generating hypotheses and the abstraction of theories to form higher levels of description of experience.

> One could . . . visualize the progressive development of scientific theory to take place as the development of the system of concepts at an increasingly higher level of "layers" or strata, each layer having fewer and fewer direct connections with the complexes of sense experience. In this way, a more phenomenological theory at the early stage of science . . . gives way to a more independent set of concepts and axioms. . . . There is of course a cost in this developmental process. By going cyclically through several stages of theories,

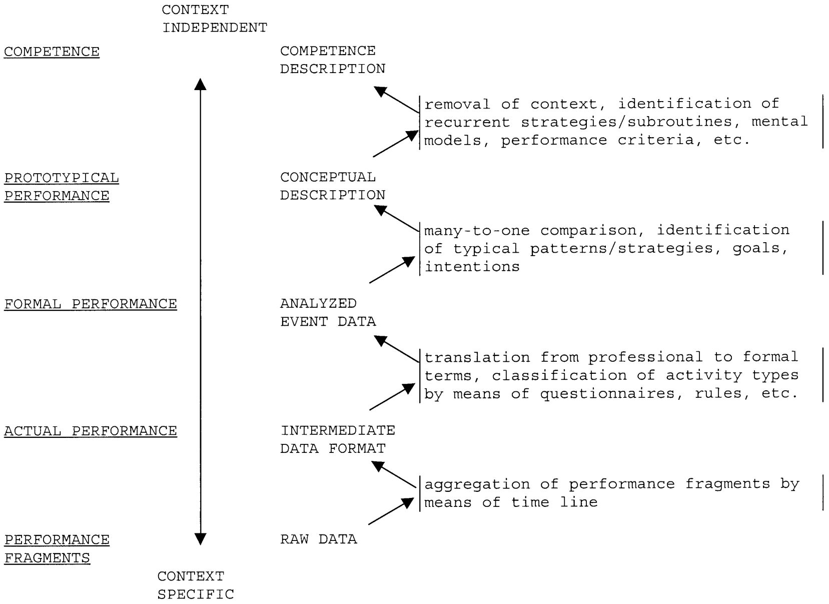
CONTEXT INDEPENDENT
COMPETENCE
COMPETENCE DESCRIPTION
removal of context, identification of recurrent strategies/subroutines, mental models, performance criteria, etc.
PROTOTYPICAL PERFORMANCE
CONCEPTUAL DESCRIPTION
many-to-one comparison, identification of typical patterns/strategies, goals, intentions
FORMAL PERFORMANCE
ANALYZED EVENT DATA
translation from professional to formal terms, classification of activity types by means of questionnaires, rules, etc.
ACTUAL PERFORMANCE
INTERMEDIATE DATA FORMAT
aggregation of performance fragments by means of time line
PERFORMANCE FRAGMENTS
RAW DATA
CONTEXT SPECIFIC

Figure 1. Data analysis methodology of Hollnagel et al. (1981). Viewed as two-way processes (instead of upward abstraction), this methodology can be viewed as a framework for epistemological analysis, which links theories to data through several stages.

> each stage is forced to use conceptions more removed from direct experience . . . [T]he contact with common sense is more and more tenuous. But the fundamental ideas and laws of science attain a more and more unitary character. (Holton, 1986, pp. 48–49)

The common feature of the two epistemologies illustrated by Figure 1 and by Holton's quotation is the separation of empirical data ("sense experience," to use Einstein's words) and abstracted findings ("axiom") into stages of analysis (see Chow, 1996; Meehl, 1978). Such a staged approach allows one to examine explicitly several critical points in the processes of experimental design and data analysis, thereby allowing the assessment of a particular study's epistemological limitations.

The steps in Figure 1 by Hollnagel et al. (1981) were originally developed for data analysis (from bottom to top). We propose that these steps can be viewed as bidirectional processes (from bottom to top and from top to bottom). Under this view, Figure 1 depicts a framework for epistemological analysis that makes explicit the top-down instantiation and bottom-up abstraction processes between an abstract theory and data in empirical research. The two-way processes in this framework can guide the instantiation of theoretical constructs and design principles, the choice of testing application domains, the synthesis of experimental results, the analysis of behavioral data from field studies, and the abstraction of results for use across different application domains.

Explicitly separating abstraction and instantiation processes into various levels leads to several important insights. First, it is evident that as one moves up the levels in Figure 1, the potential generalizability of research findings increases. At the bottom level, the description is of a single data set obtained from an experiment or a field study conducted under a specific set of conditions. Although this is a very precise level of description, it can be difficult to read much meaning into the data. The meaningfulness of the results is increased by interpreting the findings at a higher level of abstraction within a given set of theoretical concepts. The advantage of doing so is that the potential implications of the "numbers" obtained from a particular empirical encounter (an experiment or a field study) become more tangible.

There are, however, inherent sources of danger in this abstraction process. For example, the significance of the data can change as a function of the theoretical concepts that are chosen for analysis, thereby introducing an element of relativism. Furthermore, higher levels are far removed from the data, and thus certain details are being left out as a function of the theory that has been chosen to describe the results. But as Meehl (1978) observed, the experimental results are a function of the theoretical propositions, the way in which those propositions were operationalized, and the conditions under which the experiment was conducted. These dependencies will be lost in the process of abstraction.

In field studies researchers are concerned with the epistemological question of how to derive general statements or findings based on the field observations made. Observations made in the field potentially reflect the results of many factors affecting a worker's behavior. Each observation is unique to the time of observation. The first hurdle is to arrive at statements that describe the complete set, or an important subset, of observations. Necessarily, a process of abstraction is involved to remove some details while simultaneously replacing them with a description in a more abstract language. The second hurdle is to arrive at statements that can describe behavior across cases or even domains. The third hurdle is to synthesize statements from different field studies in different cases and different domains into a coherent theory of human performance. In other words, the process of analyzing data from a field study can be viewed as an upward trajectory through Figure 1.

The framework for epistemological analysis represented by Figure 1 can thus be used as a tool to examine the issue of generalization in both types of empirical studies. In the following sections we will illustrate the use of the framework by conducting an epistemological analysis of two empirical studies: one conducted in the laboratory (to show primarily the downward process for Figure 1) and the other in the field (to show primarily the upward process for Figure 1).

EXAMPLE 1: EPISTEMOLOGICAL ANALYSIS OF A LABORATORY STUDY

A cognitive engineering design principle is, in essence, a theory that prescribes a way to enhance human performance via design intervention(s). The evaluation of a design principle is thus a test of a theory, and it is a prototypical activity in cognitive engineering research. Consequently, the following example of epistemological analysis should be of broad relevance.

Vicente (1992) conducted a laboratory experiment to evaluate one of the principles of ecological interface design (EID), a theoretical framework for designing interfaces for complex human-machine systems (Vicente & Rasmussen, 1992). A critical question arising from this particular study was, To what extent can the specific results obtained in this single study be said to provide support for the EID framework?

Instead of answering this question directly, as is typically done, Vicente (1991) conducted an explicit epistemological analysis using the framework described in the previous section. Figure 2 illustrates how the framework in Figure 1 was adapted for this laboratory experiment. Note that the labels in Figure 2 differ from those in Figure 1 because the specific purpose of Vicente's analysis differed from that of Hollnagel et al. (1981). The key point, however, is that the relationship between levels is identical in both cases, and it is this general structure that is the essence of the epistemological framework. In this section we will show how this multiple-level structure was used to elucidate a subset of the downward process of experimental design (synthesis). We will also show the upward process of data interpretation (analysis) as a way to demonstrate the utility of the framework (see Vicente, 1991, for the full epistemological analysis).

Overview

The EID principle evaluated in this experiment was as follows: To provide proper support for knowledge-based behavior (KBB), an interface should be based on an abstraction hierarchy (AH) representation of the work domain (see Rasmussen, Pejtersen, & Goodstein, 1994, for a detailed description

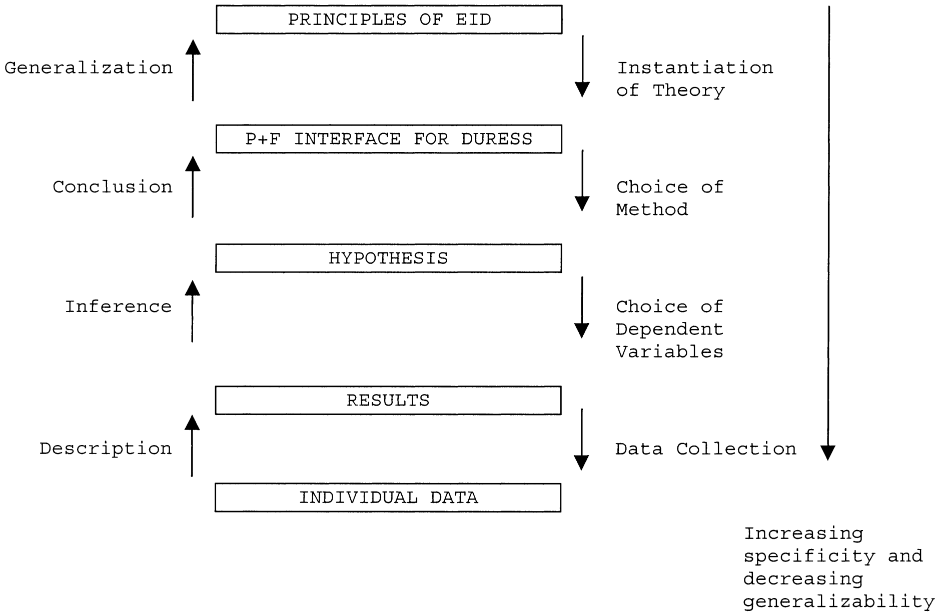

Figure 2. The epistemology of Vicente (1992), adapted from Figure 1, for guiding the epistemological analysis of an experimental study.

of KBB and AH). As shown in Figure 2, the first step in making this abstract claim more concrete is to instantiate the theory by building two interfaces, one that embodies the theoretical claim and one that does not. This design effort must inevitably be conducted within a specific context. In this case a representative process control simulation, dual reservoir system simulation (DURESS), was developed as an experimental test bed.

Doing this allowed us to move to the second level in Figure 2, which is a more concrete restatement of the principle of EID in terms of two particular interfaces for DURESS: a physical (P) interface that contains a subset of the information in the AH for DURESS and a physical plus functional (P+F) interface that contains information at all levels of the AH for DURESS. Using these terms, EID predicts that the P+F interface should result in better support for KBB than would the P interface because the former represents all levels of the AH whereas the latter does not. This theoretical claim needs to be translated into an empirically testable hypothesis, which requires a choice of experimental method.

In this study a combined diagnosis-memory task for theoretical experts in thermalhydraulics was chosen as an appropriate way to evaluate support for KBB (Vicente, 1988). As illustrated in Figure 2, the outcome of this step is a theoretically motivated empirical hypothesis: that the P+F interface should result in better diagnosis and memory performance than would the P interface.

As shown in Figure 2, the next step involved a choice of dependent variables. How should the accuracy of diagnosis and memory performance be measured? Having made this decision (see Vicente, 1992, for details), we can then describe the specific set of results that would be consistent with the empirical hypothesis.

Finally, at the bottom level, the data are collected. As shown in Figure 2, the exact conditions under which the experiment is to be conducted must be specified at this point (i.e., all of the degrees of freedom will have to have been resolved).

The following epistemological analysis will show that (a) it is impossible to test the design principle "directly" because the concepts need to be operationalized and auxiliary assumptions must be made to reach the level of empirical measurement; (b) there are many degrees of freedom that must be dealt with in empirically testing the principle; and (c) the choices that are made in dealing with these degrees of freedom have crucial implications for the upward analysis phase of experimentation (i.e., they strongly constrain the inferences that can reasonably be made from the data obtained).

Instantiating the theory. There are at least five potential sources of degrees of freedom in instantiating the EID framework. First, one must conduct an AH analysis to determine the information content of the P+F interface. This step is relatively constrained by the physical principles governing DURESS. As a result, different knowledgeable designers should develop AH representations for DURESS that are pragmatically equivalent.

Second, one must map this information content onto a particular visual form. This issue has many more degrees of freedom because conceivably many different forms can be used to represent the same information in the AH representation of DURESS. Moreover, some of these degrees of freedom are simply not addressed by the EID framework (e.g., decisions about color coding, size of labels, location of labels and components). Nevertheless, choices must be made about each of these attributes in developing an interface to be tested. Thus the degrees of freedom cannot be ignored.

The power of epistemological analysis is that it makes this point explicit and thereby helps in the identification of potential consequences for data analysis and generalization. In this case the empirical results that are obtained may be valid only for the coding scheme that is adopted. Furthermore, it will not be possible to determine, from this experiment alone, whether the results are driven more by the design principle being tested or by the decisions that were made about coding attributes.

Third, there are many degrees of freedom in the selection and design of the control interface. If the P interface is a mere "straw man," then the expected advantage of the P+F interface might not carry much theoretical or practical weight. In this experiment several steps were taken to deal with this issue in a conservative manner. For example, an attempt was made to control for the effects of interface design issues that are not addressed by the EID framework. Thus the color coding, labeling, and spatial coding were designed to be, to the extent possible, consistent across the two interfaces. Also, the P interface was chosen so that it actually represented an improvement over many existing industrial control room interfaces consisting of analog, hard-wired meters. At least with the P interface, the spatial topology of the components is explicitly represented, unlike designs that consist of a bank of meters.

Despite this effort, it will not be possible to determine from this experiment alone if the results obtained would change, qualitatively or quantitatively, if a different control interface were adopted. Again, the epistemological analysis shows the limits of empirical knowledge. One cannot know whether the results are driven more by the principle being tested or by the choice of controls and factors held constant.

Fourth, there are also substantial degrees of freedom in the choice of application domain. Here the important question is, Would the principles of EID be as effective if they were to be applied to a domain other than DURESS? As before, the issue can be ignored but not escaped. In the present study this issue was dealt with in a principled fashion by using Brunswik's (1956) principle of representative design of experiments. A target class to which we were interested in generalizing our results was first chosen (process control systems). Then a test bed was constructed so that it possessed some of the structural properties that are representative of that class of domains, albeit on a much smaller scale (Vicente, 1991). By following this procedure, the experimental findings obtained here may generalize to real-world domains that are structurally similar to DURESS. This last statement is qualified because the problem has been scaled down considerably, and thus the extent of this generalization is difficult to assess without additional experiments. Although it does not remove all of the available degrees of freedom, the use of representative design has the advantage of integrating experimental design and data generalization in a nonarbitrary and pragmatically directed manner.

Finally, there are degrees of freedom in the choice of interface designer. Here the critical question is, If we were to provide various interface designers with the principles of EID and ask them to build an interface for a given system, would they all come up with the same design? In this experiment the interface designer was one of the creators of the EID framework, a situation that will not be replicated in practice in industry. Thus it is not possible to infer from this experiment whether the results obtained are attributable to the EID framework or to the interface designer's generic skill (or lack thereof). Once more, our epistemological analysis identifies another significant limitation on the generalizability of the experiment results.

Construct operationalization. The operationalization of the theoretical concept of "degree of support for KBB" provides another source of degrees of freedom and, thus, epistemological uncertainty. One of the important issues that had to be dealt with was to ensure that experiment participants were in fact using KBB and not rule- or skill-based behavior (Rasmussen et al., 1994). To achieve this goal, it was important not to allow participants to perform the required task by merely being attuned to the perceptual properties of the display. Instead, they had to be encouraged to reason about the work domain using a mental model.

Several steps were taken to achieve these conditions, one of the most important being the choice of participants. Rather than using participants who had control experience with similar work domains (e.g., a power plant), the participants chosen were instead theoretical experts. The rationale for this choice was that theoretical experts in thermalhydraulics would have a relatively sophisticated mental model that they could use for KBB. Just as important, they would not be familiar with the perceptual properties of either interface, thereby discouraging reliance on rule- or skill-based behavior.

In addition to the choice of participants, a decision also had to be made concerning a task that would induce KBB. The most direct operationalization is that of diagnosis. Presenting participants with unanticipated events and asking them to diagnose those events is a prototypical case of KBB. The rationale for choosing the memory recall method as an indicator of degree of support for KBB is more involved and is described in Vicente (1988). Because this measure had not been used before for this purpose, its validity had to be established by empirical means. The diagnosis measure was therefore used as a criterion by which to validate the utility of the memory recall method as a measure of support for KBB.

Performance measures. The third choice point in the process of experimental design is the selection of dependent variables. As before, there are many degrees of freedom (i.e., many different ways of measuring the same theoretical constructs). Ideally, the chosen measures would satisfy three criteria: sensitivity, convergent validity, and discriminant validity (see Hammond, 1986). In this experiment a pilot study was used to narrow the possibilities for candidate measures.

Measurement conditions. The final source of degrees of freedom in the downward analysis process is the set of conditions under which the data were collected. Again, Brunswik's (1956) principle of representative design was used to make decisions. Two classes served as the targets for generalization in this case: one for the type of psychological demands placed on participants (KBB) and another for the competence level of the participants (theoretical experts).

Referring to the former, the experimental conditions were set up to induce KBB. Thus participants were faced with unfamiliar events and were not given any external aids except the interface itself. In addition, because the experimental conditions were set up in such a way as to minimize the influence of rote perceptual processing, participants' performance was, by default, primarily driven by the conceptual knowledge of thermalhydraulics that they brought to the task. As a result, the set of cognitive demands being placed on participants should be functionally similar to that experienced by operators faced with unanticipated events, in that both are required to rely primarily on KBB. The results, therefore, should generalize to this class of psychological demands and not necessarily to any other class (e.g., rule-based or skill-based behavior).

Representative sampling was also conducted on the participant side. The intention was to generalize the results to theoretical experts, so graduate students who had extensive formal training in the fundamental principles of thermalhydraulics were chosen. Note that the participants were not experts in controlling the system. Thus the results of this experiment cannot be defensibly generalized to operators of actual plants because they rarely qualify as theoretical experts. Theoretical experts were chosen instead of experts at control because the object of the study was knowledge-based behavior, the success of which is a function of the participants' conceptual model of the work domain rather than the appropriateness of their cue-action mappings.

Upward Analysis

The interpretation of experimental results follows a bottom-up path along the levels in Figure 2. The starting point is the data set obtained from the experiment, represented at the bottom level. At the next level up, these data can then be redescribed in terms of quantitative patterns. In this study these patterns were largely results of statistical tests. At the next level in Figure 2, the inferences one can make based on the observed results are represented. In this study, the P+F interface resulted in significantly better diagnosis and memory for functional variables than did the P interface. This inference led to a conclusion stated in the language of the theoretical concepts that guided this research: The P+F interface for DURESS provides better support for knowledge-based problem solving than does the P interface.

The final step in the interpretation process involves a generalization from the conclusion obtained in this experiment to other contexts. In this case the generalization, if warranted, would be that in order to provide proper support for KBB, an interface should be based on an AH representation of the work domain.

Summary

In summary, when we combine the choice points across levels, we see that an enormous number of degrees of freedom had to be accomodated in designing this experiment. Although these choices were dealt with in a principled manner wherever possible, the inescapable conclusion is that there are many significant limitations on what we can defensibly conclude from this (or any other) single empirical study.

What, then, are the potentially relevant factors that have not been accounted for in this experiment? Some of the more important degrees of freedom are (a) the criteria by which the visual forms of the two interfaces were designed, (b) the choice of the control interface, (c) the choice of test bed, (d) the competencies of the designer of the two interfaces, (e) the class of participants that was chosen, (f) the task that was chosen for the experiment, (g) the performance measures that were adopted, and (h) the way in which KBB and participant selection were operationalized.

Two points can be made about every one of these choice points. First, we do not know if the same results would have been obtained if a different decision had been made at each choice point. Second, we do not know to what extent the results obtained are attributable to the design principle being tested, the way in which these choice points were dealt with, or some combination of the two. Having conducted the epistemological analysis, however, we are aware of what we do not know and can design additional empirical studies to address these gaps.

EXAMPLE 2: EPISTEMOLOGICAL ANALYSIS OF A FIELD STUDY

Another category of prototypical cognitive engineering studies is observing human activities *in situ*, or field studies. This type of study is inductive, with the purpose of building theories of human cognition. Frequently, field studies in actual or simulated work domains are conducted to uncover strategies that reflect patterns of underlying cognition (e.g., Klein, Calderwood, & Clinton-Cirocco, 1986). Designers can gain insights from the discovered strategies to enhance and support human cognition. Training programs can be devised to transfer strategies to trainees. One of the key challenges to cognitive field researchers is to resolve the predicament that, on the one hand, strategies are inevitably context-bound, whereas on the other hand, generalization requires that strategies be as context-free as possible.

To further our understanding of the cognitive activities during the interaction between experienced practitioners and complex systems, a four-year field study was conducted in surgical operating rooms to evaluate how experienced anesthesiologists manage a complex system: the human physiology under anesthesia and surgery (Xiao, 1994; Xiao, Milgram, & Doyle, 1997). The field study focused on preparatory activities: how anesthesiologists use prior deliberations to prepare both mental and material resources for responding to anticipated events during surgical anesthesia.

The field of anesthesiology offers a window through which we can observe the types of cognitively challenging problems an experienced practitioner faces and how these problems are solved. Briefly described, the primary goal of surgical anesthesia is to provide necessary surgical conditions (unconsciousness and muscle relaxation) while at the same time safeguarding the patient (sustaining life processes and preventing unnecessary injuries to the patient). Achieving this dual goal is complicated by the fact that the anesthesiologist has limited means of monitoring and controlling human physiology and that each patient represents a different work domain for the anesthesiologist.

Overview

As in most field studies, a large quantity of data were collected through field notes and audio recordings. The challenge was to analyze the data collected and synthesize the findings in a potentially generalizable form. Among the temptations in meeting this challenge, two solutions are probably most often chosen:

1. Keeping descriptions of findings at a domain-specific level. This solution has some obvious shortcomings, despite the advantage that it may be easier to find supportive evidence directly in field data. For example, it would be difficult for researchers or designers of different domains to appreciate the findings and to “compare” notes from a different domain.

2. Producing a general theoretical statement with little information bridging field data and the statement. Readers have little to go by to assess the boundary conditions of the theoretical claim (see Woods's criticism on "A great leap from the data collected to interpretative conclusions" in Woods, 1993, p. 240).

To address the problems associated with these two solutions, the multiple-stage framework expressed in Figure 1 was adopted to establish explicit linkage between field data and findings through several levels of abstraction. As a direct consequence of the framework in Figure 1, data analysis would be aimed at answering different questions at different levels of abstraction. Similarly, data analysts would have to possess different types of knowledge to carry out the corresponding analysis. For example, in determining what happened and what was done based on collected data, one has to be literate about the subject matter of the domain under study. Like most field studies, the current study enlisted informants for this purpose. Figure 3 summarizes several important aspects in adapting the framework expressed in Figure 1.

The basic approach was to view the process of data analysis as a series of inductive abstractions (the upward progression in Figure 1), with the need for generalizable results as a guiding objective leading data analysis efforts. Note that the labels in Figure 3, like those in Figure 2, differ from those in Figure 1 because the specific purpose of Xiao's analysis also differed from that of Hollnagel et al. (1981).

Description of data in domain language. Data collected during this field study were taken from a variety of sources. Audio recordings were made during anesthetic procedures to capture "thinking-aloud" protocols, on-line verbal annotations, and answers to probing questions. Audio recordings were also made during case discussions, in which real or hypothetical cases were discussed among anesthesiologists. In addition to audio recordings, handwritten notes were made to document events and activities observed during surgery.

These data were performance fragments, "in the sense that they do not provide a coherent description of the performance, but rather the necessary building blocks or fragments for such a description" (Hollnagel et al., 1981, p. 10). A first step was to establish a coherent description of performance (to answer the questions at the bottom of the right-hand column in Figure 3). Even at this level of description, missing pieces to a coherent picture of

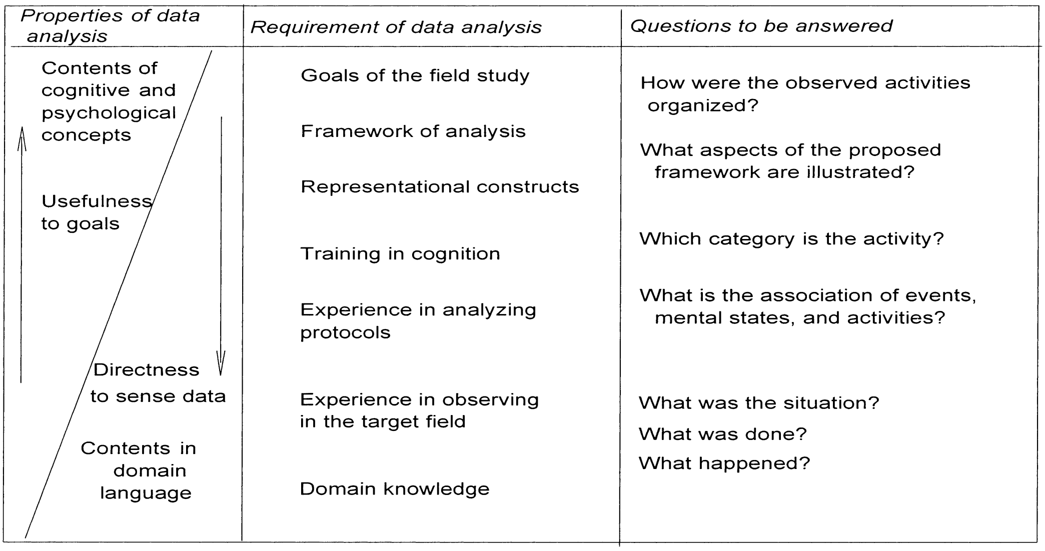
Properties of data analysis
Requirement of data analysis
Questions to be answered
Contents of cognitive and psychological concepts
Usefulness to goals
Directness to sense data
Contents in domain language
Goals of the field study
Framework of analysis
Representational constructs
Training in cognition
Experience in analyzing protocols
Experience in observing in the target field
Domain knowledge
How were the observed activities organized?
What aspects of the proposed framework are illustrated?
Which category is the activity?
What is the association of events, mental states, and activities?
What was the situation?
What was done?
What happened?

Figure 3. The epistemology of Xiao (1994), adapted from Figure 1, for guiding the epistemological analysis of a field study. Upward arrow: abstraction; downward arrow: interpretation.

events and activities had to be filled in. This step was therefore subject to uncertainties and biases on the part of the data analyst.

Description of data in domain strategies. As in many field studies carried out in high-skill domains, this field study identified a range of strategies used by anesthesiologists. These strategies described how tasks were accomplished. For example, anesthesiologists were found always to prepare syringes in advance of their use for both routine and emergency situations. If syringes were for routine use, anesthesiologists frequently laid out the syringes in the order in which they were to be used. As another example, anesthesiologists were found taping vaporizers (a control for the amount of anesthetic agent delivered to the patient) in a certain type of anesthetic procedure (total intravenous anesthetic). These strategies described in domain language could be traced back to field notes and recordings. They have the advantage that their validity could be examined by domain experts.

Description of performance in domain strategies was yet another step of abstraction, in the sense that not all field data were accounted for and that much of the details were removed. The resulting domain strategies thus constituted only one of many possible ways of describing the field data. The analyst had to answer questions such as "What is the association among events, mental states, and activities?" (questions at the middle of the right-hand column in Figure 3). Answers to these questions were essentially in the form of hypotheses.

Description of data in specific strategies. Along the direction of the upward progression of abstraction in Figure 1, the domain strategies identified could be further abstracted by removing domain language. For example, the strategy of preparing syringes was described as an off-loading strategy. The strategy of taping vaporizers during total intravenous anesthetic was described as configuring a fail-safe workspace. At this level of description, findings from the field studies can be compared with those from studies in other domains. For example, the strategy of off-loading was observed by Amalberti and Deblon (1992) in their studies of fighter pilots.

Categorization, itself an inductive process, was the analyst's main data analysis activity at this step. Again, multiple possible solutions existed to map domain strategies onto specific strategies.

Description of data in generic strategies. These specific strategies could be further abstracted and described in terms of cognitive constructs. The field study proposed a generic strategy to describe a range of the observed behavior: "Experienced practitioners reduce response complexity through anticipating future situations, mental preparation, and reorganizing the physical workspace."

Summary

Table 1 is a summary of the findings of the field study of problem-solving strategies in the domain of anesthesiology. At the bottom level, the findings are represented in domain language. Findings represented at this level can be traced back directly in the data collected from the field. In turn, these findings are difficult to generalize to domains outside anesthesiology because they do not reveal the underlying cognitive activities. After removing the domain context and adding cognitive descriptions, the domain strategies can be represented by specific strategies at the middle level. These strategies are no longer context specific and, thus, cannot be verified directly by empirical data. Instead, they were "illuminated" by examples in the empirical data. However, findings represented at this level should have wide implications in terms of design and training. At the top level, specific strategies can be abstracted into a single generic strategy, which attempts to represent a fundamental characteristic in the interaction between proficient workers and complex, dynamic task environments.

When following the framework in Figure 1 from top to bottom, we can use the generic strategy to direct the search for other kinds of specific strategies. In the case of the field study, the generic strategy was that practitioners prepare resources (both mental and material) for future interactions with the task environment. The intended result of such preparatory activities is a set of more compatible resources that will reduce the on-line response complexity –

that is, what one needs to do (physically and mentally) in response to the events occurring in the environment (Xiao et al., 1997).

One concern over the abstraction process described here is the reliability of the process. The more abstractly the findings are represented, the more likely the findings will be relevant to cases or even domains other than the ones under study and, thus, potentially, the greater the general relevance of the field study. However, the abstraction process is inductive, and multiple ways of supporting the abstraction process should be used. One way, used in several studies (Hollnagel et al., 1981; Moray, Sanderson, & Vicente, 1992), is to conduct different types of studies (e.g., observations, interviews, incident investigations). Another way is to use the findings (i.e., the end product of the abstraction process) to reinterpret the field data, to find examples to substantiate the findings, and possibly to modify and enrich the findings in the process (the downward progression in Figure 1). Further field studies, either in the domain studied (i.e., anesthesiology in this study) or in other domains, can also be conducted using the generic and specific strategies as a guide in identifying recurrent patterns in the observed behaviors.

DISCUSSION

Given the limited nature of cognitive researchers' empirical contact with the empirical world in any one study and their desire to produce findings that are generalizable beyond the exact conditions of a particular study, it would be dangerous for them not to address the philosophical concerns associated with empirical research. Nevertheless, these issues are frequently dealt with implicitly, which makes it difficult to achieve the goals researchers set for themselves. As in our example of voice input technology, the inherent limitations of empirical studies have caused confusion among researchers. Although various attempts have been made to overcome these limitations, most of these attempts have been concerned with the setting of a study – that is, laboratory versus field. Such attempts, although useful, do not address the fundamental limitations of empirical (laboratory or field) studies and the necessity of abstraction and generalization in research. Without a clear treatment of epistemology in empirical studies, one is often left with implicit assumptions that render findings with unclear limitations or boundary conditions.

Empirical studies, such as laboratory experiments and field studies, may lead researchers

TABLE 1: Findings Represented at Different Levels of Abstraction

Levels of Abstraction	Findings
Generic strategy	Reduce response complexity through anticipating future situations, mental preparation, and reorganizing the physical workspace
Specific strategies	Schedule tasks (offload) Build local models and rules Be preventive: think of probable side effects, pitfalls and predictors of these side effects Prepare necessary materials and access Rehearse pending procedures (mental simulation)
Domain strategies extracted from field data	Prepare induction and emergency syringes Pay more attention to muscle relaxation If blood pressure fluctuates too wildly, start nitroglycerin infusion Prepare nitroglycerin whenever there is a chance Tape vaporizers to prevent the use of vaporizers (in a total intravenous anesthetic) Use only short-acting drugs if surgery duration is uncertain

to believe they have learned something important – that they have made solid empirical contact with the world and have obtained stable knowledge of how the world works. As the opening quotation and the epistemological analyses make clear, however, this deep sense of security associated with empirical knowledge is unjustified. The results and conclusions obtained from any one empirical study, no matter how well designed, actually rest on a foundation of enormous epistemological uncertainty. Depressing as the outcome may be, it is important that this issue be addressed head on. Surprisingly many studies reported skip this step (such as the sanitized example used in the introduction). The consequence is that these limitations on knowledge and generalizability are left implicit, difficult to see, and thus not available for assessment by readers.

By examining the process of empirical studies, we can gain insights on ways in which experiments can be designed to gain generalizability (such as representative design). We can also examine the role of field studies and the associated constraints on the analysis of data from field. It is not our intention to require that every published paper report a detailed epistemological analysis, as we have done here. (Our two examples by no means constitute a "cookbook" of epistemological analysis.) Instead, our hope is that every empirical study will be guided by an explicit epistemological analysis. Such an analysis should allow the consumers of the research to understand the limitations of the product they are buying. The framework we have described facilitates such an analysis and thereby represents an important and valuable tool to add to the existing arsenal of cognitive engineering research methods.

ACKNOWLEDGMENTS

The writing of this paper was supported by grants from the Office of Naval Research and the Natural Sciences and Engineering Research Council of Canada. The authors thank the editors of this special section, and two anonymous reviewers whose critiques and comments on an earlier version greatly enhanced this article.

REFERENCES

Ahl, V., & Allen, T. F. H. (1996). *Hierarchy theory: A vision, vocabulary, and epistemology*. New York: Columbia University Press.

Amalberti, R., & Deblon, F. (1992). Cognitive modeling of fighter aircraft process control: A step towards an intelligent onboard assistance system. *International Journal of Man-Machine Studies*, 36, 637–671.

Banaji, M. R., & Crowder, R. G. (1989). The bankruptcy of everyday memory. *American Psychologist*, 44, 1185–1193.

Barker, R. G. (1969). Wanted: An eco-behavioral science. In E. P. Willems & H. L. Raush (Eds.), *Naturalistic viewpoints in psychological research* (pp. 31–43). New York: Holt, Rinehart and Winston.

Berkowitz, L., & Donnerstein, E. (1982). External validity is more than skin deep: Some answers to criticisms of laboratory experiments. *American Psychologist*, 37, 245–257.

Brunswik, E. (1956). *Perception and the representative design of psychological experiments* (2nd ed.). Berkeley: University of California Press.

Chapanis, A. (1967). The relevance of laboratory studies to practical situations. *Ergonomics*, 10, 557–577.

Chapanis, A. (1988). Some generalizations about generalization. *Human Factors*, 30, 253–267.

Chow, S. L. (1996). *Statistical significance: Rationale, validity and utility*. London: Sage.

Dipboye, R. L., & Flanagan, M. F. (1979). Research settings in industrial and organizational psychology: Are findings in the field more generalizable than in the laboratory? *American Psychologist*, 34, 141–150.

Driskell, J. E., & Salas, E. (1992). Can you study real teams in contrived settings? The value of small group research to understanding teams. In R. W. Swezey & E. Salas (Eds.), *Teams: Their training and performance* (pp. 101–124). Norwood, NJ: Ablex.

Fisher, D. L. (1993). Optimal performance engineering: Good, better, best. *Human Factors*, 35, 115–139.

Hammond, K. R. (1986). Generalization in operational contexts: What does it mean? Can it be done? *IEEE Transactions on Systems, Man and Cybernetics, SMC-16*, 428–433.

Henshel, R. L. (1980). The purpose of laboratory experimentation and the virtues of deliberate artificiality. *Journal of Experimental Social Psychology, 16,* 466–478.

Hoffman, R. R., & Deffenbacher, K. A. (1993). An analysis of the relations between basic and applied psychology. *Ecological Psychology, 5,* 315–352.

Hollnagel, E., Pedersen, O. M., & Rasmussen, J. (1981). *Notes on human performance analysis* (Risø-M-2285). Roskilde, Denmark: Risø National Laboratory.

Holton, G. J. (1986). *The advancement of science, and its burdens.* Cambridge: Cambridge University Press.

Klein, G. A., Calderwood, R., & Clinton-Cirocco, A. (1986). Rapid decision making on the fire ground. In *Proceedings of the Human Factors Society 30th Annual Meeting* (pp. 576–580). Santa Monica, CA: Human Factors and Ergonomics Society.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46,* 806–834.

Mook, D. G. (1983). In defense of external invalidity. *American Psychologist, 38,* 379–387.

Moray, N., Sanderson, P. M., & Vicente, K. J. (1992). Cognitive task analysis of a complex work domain: A case study. *Reliability Engineering and System Safety, 36,* 287–316.

Patterson, E. S., & Woods, D. D. (1997). Shift changes, updates, and the on-call model in space shuttle mission control. In *Proceedings of Human Factors and Ergonomics Society 41st Annual Meeting* (pp. 243–247). Santa Monica, CA: Human Factors and Ergonomics Society.

Rasmussen, J. (1985). The role of hierarchical knowledge representation in decision-making and system management. *IEEE Transactions on Systems, Man and Cybernetics, SMC-15,* 234–243.

Rasmussen, J., Pejtersen, A. M., & Goodstein, L. P. (1994). *Cognitive systems engineering.* New York: Wiley.

Vicente, K. J. (1988). Adapting the memory recall paradigm to evaluate interfaces. *Acta Psychologica, 69,* 249–278.

Vicente, K. J. (1991). *Supporting knowledge-based behavior through ecological interface design.* Unpublished doctoral dissertation, University of Illinois at Urbana-Champaign, Urbana, IL.

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

Vicente, K. J. (1997). Heeding the legacy of Meister, Brunswik, and Gibson: Toward a broader view of human factors research. *Human Factors, 39,* 323–328.

Vicente, K. J. (1999). *Cognitive work analysis: Toward safe, productive, and healthy computer-based work.* Mahwah, NJ: Erlbaum.

Vicente, K. J., & Rasmussen, J. (1992). Ecological interface design: Theoretical foundations. *IEEE Transactions on Systems, Man and Cybernetics, SMC-22,* 589–606.

Webster, M., & Kervin, J. B. (1971). Artificiality in experimental sociology. *Canadian Review of Sociology and Anthropology, 8,* 263–272.

Woods, D. D. (1993). Process tracing methods for the study of cognition outside the experimental psychology laboratory. In G. A. Klein, J. Orasanu, R. Calderwood, & C. E. Zsambok (Eds.), *Decision making in action: Models and methods* (pp. 228–251), Norwood, NJ: Ablex.

Xiao, Y. (1994). *Interacting with complex work environment: A field study and a planning model.* Unpublished doctoral thesis, University of Toronto, Ontario, Canada.

Xiao, Y., Milgram, P., & Doyle, D. J. (1997). Planning behavior and its functional roles in the interaction with complex systems. *IEEE Transactions on Systems, Man and Cybernetics, SMC-27,* 313–324.

Yan Xiao is a human factors engineer and assistant professor of anesthesiology at the Department of Anesthesiology, University of Maryland School of Medicine. He received a Ph.D. in human factors from the University of Toronto in 1994.

Kim J. Vicente received a Ph.D. in mechanical engineering from the University of Illinois at Urbana-Champaign in 1991. He is a professor of mechanical and industrial engineering and of biomedical engineering at the University of Toronto, where he is also director of the Cognitive Engineering Laboratory.