THE EFFECTS OF A COMPUTER-BASED DESIGN AID IN THE
SELECTION OF GUIDELINES WITHIN THE USI DESIGN PROCESS

by:

Jeffrey A. Fox

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of
MASTER OF SCIENCE
in
Industrial and Systems Engineering

APPROVED:  

[Signature]

Robert C. Williges, Chairman

[Signature]
Beverly H. Williges

[Signature]
H. Rex Hartson

[Signature]
Donna L. Cuomo

December, 1991
Blacksburg, Virginia
The Effects of a Computer-Based Design Aid in the Selection of Guidelines
Within the USI Design Process

by
Jeffrey A. Fox

Committee Chairman: Robert C. Williges
Industrial and Systems Engineering
(ABSTRACT)

Designing a User-System Interface (USI) is a complex task and has been approached in many ways. One approach has been to use USI design guidelines to help improve the quality and consistency of USIs. To use guidelines effectively, a general set of guidelines must be tailored to a specific application. One popular set of guidelines is the Smith and Mosier Guidelines for Designing User Interface Software (1986). This thesis investigated the effects of using a computer-based hypertext design aid (DRUID, Dynamic Rules for User Interface Design) for the selection of USI guidelines for both experienced and novice Guidelines users. Also, the relative usability of the software and hardcopy document was examined to improve the interface for the next iteration of DRUID. Both performance and variability in guideline selection strategies were studied within the context of the experimental tasks.

Results indicate that subjects effectively used both the book and hypertext presentation media. However, subjects selected more relevant guidelines with the book. The media influenced the subjects' selection strategies, and the subjects read
less when using DRUID. Subjectively, the software was preferred because it provided assistance in the selection process. The presentation of the guidelines could be improved for both media, and many of the factors which led subjects to select fewer relevant guidelines with DRUID could be overcome with improved USI design and implementation. Also, to overcome the limitations of presenting large texts on-line, "value added" features should be incorporated into on-line tools to help users better perform their tasks.
ACKNOWLEDGMENTS

This effort was in part supported by The Human Factors Engineering for User System Interface Specialty Group at The MITRE Corporation and the Electronic Systems Division (ESD) of United States Air Force. ESD funded the development of the DRUID (Dynamic Rules for User Interface Design) project and provided the opportunity to evaluate it. I would like to thank Dr. Donna L. Cuomo, my group leader at MITRE, for all her advice and support in the roles of colleague and thesis committee member. I would also like to thank Nancy Goodwin from the specialty group for her support in sustaining the effort.

I would like to express my thanks to Dr. Sidney L. Smith and Jane Mosier for their participation as the "guideline experts" in the study. I would also like to thank Dr. Smith for taking me under his wing four years ago and being my mentor.

A special thanks Dr. H. Rex Hartson and Beverly H. Williges for serving on my thesis committee. It was a pleasure having the opportunity to learn from and work with them. I would especially like to thank Dr. Robert C. Williges for accepting me as one of his students and chairing my committee. His invaluable insights and experience provided clear guidance throughout my graduate career.
# TABLE OF CONTENTS

INTRODUCTION ............................................................................................................. 1  
Design Guidelines and USI Design ............................................................................. 1  
The Smith and Mosier Guidelines ............................................................................. 2  
Who Uses Guidelines and How They Have Been Used ......................................... 3  
The Role of Guidelines ............................................................................................... 4  
Creating Design Rules ............................................................................................... 5  
  Step 1 ......................................................................................................................... 6  
  Step 2 ......................................................................................................................... 6  
  Step 3 ......................................................................................................................... 7  
Guidelines in System Acquisition ............................................................................ 10  
  Requirements Phase ................................................................................................. 11  
  Preliminary Design Phase ....................................................................................... 12  
  Detailed Design Phase ............................................................................................ 12  
  Development Phase .................................................................................................. 13  
  Test and Evaluation ................................................................................................. 13  
The Problem with Guidelines ................................................................................... 14  
Guideline Tools .......................................................................................................... 15  
Are On-Line Guidelines Useful? .............................................................................. 16  
Purpose ......................................................................................................................... 17
METHOD

Operational Definitions
Experimental Design
Subjects
Apparatus
Software
Computer Hardware
Additional Materials
Task Scenario Descriptions
Guideline Selection Procedure
Phase I
Phase II

RESULTS

Phase I
Phase II
Critical Measures
Efficiency
Other Guideline Usage Measures
Correlation
Subjective Measures
Pre-test Questionnaire
Post-test Questionnaire
Summary Questionnaire
Drawings
Appendix H: Summary Questionnaire ................................................................. 95
Appendix I: List of Guidelines Selected by Experts .............................................. 99
Appendix J: ANOVA Summary Tables ............................................................... 102
Appendix K: Scenario Interface Drawings .......................................................... 108

VITA ....................................................................................................................... 125
LIST OF FIGURES

Figure 1. Sample guideline.................................................................3
Figure 2. Factors affecting guideline selection....................................7
Figure 3. DRUID screen showing input methods..................................23
Figure 4. Selecting functions in DRUID............................................24
Figure 5. Selecting guidelines from a guideline title list......................24
Figure 6. Selecting a guideline from its own display...........................25
Figure 7. 3D plot of relative guideline selection for ACME scenario.....38
Figure 8. 3D plot of relative guideline selection for Widgets scenario...39
LIST OF TABLES

TABLE 1. Functional Areas Within Smith and Mosier Guidelines (1986) ....2
TABLE 2. Types of Guidelines Users..................................................................3
TABLE 3. Guideline Usage.................................................................................4
TABLE 4. Structured Design Process..................................................................11
TABLE 5. Presentation Medium versus Experience Level.................................20
TABLE 6. Balance of Experimental Conditions.................................................20
TABLE 7. Decision Rules for Determining the Number of Subjects....................21
TABLE 8. Guideline Usage Means.................................................................33
TABLE 9. Guideline Usage ANOVA Results.....................................................34
TABLE 10. Contingency Table of All Possible Combinations of Subjects and
          Experts Picking any Guideline............................................................35
TABLE 11. Contingency Table for ACME........................................................36
TABLE 12. Contingency Table for Widgets......................................................36
TABLE 13. Post-test Scores for Media..............................................................41
TABLE 14. Post-test Scores for Experience Level..............................................42
TABLE 15. Least Correlated Cells for the ACME Contingency Table.................59
TABLE 16. Least Correlated Cells for the Widgets Contingency Table.............59
INTRODUCTION

Design Guidelines and USI Design

Good user-system interface (USI) design has become increasingly important as more system code is devoted to the USI and as the population of computer users continues to expand. At the same time, the USIs now being developed (and those on the drawing board) are becoming more complex. To help the user interface practitioner create, manage, and communicate USI designs, numerous models, techniques, and tools have been developed.

Russell and Galer (1987) proposed that an ideal environment for human factors in USI design is in fact a combination of:

Human Factors Practitioner + Methodologies + Tools

They point out that each of these elements is important in the design of USI software. USI design guidelines are one such tool developed to help the human factors expert, but they can also be used by a non-expert. In fact, Shneiderman (1988) considers guideline documents as one of the "three pillars" of user interface development (the others being user interface management systems and usability testing).

Smith (1986) describes guidelines as a series of generally worded recommendations that are used for guidance (as their name implies) in the USI design process. It is important to distinguish guidelines from standards. Standards are generally worded requirements that must be followed in a design (often contractually). For an in-depth discussion of the differences between the two types of reference material see Smith (1986).
The Smith and Mosier Guidelines

There are many sets of guidelines available today. Some are proprietary (for in-house use only) and some are publicly available. Several of the larger and better known sets of guidelines include those by Brown et al. (1983), Engel and Granda (1975), and Smith and Mosier (1986). This thesis studies the use of the most comprehensive set of guidelines, the Smith and Mosier Guidelines for Designing User Interface Software (1986). That report contains 944 individual guidelines classified into six functional areas of USI design, as shown in Table 1 below.

<table>
<thead>
<tr>
<th>Functional Area</th>
<th>Number of Guidelines</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Entry</td>
<td>199</td>
</tr>
<tr>
<td>Data Display</td>
<td>298</td>
</tr>
<tr>
<td>Sequence Control</td>
<td>184</td>
</tr>
<tr>
<td>User Guidance</td>
<td>110</td>
</tr>
<tr>
<td>Data Transmission</td>
<td>83</td>
</tr>
<tr>
<td>Data Protection</td>
<td>70</td>
</tr>
</tbody>
</table>

Within each functional area, the guidelines are sub-divided into more specific functions. For example, three of the nine of the functions listed under "Data Entry" are "Position Designation," "Text," and "Graphics." Therefore, the Guidelines' structure allows a user to find specific guidelines based on the functional and task-specific requirements.

An individual guideline in Smith and Mosier includes a reference number, a title, and the guideline statement (a single sentence). Most of the guidelines are also annotated with information to help in understanding and applying them. The annotation can include examples, comments, exceptions, references and cross-references. A typical Smith and Mosier guideline for data display tables follows:
2.3.7 Consistent Label Format
Adopt a consistent format for labeling the rows and columns of displayed tables.

EXAMPLE: Each column label might be left-justified with the leftmost character of the column entries beneath it.

COMMENT: Consistent left justification of column labels will prove especially helpful when columns vary in width.

REFERENCES: Hartley, Young, and Burnhill, 1975.

SEE ALSO: 2.0.6, 4.0.6.

---

Figure 1. Sample guideline.

Who Uses Guidelines and How They Have Been Used

Despite the overwhelmingly size of the Smith and Mosier Guidelines, it remains a popular reference with well over 1000 copies in circulation. The Guidelines have been used by a variety of professions in numerous ways. A survey published by Mosier and Smith (1986) provides an in-depth account of guideline usage. That 1986 survey showed the following breakdown of the 130 respondents, as shown in Table 2.

TABLE 2. Types of Guidelines Users

<table>
<thead>
<tr>
<th>Occupation</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>human factors specialist</td>
<td>51</td>
</tr>
<tr>
<td>system analyst</td>
<td>15</td>
</tr>
<tr>
<td>software designer</td>
<td>10</td>
</tr>
<tr>
<td>manager</td>
<td>8</td>
</tr>
<tr>
<td>researcher</td>
<td>8</td>
</tr>
<tr>
<td>other users</td>
<td>4</td>
</tr>
<tr>
<td>system engineer</td>
<td>3</td>
</tr>
<tr>
<td>student</td>
<td>1</td>
</tr>
</tbody>
</table>
How the *Guidelines* were used also varied. Table 3, below, shows the breakdown of guideline usage as reported by Mosier and Smith (1986) (some users reported more than one use).

**TABLE 3. Guideline Usage**

<table>
<thead>
<tr>
<th>Uses for Guidelines</th>
<th>Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>to establish requirements in advance of design</td>
<td>46</td>
</tr>
<tr>
<td>to aid during design</td>
<td>62</td>
</tr>
<tr>
<td>to evaluate a proposed design</td>
<td>41</td>
</tr>
<tr>
<td>to evaluate a completed design</td>
<td>25</td>
</tr>
<tr>
<td>to evaluate an operational system</td>
<td>18</td>
</tr>
<tr>
<td>other uses</td>
<td>24</td>
</tr>
</tbody>
</table>

Those statistics illustrate that there are a number of ways to use design guidelines. The time (in the lifecycle) and method of using guidelines may be affected by the following factors: attitudes towards guidelines, familiarity with guidelines, stage in design process, design goals, applicability of guidelines, usability of guidelines, lack of recommended procedures, and the size of system being developed or evaluated.

*The Role of Guidelines*

Some USI designers and researchers have criticized the limited usefulness of guidelines and standards in systems design (Gould and Lewis, 1985; Reaux and Williges, 1988; Thovtrup and Neilsen, 1991; and Woods and Eastman, 1989). Most of the criticisms deal with the user's ability (or inability) to properly interpret and/or apply general guidelines to specific applications during systems design.

Though design guidelines are not perfect, they can provide the basis for a structured approach throughout the system lifecycle. However, guidelines are
designed to provide guidance only and should not be directly used as the baseline for a detailed design specification. Guidelines are intended to be applicable to a wide variety of systems and situations, so they must be robust enough to allow a fairly high degree of interpretation.

Questions also have been raised about the validity of the data on which guidelines are based. In Smith and Mosier, guidelines were based on experimentally proven data whenever possible. Unfortunately, much of the needed data for USI design issues has not yet come out of the labs or academia. Smith (1986) points out that all of the data needed for "proven" guidelines for complex systems probably could never be derived from experimentation due to the complexities (number of variables and their confounds) involved in such experiments. Unfortunately, USI designs must be created now, before "all the facts are in." Thus, until concrete data are available, Smith states that guideline development must, to some degree, rely on lessons learned, relevant design experience, and professional judgement as the basis for some design guidelines. However, guidelines do need to be updated frequently to utilize new data as they become available.

Guidelines, as proposed here, are a single set of tools in the USI designer's "bag of tricks." The use of guidelines should be only a single step in the system design process, not a replacement for any of the other important human engineering tasks such as needs analysis, task analysis, and usability testing (Eason 1983; Potter, Cook, Woods, and McDonald, 1990; and Tijerina, 1986). Rather, guidelines need to be used in conjunction with other USI design tools and techniques.

Creating Design Rules

Because of the factors discussed in the previous section, guidelines must be "tailored" to be used effectively for any specific USI. Tailoring is the process of
creating application- or system-specific sets of detailed USI specifications (or rules) from the entire set of design guidelines. The goals of tailoring guidelines are: (1) to get human factors principles into the USI design in an organized manner, (2) to help ensure design consistency, (3) to provide the software developers with a specification that will be relatively easy to implement, and (4) to provide a description to test against the USI during the design process and after design completion. Most of the procedures presented here are based on work accomplished by Mosier (1988) at The MITRE Corporation.

**Step 1.** The first step in creating design rules is reviewing (at least browsing) the guidelines. The intent of this step is to become familiar with the material in that document. By understanding the organization of the book and how the guidelines are stated, a designer can decide how best to use the guidelines.

**Step 2.** The second step is selecting a subset of guidelines that apply to a particular system design. For any particular USI design, only a subset of the guidelines will apply. As Steinbach and Zoltan-Ford (1990) point out, the idea that "the more guidelines followed, the better [the design will be]" is not true. For example, a designer who is creating a USI without any graphics can ignore those guidelines that specifically deal with graphical aspects of the USI. A designer may also note that some guidelines conflict under certain circumstances, so care must be taken to resolve conflicts. One additional reason for selecting only a subset of guidelines is that some guidelines may recommend designs that may either be impossible to create with a given hardware/software suite or that may just be too costly to implement for a limited improvement in the design.

Knowledge of the user population, task and environmental factors, and the characteristics of the target hardware and software platforms are all factors in
translating guidelines for a particular system. Figure 2 shows an expanded list of factors to be considered when selecting which guidelines apply to a specific design.

![Diagram](image)

**Figure 2.** Factors affecting guideline selection.

**Step 3.** After a subset of guidelines has been selected, the third step requires translating the guidelines into specific design rules for a particular application. The design rules are statements that can be directly implemented by the software developers without interpretation. To create the rules, the USI designer must interpret the intent of the selected guidelines and then decide how they apply to the particular system. In making such decisions, the designer should consider any design constraints or alternatives. Again, at this step, knowledge of the user population, task and environmental factors, and the characteristics of the target hardware and software platforms are all factors in the guideline translation for a particular system.

For example, assume a USI designer is interested in specifying how a system should incorporate color coding. The designer would first select those guidelines that relate to color coding. In Smith and Mosier, there are a number of guidelines that
provide information on color coding. One such guideline states that "[A designer should] employ color coding conservatively, using relatively few colors and only to designate critical categories of displayed data." In other words, the overuse of colors is detrimental to USI design, and colors that are used should code specific categories of data. Another related guideline suggests that a designer should "Choose colors for coding based on conventional associations with particular colors." The intent of that guideline is to remind the designer that if particular colors have special meanings to the user population, those colors should have the same meaning in the USI software.

Continuing the example, also assume that the target population for that software currently uses the color green to mean "normal," yellow to mean "warning," and red to mean "danger." For that system, the USI will be implemented on a workstation that supports high resolution graphics with up to 256 colors. Based upon this knowledge, the selected guidelines could be translated into the following representative subset of rules:

The USI software shall use four colors.

Red shall only be used to indicate danger conditions.

Yellow shall only be used to indicate warning conditions.

Green shall only be used to indicate normal conditions.

Black shall be used for standard data display and data entry.

It should be stated here that there is no "cookbook" formula for translating guidelines into rules, just as there is none for choosing guidelines. Smith (1986) makes the analogy that creating design rules is similar to translating mission and functional requirements for any part of the system into detailed design requirements (specifications). The translation of the guidelines into rules is also based on experience and expertise. The USI designer who generates the rules should also take
advantage of the knowledge of the end-users and of the software engineers who will
develop the system by including them in a team that develops the design rules.

Though there are no formulas for creating rules, design rules should possess
certain characteristics. Design rules should be as specific as possible and should be
easy to follow. The rules should be directly usable by the software engineers who
will implement them. And of course, the rules should accomplish the human
factors and systems engineering goals.

As shown above, one guideline can produce multiple rules. Sometimes there
are one-to-one translations and other times several guidelines taken together will
result in only one rule. To date, no study has been conducted on the number of rules
translated from one or more guidelines, but probably one of the largest factors is at
what level of specificity (how concretely) a guideline is stated. The guideline that
states "Compose sentences in the active rather than passive voice" is very specific and
can be implemented with little tailoring. On the other hand, the guideline that states
"Ensure that computer-generated displays of textual data, messages, or instructions,
will generally follow design conventions for printed text" requires interpretation and
could possibly be translated into numerous rules.

In a related study, Reaux and Williges (1988) investigated the effects of
guideline specificity and its effect on how well guidelines were used to evaluate
designs. In that study, Reaux and Williges dichotomized a subset the 1984 version of
the Smith and Mosier Guidelines into those that were concrete and those that were
abstract. They found that subjects could more effectively determine design violations
of concrete guidelines. Also Tetzlaff and Schwartz (1991) found that designers
sometimes had difficulty in interpreting general guidelines.

Design rules help elevate these problems. It should also be noted that
guideline selection and rule generation is an iterative process. It is quite unlikely that
the designers will select several hundred guidelines and then translate them into rules without making some mistakes (e.g., missing some guidelines or rules, wording rules too vaguely, or creating rules that cannot be implemented). Also, rules and guidelines may change because of changes in design requirements or functionality.

*Guidelines in System Acquisition*

To be used effectively, guidelines should be used in concert with the entire system engineering effort throughout a project's lifecycle. There are many models that have been presented for software and USI design, ranging from strict linear steps to ad hoc development (Boehm, 1988; Lundel and Notess, 1991; Mantei and Teorey, 1988; Thimbleby, 1984). The method/model described here attempts to take advantage of the structure of the waterfall model of system design while incorporating some of the benefits of the spiral design model. This is the general model that is encouraged by the Electronic Systems Division (ESD) of the United States Air Force and The MITRE Corporation in acquiring USI software. For a more detailed presentation of the model, see Goodwin (1988). The ESD/MITRE model, and variants of it, are now used widely throughout the Department of Defense and other contracting agencies. It is part of this model that will be studied in the thesis.

A summary of the structured ESD/MITRE design process is shown in Table 4. That table shows a simplified view of the ESD/MITRE model by breaking it into four major phases: requirements analysis, design (preliminary and detailed), development, and test and evaluation. In each phase, the table displays some representative USI tasks (not exhaustive), guideline-related tasks, and guideline outputs. It should be noted that it is often necessary to combine or eliminate some of these steps because of limiting constraints such as time or available funds. The model is simplified for concise presentation. It should be noted that there is iteration
and feedback within (and between) the stages of the process. Also, other techniques such as prototyping and user group meetings are used in conjunction with the guidelines.

**TABLE 4. Structured Design Process**

<table>
<thead>
<tr>
<th>Phase</th>
<th>USI-related Tasks</th>
<th>Guideline Tasks</th>
<th>Guidelines Output</th>
</tr>
</thead>
<tbody>
<tr>
<td>Requirements Analysis</td>
<td>1) Perform needs analysis</td>
<td>Review and select functions</td>
<td>Functional Requirements Checklist</td>
</tr>
<tr>
<td></td>
<td>2) Perform functional and task analyses</td>
<td>Review and select critical guidelines</td>
<td>Preliminary Design Checklist</td>
</tr>
<tr>
<td>Design (Preliminary)</td>
<td>3) Create initial design</td>
<td>Update function selection and update guideline selection</td>
<td>Revised Functional Requirements Checklist</td>
</tr>
<tr>
<td></td>
<td>4) Develop low fidelity prototype</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Design (Detailed)</td>
<td>5) Formalize design</td>
<td>Update guidelines</td>
<td>Design Requirements Checklist</td>
</tr>
<tr>
<td></td>
<td>6) Create detailed design</td>
<td>Translate guidelines into rules</td>
<td>Detailed Design Specification</td>
</tr>
<tr>
<td></td>
<td>7) Develop medium to high fidelity prototype</td>
<td>Update design rules</td>
<td>Revised Design Specification</td>
</tr>
<tr>
<td>Development</td>
<td>8) Develop system</td>
<td>Follow design rules and update where necessary</td>
<td>Updated Design Specification</td>
</tr>
<tr>
<td>Test and Evaluation</td>
<td>9) Evaluate design</td>
<td>Test against design rules</td>
<td>Design Evaluation Checklist</td>
</tr>
</tbody>
</table>

**Requirements Phase.** Systems design must begin with proper up-front analyses, such as needs assessment, functional analysis, function allocation, and task analysis. From such analyses, data on operational concepts, detailed requirements, design constraints, usability data, training needs, etc. should be developed as the basis for the design. This paper does not intend to explain those processes or claim to provide a complete list of available analyses. The intent here is to show that the
designer needs to perform the classic analytic techniques before selecting guidelines for a design.

Once functions have been determined and allocated, the USI designer can select those USI functions that apply to the system (e.g., Data Display Tables). That list of applicable functions is called a "Functional Requirements Checklist."

**Preliminary Design Phase.** The initial design concepts are created in the preliminary design phase. During this phase, the USI designer should review the individual guidelines and begin tailoring the guidelines to the system being designed. Tailoring includes creating a first-draft subset of guidelines for the USI design. The list of selected guidelines is the "Preliminary Design Checklist." For smaller systems or when requirements are well established, the function selection can also occur at this step, and the results incorporated into the Preliminary Design Checklist.

After creating that checklist, it is helpful to use a prototyping tool in evaluating the guideline selection. The prototype needs only to be of low fidelity, even paper and pencil, to show the basic concepts of the design. Quite likely, the prototype will help the USI designer decide if the proper guidelines were selected or if the preliminary list needs to be modified. Based on the evaluation of the prototype (or any other method of design review), a revised Preliminary Design Checklist should be generated. An entire function and its guidelines may have been missed or should not have been included. If so, the Functional Requirements Checklist should also be updated.

**Detailed Design Phase.** During the detailed design phase, the USI design is solidified. Any changes in the system requirements or functionality should involve updating the selected guidelines. Once a final set of guidelines has been selected, those guidelines should be recorded in a "Design Requirements Checklist". 
The design rules for a system are based on the guidelines contained in the Design Requirements Checklist. The design rules should be generated in the manner recommended earlier. The list of design rules that are the basis for the design is the "Detailed Design Specification."

After the initial set of rules has been generated, it is useful to create a medium to high fidelity prototype to refine and validate the design rules. The rules should be updated based on the prototype design, and the Design Specification should be revised accordingly. After revising, adding, or subtracting any necessary rules, the prototype may be updated to reflect the rule changes. This iterative process should continue until the final rule set describes the design in enough detail so that a software engineer can build the user interface software from that set. The final set of rules can then be "frozen" as the Design Specification to use in developing the system.

*Development Phase.* During the actual development of the system, it is quite likely that the list of rules may change. Hardware and/or software configurations may change, new features may be added, or existing ones taken away. When such changes occur, a formal procedure for changing the design rules must be established and followed. All parties affected by the change must be informed of any changes and their potential impact. The Design Specification must also be updated and may need to be approved by the proper oversight authorities.

*Test and Evaluation.* During the development of the system, design reviews of the USI can be conducted to ensure that the rules are being properly applied and followed. For this and subsequent evaluations and tests, the Design Specification can be used as a "Design Evaluation Checklist" by adding an extra column to mark design compliance with each rule. Thus, the Evaluation Checklist can be used also to provide quality control during system development.
At the completion of system development, the Evaluation Checklist can be used for formal in-depth test and evaluation to determine if the software meets the specifications. The checklist can also be used to establish the fulfillment of contractual obligations.

The Problem with Guidelines

Unfortunately there are some acknowledged problems with the approach described above (Chapanis and Budurka, 1990; Smith and Mosier, 1986). Because there are no formulae or defined sequences of steps to follow, there will undoubtedly be variability in the guideline selection. Any two USI designers will most likely pick different sets of guidelines and rules for any given system. USI designers must apply their knowledge of the factors presented in Figure 1 on a case-by-case basis in selecting guidelines. However, all variability should not necessarily be considered bad. The selection provides for individual creativity within a structured process.

Additionally, designers must draw upon their own knowledge of systems and USI design in this process. Rouse and Boff (1987) note that in design, it is not only the nature of the problem, but also the nature of the designers, effects of the organization, available support tools, and the acceptance of the tools that are factors that come into play in the design process.

Compounding these problems is the sheer size of the Smith and Mosier Guidelines. Altogether, the Smith and Mosier report is 478 pages long, of which approximately 375 pages are devoted to the guidelines. The remaining pages provide aids in understanding, accessing, and applying the material. Despite the human engineering of that document, any reference document of that size becomes cumbersome to use.
Guideline Tools

Many practitioners and researchers have called for on-line tools to provide better access to and better utilization of large guideline documents (Normore and Tijerina, 1986; Smith and Mosier 1984; Perlman, 1989b; Tijerina, 1986; Thovtrup and Neilsen, 1991; Woods and Eastman, 1991). Several attempts have already been made to address the issue of the size of the Guidelines. To aid in the access and application of that document, several hypertext tools have been developed on different platforms. RIPL is on the DEC VAX (Reaux and Williges, 1988), NaviText SAM is on the IBM PC (Perlman, 1987), and Dynamic Rules for User Interface Design (DRUID) is on the Macintosh (Fox and Smith, 1989a; Fox and Smith, 1989b; Fox, 1991).

Among its other capabilities, RIPL allows users to browse the Guidelines online. In addition to presenting a hypertext version of Smith and Mosier, DRUID and NaviText SAM provide aids that allow users to tailor the guidelines and create online checklists. Both systems provide database-like capabilities which allow users to select pertinent guidelines, store those guidelines, and generate guideline checklists. One distinguishing factor of DRUID is that it not only provides command line input, but also provides direct manipulation hypertext functionality. Also, DRUID encourages the ESD/MITRE guideline selection process and generates the checklists described in the previous section.

A still more useful tool would be an on-line assistant with expert knowledge of the guidelines and how they can be applied (Reaux and Williges, 1988; Smith, 1988; Tijerina, 1986). Such a tool could work with a designer to improve the guideline selection process. The advantage of such a tool would be that it would combine the tool’s knowledge of the guidelines with the designer’s knowledge of the USI design and the unique factors concerning the system’s implementation.
Are On-Line Guidelines Useful?

Before more sophisticated on-line assistants can be built, several questions need to be answered. First, does placing design guidelines on-line aid in their usability? In comparing RIPL to the printed guidelines, Reaux and Williges (1988) showed that when using guidelines for prototype evaluation, there was no overall significant difference in task completion time or task performance based on presentation media. However, Reaux and Williges did find that the subjects who used the on-line version of the Guidelines spent more time navigating through and reading the guidelines (instead of analyzing the USI) than did the hard-copy users. This discouraging result may possibly be attributed to the particular implementation (screen design, access time, browsing facilities, etc.) and to typical CRT readability factors.

On the other hand, Perlman has performed a number of studies comparing NaviText SAM to the paper report. Perlman and Moorehead (1988) reported that "for equivalent tasks with similar outputs, SAM is much faster than working with paper." Unfortunately, neither the specific tasks nor the numeric results were published.

A second question (really set of questions) is whether placing guidelines on-line will influence the designers performance in selecting guidelines or their strategies for guideline selection. In the studies conducted by Reaux and Williges (1988) and Perlman (1989a), users of both systems changed their strategies on accessing the material (e.g., percent of time using the index, table of contents, etc.) when using the software systems.

The changes may be attributed to several factors. The largest factor was probably the subjects that were used in the studies. Reaux and Williges used college students (computer science majors with less than three months of classroom USI design experience) who had previous experience with the Guidelines. Perlman did
not identify his users. The experimenters noted that their results may differ with USI experts and recommended further studies in this area. Additionally, the results may also vary with the level of experience with using guidelines for the tasks under study.

Another factor is the tasks that have been studied. Perlman (1989a) tested the speed with which users could find guidelines relevant to a specified topic (i.e., information retrieval). Reaux and Williges tested the use of a small subset of the Guidelines for design evaluation. To date, there has been no published study on using the Guidelines for USI design and requirements definition, which Smith and Mosier (1986) found to be one the largest areas of guideline usage. It is quite likely that the influence of the presentation medium (hardcopy or software) may depend on the task.

Purpose

The greatest benefit of having an on-line tool incorporating the Guidelines would be to help the USI designer better utilize the material in the systems design process. It is widely acknowledged that for human factors to be effective in system engineering, it should be incorporated from the beginning of the system lifecycle. The earlier in the design that USI principles and guidelines can be applied, the more impact they will have on the final system. Thus, an on-line tool that can provide effective support in the design phase, as well as throughout the remainder of a system's lifecycle, could greatly improve that system's USI design and success.

This thesis studies how USI designers use the Smith and Mosier Guidelines in USI design scenarios developed specifically for this experiment and the impact of using an on-line tool, DRUID, in aiding guideline selection. Data are collected in an attempt to isolate media-related factors that influence performance when using the
software. Collecting such data could then be used as a summative evaluation of DRUID to improve the design of the presentation media for future updates.

The research also examines how the participant's level of experience with the Guidelines influences how well the Guidelines and DRUID serve as a design aid. DRUID was selected as the tool for this study because it is the most advanced computer implementation of the Guidelines, offering a modern graphical interface to that material. DRUID maintains the hardcopy's structure and organization. Also, DRUID was designed with the aid of the Guidelines' authors and presents the entire contents of that document. Additionally, the thesis investigates the variability in guideline selection by determining the correlation between guidelines selected by a panel of Guidelines experts and by subjects (both experienced and novice).
METHOD

Operational Definitions

This section describes some of the terms used in the remainder of this document. Design Guidelines are generally stated recommendations for user interface software. Guidelines are based on principles, scientific data, and experience. Often, guidelines are accompanied by examples, added explanations, or other commentary. For this experiment, the guidelines are from Smith and Mosier (1986).

Design Rules are design specifications for a particular application. The design rules are usually written so that they can be directly implemented by software engineers on a particular project without interpretation. Design rules should be based upon established guidelines, principles, and system knowledge.

Design Checklists are lists of design guidelines or rules that form a basis for USI design specification or evaluation.

Experimental Design

The research was conducted using a 2x2x2 mixed-factors design. The presentation media (the printed Guidelines and DRUID) and the two similar scenarios (USI descriptions) are the within-subjects factors. The level of guideline experience (novice or experienced) is the between-subjects factor. Table 5 shows the experimental design. Each subject selected guidelines about one scenario from the hardcopy document and selected guidelines for the other scenario from DRUID. Table 6 shows the counterbalancing used in the experimental design.
TABLE 5. Presentation Medium versus Experience Level

<table>
<thead>
<tr>
<th>Level of Guideline</th>
<th>Presentation Medium</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Hardcopy</td>
</tr>
<tr>
<td>Experienced</td>
<td>4</td>
</tr>
<tr>
<td>Novice</td>
<td>4</td>
</tr>
</tbody>
</table>

TABLE 6. Balance of Experimental Conditions

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Experienced Users</th>
<th>Novice Users</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>Trial 1 DRUID - Scenario 1</td>
<td>Trial 1 DRUID - Scenario 1</td>
</tr>
<tr>
<td></td>
<td>Trial 2 Hardcopy - Scenario 2</td>
<td>Trial 2 Hardcopy - Scenario 2</td>
</tr>
<tr>
<td>B</td>
<td>Trial 1 DRUID - Scenario 2</td>
<td>Trial 1 DRUID - Scenario 2</td>
</tr>
<tr>
<td></td>
<td>Trial 2 Hardcopy - Scenario 1</td>
<td>Trial 2 Hardcopy - Scenario 1</td>
</tr>
<tr>
<td>C</td>
<td>Trial 1 Paper - Scenario 1</td>
<td>Trial 1 Paper - Scenario 1</td>
</tr>
<tr>
<td></td>
<td>Trial 2 DRUID - Scenario 2</td>
<td>Trial 2 DRUID - Scenario 2</td>
</tr>
<tr>
<td>D</td>
<td>Trial 1 Paper - Scenario 2</td>
<td>Trial 1 Paper - Scenario 2</td>
</tr>
<tr>
<td></td>
<td>Trial 2 DRUID - Scenario 1</td>
<td>Trial 2 DRUID - Scenario 1</td>
</tr>
</tbody>
</table>

Note: There are four different treatment combinations per experience level. Each treatment contains two conditions in a specific order. Because of the nature of this study, the first condition dictates the second. Thus, there are only four unique sets of treatments, and the experiment cannot be completely counterbalanced.
Subjects

Due to the difficulty and expense of running experienced USI designers, the experimental design was set-up to minimize the total number of required subjects. There were four subjects at each experience level. After the initial set of eight subjects were run, the following predetermined "critical" dependent measures were tested for their level of significance using analysis of variance (ANOVA) at the 0.05 level. These tests were used to determine whether it would be necessary to run a second set of eight subjects.

- Time to complete task - Time spent reviewing and selecting core guidelines
- Total number of core (relevant) guidelines selected
- Guideline Usage - Number of guidelines selected from the title lists versus individually

Based on the predefined decision rules shown in Table 7, it was determined that a second set of 8 subjects (4 per level) was not required. The actual data are shown in the results section.

<table>
<thead>
<tr>
<th>Initial Result</th>
<th>Rule</th>
</tr>
</thead>
<tbody>
<tr>
<td>If 2 or more of the critical measures are within 0.05 of the predefined alpha level of 0.05 then</td>
<td>Continue experiment - run 8 additional subjects</td>
</tr>
<tr>
<td>If 2 or more of the critical measures are either significant at the alpha level or exceed the alpha level by more than 0.05 then</td>
<td>Terminate experiment - do not run additional subjects</td>
</tr>
</tbody>
</table>

All subjects in the experiment had at least one year of experience in USI design (outside of academic training) and at least some hands-on experience using the
Macintosh computer. Subjects included human factors practitioners, systems analysts, and software engineers. All the experienced guideline users had previously used the Guidelines in the design and/or evaluation of a USI design. The range of previous experience was controlled to include subjects who have previously used the Guidelines at least 5 times in the design or evaluation of a USI. Designers who have never used the Guidelines before the experiment, or who only had cursory exposure to the Guidelines before the experiment, were assigned to the novice group. All subjects were either MITRE employees or Air Force personnel from ESD who work with MITRE in USI design activities. One subject was eliminated from the experiment because that person was not properly able to understand the task scenario requirements. Since being able to understand task requirements was critical for performing the guideline selection task, that subject was replaced with another subject.

Subject selection and assignment was made from the results of a pre-test questionnaire (Appendix A). The questionnaire was also used to collect other relevant information, such as USI and system design experience and USI training.

Apparatus

Software. A beta version of DRUID 2.0 was used as the computer-based guideline presentation medium. DRUID is a tool developed to improve the design and acquisition of user interface software by providing aids for specifying guidelines. DRUID is based on, and contains, the entire contents of the Smith and Mosier Guidelines for Designing User Interface Software (1986).

DRUID provides the capability to review design guidelines on-line. DRUID provides a unique method of accessing the Guidelines by offering both command line and direct manipulation access to that material. A printout of the display that shows
DRUID's contents illustrates the multiple input methods (Figure 3). As shown in this figure, the user can navigate to the functional area "2 Data Display" either by typing that area's number "2" or by using the mouse to select that option from the menu list.

![Guideline Contents (SC) diagram]

**Figure 3.** DRUID screen showing input methods.

Beyond that hypertext functionality, DRUID allows users to select and weight the importance of relevant functions and guidelines for a system design application. In other words, DRUID provides on-line checklist capabilities. DRUID provides the user with the ability to select functions by marking those that are desired, as shown in Figure 4.
Likewise, a DRUID user can tailor the entire set of guidelines by marking individual guidelines. Within DRUID, a user can select guidelines from an abbreviated list of all the guidelines within a given function (Figure 5) or directly from an individual guideline display (Figure 6).

Figure 4. Selecting functions in DRUID.

Figure 5. Selecting guidelines from a guideline title list.
Figure 6. Selecting a guideline from its own display.

DRUID also provides the capability to rate design compliance with the guidelines. DRUID can then output the sections (checklists) in a variety of hardcopy and softcopy checklist formats. DRUID currently runs under HyperCard™ 2.1 on the Macintosh™ computer.

Computer Hardware. DRUID was run on a Macintosh IIcx.

Additional Materials. A hardcopy of the Smith and Mosier Guidelines was provided to the subjects only when that presentation medium was being studied. When DRUID was tested, a copy of the DRUID 2.0 Users' Manual was provided for reference.

One video camera was used to record the test sessions. The camera was focussed on the computer screen when DRUID was used, and on the hardcopy Guidelines when that medium was used.

Data logging software was used to record frequency data. The experimenter also collected additional data (such as task strategies and comments) with paper and pencil.
Task Scenario Descriptions

For this experiment, the subjects selected guidelines they thought relevant for two individual systems descriptions (scenarios). One description is about the requirements for a computerized personnel information system for the personnel director of ACME Tool Company. The other description is about a computerized sales analysis system for the head of marketing of American Widgets, Inc. Those two scenarios are presented in Appendix B. The scenarios provided the subjects with the minimum necessary requirements from which a USI could be designed. The scenarios contain information on the users, the task requirements, the environment, and any hardware or software constraints. The two scenarios were designed to be approximately equivalent in terms of difficulty in understanding the USI requirements and the number of relevant guidelines.

Guideline Selection Procedure

Phase I. The initial phase of the experiment was conducted to develop a baseline set of necessary (relevant) guidelines for each task scenario description. The baseline guidelines were selected by three guideline experts. Two of the experts were the authors of the 1986 version of the Guidelines, Sidney L. Smith and Jane N. Mosier. The other expert was Jeffrey A. Fox, the author of the DRUID software. Each expert read and signed the consent form presented in Appendix C.

The three experts then independently selected the guidelines that they judged to be relevant for each task scenario. The experts used the DRUID software to select all of the guidelines. Those guidelines that were selected by all three experts were considered "core" guidelines. Requiring a complete consensus makes the core set of guidelines conservative.
Those core guidelines were then used as the "necessary" or "mandatory" set of guidelines for the second phase of the experiment. Likewise those guidelines that were not selected by any of the experts were considered not applicable or relevant to the scenarios. The core guidelines form the basis for evaluating the subject's selection in Phase II.

Phase I was also used to refine the task scenario descriptions. Changes were made to the descriptions based on the relative number of guidelines selected and comments and opinions of the experts. The goal was to ensure that each task description had all the necessary data for the subjects to perform their tasks, and to ensure that both tasks are roughly equivalent, as defined above.

Phase II. Pilot studies were conducted to refine the experimental procedures. Participants in the pilot studies performed the full experiment as described below. The pilot subjects were also used to test the data collection methodology and to improve the questionnaires and scenario descriptions as necessary.

All subjects first read and signed the consent form presented in Appendix C. Next, each subject received a training presentation on USI guidelines and their applications. That presentation was conducted by the experimenter. Subjects also received training on how the media presents the guideline information (structure, guideline material, and design aids). For DRUID, the session also included material on how to use the software to manage files and input data.

After the demonstrations, the subjects were given five minutes to use the medium as they wished and to ask questions. Next, the subjects performed a set of test tasks on that presentation medium (Appendix D). The tasks represent a cross section of the skills necessary to use the appropriate presentation medium effectively. All subjects received feedback on incorrect responses at the completion of the tasks. A criterion of 80% correct was set for passing the test. Those subjects that did not
reach criterion received additional training and were then given a second set of test tasks.

All the subjects reached criterion on the initial test for the book. All except one subject met the initial criterion with the DRUID software. The one subject that did not reach criterion was given additional training and was then re-tested. That subject passed the second test. Thus, all the subjects successfully completed the training portion and continued with the experiment. Each subject received a five minute break after completing the training.

Following the break, each subject read an instruction sheet that described the tasks for that trial. A copy of the instruction sheet is provided in Appendix E. After reading the instructions, the experimenter answered the subjects' questions.

The subjects then read one of the two task scenario descriptions (Appendix B). Questions asked to clarify the scenario description were answered at any time during the testing. There were no time constraints on completing tasks. However, the instructions emphasized that the guidelines should be selected as efficiently and accurately as possible.

To make the experiment more manageable, the subjects were instructed to select guidelines only from the "Data Display" section of the Guidelines. That section was chosen because it is the largest single section of the guidelines (298 of 944 total guidelines). Also, the "Data Display" guidelines have the widest topical coverage and are probably the easiest to understand.

When subjects used DRUID, they were instructed to enter their selection into the software. Subjects typed in a "Y" (for Yes) to indicate selection of relevant guidelines, and typed in a "?" for guidelines they are unsure of, as was previously shown in Figure 5.
When using the hardcopy method, subjects used an unmarked copy of the Guidelines and a preformatted answer sheet. The first page of an unmarked answer sheet is presented in Appendix F. The subjects were instructed to mark their selections in the space provided to the right of the reference number on the answer sheet. It should be noted that the answer sheet was created specifically for this experiment. No such sheet is available for the typical book user. In that respect, an unfair advantage is given to the paper medium. However, it was decided that subjects may not select all the guidelines they desire if they had to write down a large quantity of guideline numbers. Also, the subjects were allowed to mark the copy of the Guidelines in any way they wished.

Several blank pieces of paper were also provided to the subjects. The participants were informed that they may use the paper anyway they wished to help them with the guideline selection task.

After studying the task scenario description, the subjects began selecting guidelines. The subjects informed the experimenter when they had completed the selection task. After completing guideline selection for each medium, the subjects were asked to complete a post-test questionnaire (Appendix G). The questions were directed toward assessing the usability and usefulness of the guidelines themselves and the usability of the presentation medium. The usability questions were modeled after those presented by Ravdin and Johnson (1989).

Following the completion of both tasks, a summary questionnaire was presented to the subjects (Appendix H). On that questionnaire, subjects answered questions comparing the two presentation media.

Throughout the experiment, subjective data were recorded via verbal protocol. The instructions encouraged the subjects to verbalize their decision making processes while selecting guidelines.
RESULTS

Phase I

In phase I of the experiment, experts selected those guidelines they thought were applicable to the two scenarios. A list of all the guidelines that were independently picked by all three experts (the core guidelines) is presented in Appendix I. There were a total of 33 core guidelines selected for the ACME scenario and 28 core guidelines for the Widgets scenario. After the experts' results were tallied, a meeting of the experts was held to discuss the results. At that meeting, the experts agreed that the original selections would be used as the basis for the core set because each of the experts felt confident and justified about their selections.

Phase II

Critical Measures. As defined in the "METHOD" section, the three critical measures for this experiment were the total number of core guidelines selected, time to complete the task, and the number of guidelines that were picked from the abbreviated title lists instead of the complete guidelines.

Those measures were evaluated using an analysis of variance (ANOVA), with experience level, presentation medium, and scenario as the three main factors. Because the experiment was not completely counterbalanced, a three-way between subjects ANOVA was constructed so that two-way and three-way interactions could be measured. Without this type of ANOVA, the Media x Scenario effect and Media x Skill x Experience Level effects could not be tested. However, the consequence of such a procedure is that sensitivity is lost by using only the residual as the error term. The three-way ANOVA produced no significant results for any main effects or any interactions. The summary table for that ANOVA is presented in Appendix J.
Because the Media x Skill x Experience Level interaction was not significant, the data were pooled once across scenarios (Experience x Medium) and once across media (Experience x Scenario) for further analysis using two two-way ANOVAs. All the ANOVA summary tables for this section are presented in Appendix J. There were no significant two-way interactions recorded for any measure. Thus, the F and p-values for interactions are not reported in this results section. However, they are reported in the ANOVA summary tables in Appendix J for reference.

Because of the unequal number of core guidelines (33 for ACME and 28 for Widgets), the subjects' percentage of core guidelines selected was used as the primary dependent measure for how well the subjects performed. This measure is known as recall in information retrieval literature and is defined as the number of items retrieved and relevant divided by the total number of relevant items in the collection (Salton, 1975). In this case, the items are guidelines and the collection is the 298 guidelines in the Data Display section of the Guidelines. There was a significant difference in performance for presentation medium, $F(1,6) = 25.126, p = 0.002$. All eight subjects selected a higher percentage of core guidelines when using the book (mean = 91.2%) than when using DRUID (mean = 83.4%). For experience level, experienced (mean = 91.0%) versus novice (mean = 83.6%), was not significant, $F(1,6) = 1.344, p = 0.290$. There was also no significant effect for scenario, $F(1,6) = 0.10, p = 0.924$. The mean for ACME was 87.1% and the mean for Widgets was 87.75%.

There were no significant main effects or interactions for the time to complete the task. On average, the experienced subjects took 61.71 minutes to complete a task scenario and the novice subjects took 58.31 minutes; $F(1,6) = 0.276, p = 0.618$. The mean time to complete a task with DRUID was 65.90 minutes and the mean time using the book was 54.12 minutes; $F(1,6) = 2.849, p = 0.1424$. The mean times to complete the respective scenarios were 61.02 minutes for ACME and 59.01 minutes.
for Widgets; \( F(1,6) = 0.057, p = 0.8196 \). Pooled across all conditions, the mean time to complete a scenario was 60.02 minutes. This time ranged from a low of 30.43 to a high of 84.87 minutes.

Using the percent of guidelines that were picked from the list as the final critical dependent measure, the medium main effect had a significance level slightly higher than the 0.05 criterion, \( F(1,6) = 5.483, p = 0.058 \). The mean percentage for DRUID was 19.3\%, while the mean for the book was 3.6\%. Across all subjects, a total of 152 guidelines was picked from the title lists within DRUID, and a total of 28 guidelines was picked from the title lists of the book. Again, experience level (experienced = 12.1\% and novice = 10.8\%), \( F(1,6) = 0.014, p = 0.911 \) and scenario (ACME = 9.7\% and Widgets = 13.2\%), \( F(1,6) = 0.141, p = 0.720 \) were not significant.

The two task scenarios were designed to be equivalent as possible. This was successfully accomplished because none of the critical dependent measures produced any significant effects for scenario as either a main effect or any interactions. Thus, the remaining guideline usage measures were calculated for the main effects of experience and medium only.

**Efficiency.** Another important measure was task efficiency. Efficiency was calculated two ways. The first way was in terms of the time it took to select the core guidelines (core guidelines per minute). There was a significant difference in efficiency for the media, \( F(1,6) = 6.007, p = 0.049 \). The mean rating for DRUID was 0.4 guidelines/minute, while the mean rating for the book was 0.545 guidelines/minute. There was practically no difference due to experience level, \( F(1,6) = 0.003, p = 0.995 \). Experienced guideline users averaged 0.470 guidelines/minute while novice users averaged slighted more at 0.475 guidelines/minute.

Efficiency was also measured by calculating an information retrieval measure called *precision*. Precision is defined as the number of relevant items retrieved from
a collection divided by the total number of items retrieved from a collection (Salton, 1975). The mean precision rate for DRUID was 0.342 and the mean precision rate for the book was 0.390; $F(1,6) = 3.147, p = 0.126$. For experience level, the mean precision rate was 0.329 for the experienced group and was 0.403 for the novice group. Thus, the precision measure resulted in no significant differences for either medium or experience level. Individual precision rates ranged from 0.194 to 0.500.

As a way of measuring guideline comprehension, the number of guidelines the subjects marked as unsure (either because the subject did not understand the meaning of the guideline or could not tell if it was relevant) was collected and analyzed. There was no significant difference for medium, $F(1,6) = 0.1724, p = 0.237$; or experience level, $F(1,6) = 0.661, p = 0.447$. An average of 4.0 guidelines were marked as unsure with DRUID, while 1.875 guidelines were marked as unsure with the book. The mean number of guidelines marked as unsure was 2.13 for the experienced subjects and 3.75 for the novice subjects.

*Other Guideline Usage Measures.* Additional measures were collected on how often the subjects used the media-supplied design aids, such as cross-references and the glossary. Most of the aids were used quite infrequently and produced no significant results. The results of those measures are listed in Tables 8 and 9.

**TABLE 8. Guideline Usage Means**

| Dependent Measure | Experience Level | Medium | | | |
|-------------------|-----------------|--------|--------|--------|
|                   | Experienced     | Novice | DRUID  | Book   |
| Cross-references  | 1.000           | 1.375  | 1.125  | 1.250  |
| Glossary          | 0.000           | 1.125  | 0.250  | 0.875  |
| Index             | 0.750           | 0.000  | 0.375  | 0.375  |
| Samples           | 0.375           | 0.750  | 0.500  | 0.625  |
TABLE 9. Guideline Usage ANOVA Results

<table>
<thead>
<tr>
<th>Dependent Measure</th>
<th>Factor</th>
<th>F-value</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cross-references</td>
<td>Experience</td>
<td>0.184</td>
<td>0.683</td>
</tr>
<tr>
<td></td>
<td>Media</td>
<td>0.038</td>
<td>0.852</td>
</tr>
<tr>
<td>Glossary</td>
<td>Experience</td>
<td>4.765</td>
<td>0.718</td>
</tr>
<tr>
<td></td>
<td>Media</td>
<td>1.271</td>
<td>0.303</td>
</tr>
<tr>
<td>Index</td>
<td>Experience</td>
<td>1.000</td>
<td>0.356</td>
</tr>
<tr>
<td></td>
<td>Media</td>
<td>N/A</td>
<td>N/A</td>
</tr>
<tr>
<td>Samples</td>
<td>Experience</td>
<td>1.174</td>
<td>0.320</td>
</tr>
<tr>
<td></td>
<td>Media</td>
<td>0.045</td>
<td>0.839</td>
</tr>
</tbody>
</table>

**Correlation.** A way of assessing the variability in guideline selection is to measure the degree of correlation between the number of experts that selected each guideline and the number of subjects that selected the guidelines. In other words, for each guideline in the data display section, anywhere from zero to three experts may have selected it. Total agreement occurred when no experts picked a guideline and where all three experts selected a guideline. Where one or two experts selected a guideline, there was disagreement. Likewise, for the subjects, anywhere from zero to eight subjects may have selected the same guideline for a given scenario. Again, total agreement occurred when all eight subjects selected a guideline and when zero subjects selected a guideline. Disagreement occurs when anywhere from one to seven subjects select a guideline. The correlation between experts and subjects on the degree of agreement (certainty) that a guideline should be selected can be calculated by tallying (for each guideline) how many experts selected it and how many subjects also selected it (i.e., 3 experts selected a particular guideline, while only 6 subjects selected that same guideline), independent of presentation medium.

A table of all possible outcomes can then built (e.g., a contingency table of all combinations of how many experts and subjects could pick a guideline). Such a
contingency table is shown as Table 10. The number of experts that picked any particular guideline is presented on the X-axis and the number of subjects picking the same guideline is shown on the Y-axis. The values for each cell in that table represent the number of guidelines for which any particular combination of experts and subject selection (i.e., the frequency). The contingency tables for the ACME and Widgets scenarios are shown in Tables 11 and 12 respectively.

TABLE 10. Contingency Table of All Possible Combinations of Subjects and Experts Picking any Guideline

<table>
<thead>
<tr>
<th>Number of subjects selecting the same guideline</th>
<th>Number of experts selecting any guideline</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Σ</td>
<td></td>
</tr>
</tbody>
</table>
### TABLE 11. Contingency Table for ACME

<table>
<thead>
<tr>
<th>Number of subjects selecting the same guideline</th>
<th>Number of experts selecting any guideline</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>67 1 0 0 68</td>
</tr>
<tr>
<td>1</td>
<td>5 2 3 0 10</td>
</tr>
<tr>
<td>2</td>
<td>12 1 2 2 17</td>
</tr>
<tr>
<td>3</td>
<td>0 1 6 0 7</td>
</tr>
<tr>
<td>4</td>
<td>1 4 18 1 24</td>
</tr>
<tr>
<td>5</td>
<td>1 13 1 16</td>
</tr>
<tr>
<td>6</td>
<td>0 7 4 11</td>
</tr>
<tr>
<td>7</td>
<td>0 12 11 23</td>
</tr>
<tr>
<td>8</td>
<td>0 4 15 19</td>
</tr>
<tr>
<td>Σ</td>
<td>86 10 65 34 195</td>
</tr>
</tbody>
</table>

### TABLE 12. Contingency Table for Widgets

<table>
<thead>
<tr>
<th>Number of subjects selecting the same guideline</th>
<th>Number of experts selecting any guideline</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>45 4 3 0 52</td>
</tr>
<tr>
<td>1</td>
<td>5 5 14 0 24</td>
</tr>
<tr>
<td>2</td>
<td>11 1 2 0 14</td>
</tr>
<tr>
<td>3</td>
<td>5 3 2 0 10</td>
</tr>
<tr>
<td>4</td>
<td>11 5 6 2 24</td>
</tr>
<tr>
<td>5</td>
<td>4 10 4 4 22</td>
</tr>
<tr>
<td>6</td>
<td>1 1 13 3 18</td>
</tr>
<tr>
<td>7</td>
<td>0 1 8 2 11</td>
</tr>
<tr>
<td>8</td>
<td>0 0 3 17 20</td>
</tr>
<tr>
<td>Σ</td>
<td>82 30 55 28 195</td>
</tr>
</tbody>
</table>
Because the scenario descriptions explicitly stated that the USI designs could not include graphics, no subject looked at the graphics-related guidelines. Thus, the 103 graphics guidelines were not included in correlation analysis, reducing the total set of guidelines from 298 to 195. If those graphic guidelines were included in the analysis, the scores for the "0,0" cells would increase from 45 to 148 for ACME, and from 67 to 170 for the Widgets scenario. Since no real decision making happened for those individual guidelines or their functions, including the extra scores would artificially place too much weight into the "0,0" cell and obscure the measures.

Using the Pearson Product-Moment statistic, the correlation between the guideline experts and subjects for the ACME scenario was $R = 0.88 \ (R^2 = 0.77)$, and for the Widgets scenario $R = 0.68 \ (R^2 = 0.46)$. Correlation cannot be measured for presentation media because of the incomplete counterbalancing discussed in the methods section.

The correlation can also be shown graphically by plotting the data in three dimensions. The same correlation data are presented in Figures 7 and 8, but this time plotting the frequency the cells as the third dimension on the Z-axis.
Figure 7. 3D plot of relative guideline selection for ACME scenario.
Subjective Measures.

Pre-test Questionnaire.

The pre-test questionnaire was used only to screen subjects and collect demographic data. Therefore, only an abbreviated summary of that data are presented here to provide some insight into the type of subjects that participated in the experiment.

All of the subjects had at least two years of experience in systems and UI design; five of the eight subjects had over five years experience in both of those disciplines. The types of interfaces with which the subjects had experienced were quite diverse, ranging from ASCII-only to speech-based to modern graphical displays.
Four of the subjects had formal USI design courses in college and five of the subjects had taken professional short courses (there was an overlap). All of the subjects had previously used a Macintosh computer. Only one subject had less than six months experience using the Macintosh. However, that subject had over a year's experience using an IBM PC with Windows™.

Seven of the eight subjects said they had used the Smith and Mosier Guidelines before the experiment. The experienced users had used the Guidelines numerous times for design, evaluation, and reference. The novice users had used the Guidelines in school and only in a cursory manner at work (for occasional reference).

The subjects were biased towards wanting to work with formal specifications in the design process. When asked if they preferred to work from a "formal USI specification when designing USIs," seven of the eight subjects answered yes. The one subject that answered no responded that "...[specifications] are too specific."

**Post-test Questionnaire.**

A post-test questionnaire was administered after each trial to collect the subjective ratings on the presentation medium. Because the responses utilized Likert scales with anchors, the data were treated as an interval scale. The scores were analyzed using an analysis of variance (ANOVA) for medium x experience level. Tables 13 and 14 summarize the post-test ratings. Table 13 provides the average scores, F-values, and p-values for each question based on presentation medium. Table 14 provides the same data, but based on the experience level. Again, there were no significant interactions to report.
### TABLE 13. Post-test Scores for Media

<table>
<thead>
<tr>
<th>QUESTION</th>
<th>Averages</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. How confident are you that you found all the guidelines that applied to this scenario? 1 = Very Unsure  -&gt; 7 = Very Confident</td>
<td>5.375</td>
<td>5.625</td>
<td>1.000</td>
<td>0.356</td>
</tr>
<tr>
<td>2. Did you lose track of where you were in the guidelines during the task? 1 = Never  -&gt; 7 = Always</td>
<td>3.625</td>
<td>2.500</td>
<td>4.765</td>
<td>0.072</td>
</tr>
<tr>
<td>3. Did you have to remember a lot of information while performing the task? 1 = Very Little to remember  -&gt; 7 = Too much</td>
<td>3.000</td>
<td>3.500</td>
<td>1.600</td>
<td>0.253</td>
</tr>
<tr>
<td>4. Recording the selected guidelines was: 1 = Very Difficult  -&gt; 7 = Very Easy</td>
<td>6.125</td>
<td>6.250</td>
<td>0.300</td>
<td>0.868</td>
</tr>
<tr>
<td>5. How would you rate the actual guideline text (not the presentation medium) in terms of visual clarity (organization, ambiguity, and readability)? 1 = Very Unsatisfactory  -&gt; 7 = Very Satisfactory</td>
<td>4.875</td>
<td>4.500</td>
<td>0.262</td>
<td>0.627</td>
</tr>
<tr>
<td>6. How would you rate the presentation medium (document or software) in terms of visual clarity (organization and readability)? 1 = Very Unsatisfactory  -&gt; 7 = Very Satisfactory</td>
<td>5.625</td>
<td>5.500</td>
<td>0.111</td>
<td>0.750</td>
</tr>
<tr>
<td>7. How would you rate the document or software in terms of consistency? 1 = Very Inconsistent  -&gt; 7 = Very Consistent</td>
<td>6.375</td>
<td>5.375</td>
<td>3.429</td>
<td>0.114</td>
</tr>
<tr>
<td>8. How well did the document or software meet your needs? 1 = Very Poorly  -&gt; 7 = Very Well</td>
<td>6.000</td>
<td>5.125</td>
<td>5.444</td>
<td>0.058</td>
</tr>
<tr>
<td>9. How would you rate the document or software in terms of providing aids in performing the task? 1 = Very Unsatisfactory  -&gt; 7 = Very Satisfactory</td>
<td>5.875</td>
<td>4.875</td>
<td>2.400</td>
<td>0.172</td>
</tr>
</tbody>
</table>
TABLE 14. Post-test Scores for Experience Level

<table>
<thead>
<tr>
<th>QUESTION</th>
<th>Averages</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1. How confident are you that you found all the guidelines that applied to this scenario? 1 = Very Unsure -&gt; 7 = Very Confident</td>
<td>5.000</td>
<td>6.000</td>
<td>1.000</td>
<td>0.356</td>
</tr>
<tr>
<td>2. Did you lose track of where you were in the guidelines during the task? 1 = Never -&gt; 7 = Always</td>
<td>2.875</td>
<td>3.250</td>
<td>0.097</td>
<td>0.766</td>
</tr>
<tr>
<td>3. Did you have to remember a lot of information while performing the task? 1 = Very Little to remember -&gt; 7 = Too much</td>
<td>3.625</td>
<td>2.875</td>
<td>0.425</td>
<td>0.539</td>
</tr>
<tr>
<td>4. Recording the selected guidelines was: 1 = Very Difficult -&gt; 7 = Very Easy</td>
<td>5.875</td>
<td>6.500</td>
<td>2.143</td>
<td>0.194</td>
</tr>
<tr>
<td>5. How would you rate the actual guideline text (not the presentation medium) in terms of visual clarity (organization, ambiguity, and readability)? 1 = Very Unsatisfactory -&gt; 7 = Very Satisfactory</td>
<td>4.750</td>
<td>4.625</td>
<td>0.017</td>
<td>0.900</td>
</tr>
<tr>
<td>6. How would you rate the presentation medium (document or software) in terms of visual clarity (organization and readability)? 1 = Very Unsatisfactory -&gt; 7 = Very Satisfactory</td>
<td>5.250</td>
<td>5.875</td>
<td>0.556</td>
<td>0.484</td>
</tr>
<tr>
<td>7. How would you rate the document or software in terms of consistency? 1 = Very Inconsistent -&gt; 7 = Very Consistent</td>
<td>6.000</td>
<td>5.750</td>
<td>0.231</td>
<td>0.648</td>
</tr>
<tr>
<td>8. How well did the document or software meet your needs? 1 = Very Poorly -&gt; 7 = Very Well</td>
<td>5.375</td>
<td>5.750</td>
<td>0.380</td>
<td>0.560</td>
</tr>
<tr>
<td>9. How would you rate the document or software in terms of providing aids in performing the task? 1 = Very Unsatisfactory -&gt; 7 = Very Satisfactory</td>
<td>5.000</td>
<td>5.750</td>
<td>0.730</td>
<td>0.426</td>
</tr>
</tbody>
</table>
The post-test questionnaire also consisted of eight additional open-ended questions. The results are summarized below. The responses are divided by media type. If a response was given more than once, the frequency of that response is indicated in parentheses after that response. The maximum frequency score for any individual response is eight, the total number of subjects. Some of the responses listed below have been reworded for clarification. Where illustrative, the listed responses use the subject's exact wording.

Question #10: Did the document or software provide you with all the necessary information you needed to perform the task? If not what information was missing?

**Hardcopy:**
Six of the eight subjects answered "yes" to this question. Subjects noted that expanded information (more annotation) would have made the guidelines more usable.

**DRUID:**
Seven of the eight subjects answered "yes" to this question. The one subject that said "no" suggested that DRUID could have provided aids in eliminating non-relevant guidelines.

Question #11: List three things you liked about the document or software.

**Hardcopy:**
- Easy paging. (3)
- The examples. (3)
- The cross-references. (2)
- Guideline organization. (2)
- Useful topical index. (2)
- Definitions combined with guidelines. (2)
- The "carrot" markers that show related guidelines.
- Easy to scan pages.
- Wide margins for marking guidelines.
- Easy to identify sections.
- Easy to orient yourself.
- Page layout.

**DRUID:**
- Easy to navigate through the guidelines. (5)
- On-line checklist. (4)
Speed of navigation (3)
Checklist printouts. (2)
Review guidelines at various levels of detail. (2)
(Condensed title lists versus individual guidelines.)
Automated cross-referencing. (2)
Easy to scan information.
Forgiving of errors.
Availability of definitions.
Consistent with current work formats.
Good location cues.
"Good use of hypermedia."

Question #12: List three things you did not like about the document or software.

**Hardcopy:**
Samples do not show which guidelines are violated.
Need way to "weed-out" non-applicable guidelines.
Guideline titles could provide more information.
Hard to produce a "tailored" report.
Lost track of Guidelines structure.
Similar guidelines put in different functions.
"Coding" guidelines are hard to apply.
Lots of page turning.
Format of only 2-4 guidelines per page.
Physically having to page through "such a large" document is cumbersome.
Need more definitions.
Need more examples.
Some guidelines conflict.

**DRUID:**
Wanted to see more than one guideline at a time. (3)
Poor audio cues. (2)
Slow response time. (2)
Software did not automatically eliminate the guideline within a function
when the function was eliminated.
When marking individual guidelines, user wanted to use return key to
move to the next guideline (instead of clicking on the "next" button).
Wanted to type in notes (annotate guidelines).
Problem with selecting from title list (ending up selecting the wrong line
from the list).
No automatic tailoring of guidelines based system constraints.
Got lost -- wanted a map.
Guideline wording confusing -- wording needs to be simplified.
Need a hierarchic browser for novice.
Question #13: Describe any parts of the document or software which were confusing or difficult to understand.

Hardcopy:
Some guidelines are hard to understand. (2)
Function definitions did not provide enough information to judge if guidelines were appropriate.
Confusion sections -- "Coding" and "Display Control."
"Panning" versus "scrolling" guidelines were confusing.

DRUID:
Problems with the abbreviations. (2)
Poor audio cues.
Variety of input mechanisms.

Question #14: Did the document or software help you in selecting guidelines. If so, how?

Hardcopy: (seven out of eight subjects said "yes")
Structure of the document. (3)
Provides single source of information.
Can easily scan necessary level of detail for guidelines.
Some examples clarified the intent of guidelines.
Reduced the possibility of forgetting guidelines.

DRUID: (seven out of eight subjects said "yes")
Easy to see overview of selections. (2)
Made it "easy and painless" to explore links (cross-references). (2)
Easy to make changes.
Provided organized approach to selecting and tracking guidelines.

Question #15: Did the document or software hinder you in any way in selecting guidelines?

Hardcopy: (two out of eight subjects said "yes")
Some guidelines too vague to interpret.
There was a lot of page turning.
Lends itself to "sequential" rather than "jumping" around.

DRUID: (three out of eight subjects said yes)
Slow response time. (2)
Too much information.
Can only view one guideline at a time.
Had problems selecting items from the lists.
Question #16: Did the presentation media (document or software) influence your design process? How?

**Hardcopy**: (two out of eight subjects said "yes")
- Tended to go through guidelines "sequentially" rather than "jumping" around.
- There was a lot of page turning.
- "Seems like it imposes such a tedious process on you that you want to simplify your design to where guidelines won't apply."

**DRUID**: (two out of eight subjects said yes)
- Tried to pick or disqualify guidelines based on title alone to save time.
  "Because if I was unsure of what the terms meant, or what the area covered, I could ask the system and get a rapid answer."

Question #17: Did you find the function definitions useful in selecting guidelines?

**Hardcopy**: (five out of eight subjects said "yes")
- Put guidelines into context -- lessened ambiguity of guidelines for that function. (4)
- Tended to read them because they were there.

**DRUID**: (six out of eight subjects said "yes")
- Clarified the meaning of guidelines. (4)
- Refreshed my memory.
  "Informed what the guideline author was thinking."

Summary Questionnaire.

A summary questionnaire was distributed to each subject after completing the second experimental task. That questionnaire provided the subjects with the opportunity to compare the two presentation media subjectively. Also, the questionnaire asked the subjects about their design process to see how they selected guidelines and how the media affected that process.
**Question #1** Did you "picture" the design in your mind before, during, or after the task? If you answered "before" or "during" in the last question, did you refine your design as you selected more guidelines?

Six out of eight subjects pictured the design before selecting any guidelines. The remaining two started to picture the design while selecting guidelines. However, only one subject actually started drawing before selecting guidelines. One other subject sketched the design while selecting the guidelines.

**Question #2:** Which task, if any, was more difficult to perform?

The subjects were evenly divided on this question: two subjects said the ACME scenario, two subjects said the Widgets scenario, and the remaining four subjects said they were equal in difficulty.

**Question #3:** Did the size of the systems for which you selected guidelines affect your guideline selection? (for example: total number selected, type selected)

Four of the subjects reported being influenced. Each of those subjects said that they selected a smaller number of guidelines because of the limited functionality of the scenarios.

**Question #4:** Which presentation medium did you prefer?

The majority of the subjects (five) preferred the software, DRUID; while the other three subjects said it depends on the task. The following reasons were given for the DRUID preference:

- **Easy of use.** (3)
- **Automatically creates checklists.** (2)
- **Speed.** (2)
- **Compact size.**
- **Easy to follow links.**

Those people that said it depends, stated the following reasons:

- For small tasks the book is faster, but would prefer DRUID for large tasks.
- Book seems faster, but selecting and tailoring is preferred with DRUID.

Another common comment, for the subjects that preferred DRUID and for those who said it depends, was that as they got more experience with DRUID, they thought they would grow to like it even more.
Question #5: Which presentation media was easier to use?

Four subjects said DRUID was easier to use for the following reasons: 
  Had better navigation tools. 
  Had automatic checklist generation. 
  Easier to use than physically paging through the book. 
  "It's more natural -- paper is too cumbersome." 
  There was less searching because of the links. 

Two subjects preferred the book because: 
  More learning was necessary for DRUID. 
  With checklist (answer sheet) provided, it was faster to scan and select guidelines. 

The remaining two subjects said it depended on: 
  The size of the task (prefer book for smaller tasks and DRUID for larger ones). 
  Prefer book now only because of lack of experience. 

Question #6: Did the presentation media influence your selection strategy? 

Only three of subjects responded that there was an influence. The following influences were stated: 
  With the software, there was less inclination to go through the guidelines sequentially. (2) 
  "Using the paper was a such a drag, I found myself compromising the design just to get through it." 

Question #7: Did the presentation media influence your overall design process? 

All eight subjects answered "no" to this question. 

Drawings. Upon completion of each post-test questionnaire, the subjects were asked to provide a rough sketch of their design. The subjects were asked to provide only enough detail to communicate the USI's "look and feel." Those sketches are presented in Appendix K. The sketches range in detail from brief to quite elaborate. For each scenario, the common attributes of those designs are listed below with their frequencies shown in parentheses (the maximum frequency is eight). Where
relevant, the numbers for the core guidelines that relate to those design features are listed. The titles of the core guidelines are listed in Appendix I.

ACME:

The interface was divided into areas by function. (8)

2.0-5

The comment text field was placed at the bottom of the display. (8)

The fields and their labels had a consistent format and did not use jargon. (8)

2.0-6, 2.0-12, 2.2-1 through 2.2-9, 2.5-1

The primary fields were the employee name and number, which were the required input fields. (7)

2.7.1-1

The title was placed at the top of the display. (6)

2.5-10

The controls were placed at the bottom of the display. (6)

2.5-11

All data were displayed on one screen (5)

Fields that had pre-defined formats, such as for the social security number, had any necessary markers embedded in the fields. (3)

2.2-14

The comment field was represented in one of two manners:

As a scrolling field (4) or a multi-page field. (4)

2.7.2-2

Of the subjects that chose the multi-page format, most (3/4) chose to indicate the current page number "X of Y."

There were only two proposed input techniques:

Keyboard input. (5)

Keyboard and function key combination. (3)
Widgets:
The sketches clearly divided the interface into areas by function. (8) 2.0-5

The fields and their labels had a consistent format and did not use jargon. (8) 2.0-4, 2.0-5

The title was placed at the top of the display. (8) 2.5-10

All data were displayed on one screen. (7) 2.5-7, 2.7.2-1

Critical data were highlighted as required. (5) 2.6-1, 2.6-12, 2.6-13

Numeric data were properly aligned. (4)

The locations of the display controls varied as followed:
  bottom (5)
  top (2)
  separate screen (1) 2.5-11

The input method also varied:
  Keyboard only (6)
  List selection (1)
  Could not tell (1)

The months for the sale period were displayed either vertically and horizontally in well defined rows and columns.
  horizontally (3)
  vertically (3)

  Of these, all had the reference data in the first column. 2.5-3, 2.5-12, 2.5-13
DISCUSSION

Presentation Medium

For most dependent measures, there were no significant differences in user performance based on the presentation medium (book or hypertext software). However, one critical measure, percentage of core guidelines selected (recall), showed a significant difference $F(1,6) = 25.126, p = 0.0024$. When using the book, the subject's averaged 91.2% of the core guidelines, but averaged only 83.4% of the core guidelines with DRUID. However, most other measures showed little difference in performance.

The difference in the number of core guidelines selected cannot be attributed to a single cause. It is believed that a variety of factors led to the book's higher score. Evaluating both the paper document and software provided an efficient means for identifying the factors that influenced the users' relative performance. Those factors are presented below.

Familiarity with Medium. A problem that must be overcome with any hypertext system is experience. Anytime hypertext is compared to a document it must be understood that most subjects have been trained on how to use a book effectively since elementary school. Even with several hours or days of training with a hypertext system, a book will have the advantage of greater familiarity. In the post-test and summary questionnaires, the subjects noted that they felt more comfortable using the book.

Image Quality and Text Formatting. Independent of DRUID, it has been shown that user performance for software-displayed text is not as good as when displayed on paper. Gould and his colleagues (Gould et al., 1987) have shown that people read 15-25% more slowly from CRTs than from paper except under ideal conditions. A
detailed examination of those reasons goes beyond the intent of this discussion; however, those authors note that the difference was based on the relative image qualities and text formatting of the media. In this experiment, no adjustments were made for image quality. DRUID was not designed to maximize that aspect of the USI design. Some of the image quality and text formatting aspects were compromised for the sake of implementing the software's hypertext capabilities, such as using a monospace font (Monaco) and displaying text in a small point size (because of HyperCard's small window limitations). On the other hand, the paper version of the Guidelines has been well engineered. Over the many iterations of that document, user comments have been incorporated to make the book more usable. In its current form, the book makes extensive use of formatting to aid the user. Also, the book is printed with high quality typesetting on superior stock matt paper.

Display Size. DRUID's display area may have also affected the subjects' performance. Hansen and Hass (1988) showed that the size of the screen was also a factor in CRT usability. Unfortunately, DRUID was limited to HyperCard's maximum screen size of nine inches diagonally. This placed DRUID at a distinct disadvantage. With the limited screen real estate, DRUID only displays one guideline at a time; and frequently, the user must scroll through the material because lengthy guidelines cannot all fit on the screen at once.

That effect of having such a small display size showed up in the responses to the post-test questionnaire, where three of the eight subjects commented that they wanted to see more than one guideline at a time. Also, from direct observation, it was interesting to note that the subjects rarely scrolled through those guidelines that did not fit on the screen. The subjects did not look beyond what was initially presented to them. Also, during the experiment, subjects commented that they sometimes lost context among guidelines because they could only see one guideline
at a time. An interesting side note is that two subjects commented that whenever they got lengthy electronic mail or word processing documents, they sent the file to the printer rather than read it from the screen.

Response Time. Another factor in the relative performance of the media was DRUID's response time. It took DRUID up to three seconds to "turn" between individual guidelines. During the experiment, most of the subjects commented on how slowly DRUID updated. Several of the subjects also commented on DRUID's inadequate response time on the post-test questionnaire.

Software Errors. A "bug" in the software also influenced several of the subject's performance. When guidelines were selected from their title lists, as shown in Figure 4, the software would sometimes record the wrong selection. That was due to the fact that some subjects would occasionally click on the first entry in the list, and then quickly move the pointer downward across the screen. Unfortunately, the software did not always update fast enough to catch where the subject actually "clicked." For example, instead of taking the user to the first guideline within the function (the actual selection), the software saw that the pointer was on the sixth guideline (when the software sampled), and thus brought the subject to the sixth guideline display instead of the first guideline for that function. Several times the subjects did not notice that they were brought to the wrong guideline, causing them to miss reviewing the preceding guidelines within those functions.

Selection Strategy. Interestingly, the times it took to complete the tasks were roughly equivalent for both media (01:05:55 for DRUID versus 00:54:07 for the book). Also, the subjects precision rates were not significantly different for the two media.

How then, did the subjects complete the tasks in the same amount of time? They altered their selection strategy. The subjects tried to reduce the amount they read to a minimum when using the software. This was demonstrated in two ways.
First, subjects selected more guidelines from the condensed list of guideline titles when using DRUID ($F(1, 6) = 5.483, p = 0.0577$). With DRUID, subjects picked on average almost 20% of the guidelines from the title lists. With DRUID, five subjects used the lists, while only one subject used the list in the book.

The change in strategy of the subjects picking from the title lists in DRUID also contributed to the subjects' poorer performance when using DRUID. Because guidelines were selected from the title lists, the subjects did not have as much information about those guidelines when determining their relevance. However, it is impossible to determine if a subject would have selected a missed core guideline if he/she read it from its own full display. From direct observation, it appeared as though the subjects chose those items from the list that they felt comfortable with (via individual paging), but when they were unsure of guidelines they would switch back to the full guidelines display. When subjects again reached a guideline they felt comfortable with, that they would often "jump" back up to the title list to expedite the selection process.

That accessing of guidelines at different levels of detail caused the subjects to perform more complex navigation strategies. Sometimes the subjects would become confused as to where they were. This problem occurred most often when subjects went from the first guideline of section 2.7 ("2.7-1") to the first guideline of the next section, 2.7.1 ("2.7.1-1"). In that particular case function 2.7 (Display Control) has several sub-functions, so section 2.8 does not directly follow 2.7. The sometimes confusing numbering scheme of the guidelines did not seem to cause as many problems when using the book, since the subjects tended to navigate in a more linear (sequential) manner.

Second, when viewing the individual guideline displays, the subjects rarely read the annotation associated with the guidelines. From the videotapes of the
experiment, it was noticed that the subjects quite often moved the pointer from word to word on the screen to assist them in reading. Thus, it was easy to determine that the subjects often were not reading the annotation material that was provided to put the guideline into context.

Summarized below are the hypothesized reasons for the book's higher score for selecting relevant guidelines. Those factors need to be addressed in potential future updates to improve the software's effectiveness for selecting guidelines.

- Familiarity with the medium
- Image quality and text formatting
- Display area
- Response time
- Software errors
- Amount of text read
- Navigation issues

Subjectively however, most of the subjects preferred using DRUID and said that if they had to perform a similar task again they would use DRUID. Though statistically there were no differences (at the 0.05 level) in each of the ratings of the presentation medium, on six of the nine questions, DRUID received a higher mean score than the book. In general, it seems that it was the software's aids for selecting the guidelines that most contributed to these results. In particular, subjects noted that the software better met their needs ($F(1,6) = 5.444, p = 0.0584$) and that the software helped them keep tract of where they were when selecting guidelines ($F(1,6) = 0.4.765, p = 0.072$) despite the subjects getting lost between sections 2.7 and 2.7.1. It is also important to note that several subjects thought they would perform even better with DRUID once they gained more hands-on experience.
Level of Guideline Experience

The level of experience with the Guidelines did not seem to have any significant influence on any of the measures used in this experiment. Thus, the novice Guidelines users, in general, performed as well (efficiently and effectively) as subjects who had considerably more experience with the Guidelines material.

There are several possible explanations for this rather unexpected result. First, all subjects received the same level of training with the Guidelines and the presentation media. Also, all subjects reached criterion before performing the experiment, as dictated by the experimental design. Based on those factors, one could argue that both presentation media were sufficiently well designed so that they provided people inexperienced with the guideline content enough support to perform a guideline selection task. While at the same time, the media provided enough flexibility for an experienced Guidelines user to perform such a task without any hindrance.

A second factor that must be considered is the nature of the tasks used in the experiment. Though based on realistic concepts, the guideline selection tasks were simplified so that they could be conducted in a reasonable amount of time. For each scenario, the subjects had to pick guidelines from only the data display section of the Guidelines. That simplification eliminated the need to look at approximately two-thirds of the total set of guidelines. And because all the guidelines were from one section of the material, the need to cross-reference among sections (which can be difficult) was minimized. Also, the requirements of the software described in the scenarios were devised to maintain simplicity (no windows, graphics, etc.). Perhaps when selecting guidelines for a large complex system, the effects of the level of guideline experience would become more evident. However, collecting those
measures would be quite difficult to do given the length of time it would take to perform a final selection of guidelines (possibly weeks).

A third factor is that all the subjects had a rather liberal bias in the selection of guidelines. That is, when unsure if a guideline was needed, they tended to include it rather than leave it out. The precision measure detected that bias. The average precision rate for the experienced Guidelines users was 40.3% and was 32.9% for the novice users. There are several explanations for this result. The subjects probably knew that they were being evaluated on their ability to select the correct guidelines. However, it was probably not as obvious to the subjects that more subtle data, such as efficiency and selection strategies, were also being collected. Based those assumptions, it was in the subjects' best interest to select as many guidelines as possible.

Another reason for that liberal bias is that several of the subjects said their choice of guidelines should be judged as an initial selection, and that the final set of guidelines would vary. Several of the subjects commented during the experiment that they would later go back and make additional passes at the guidelines based on factors such as prototyping and collecting more task-related requirements.

One additional reason for the bias comes from the subjects' experience in systems design work. It was mentioned several times that based on experience in working with programmers, it was better to "over specify" the design, otherwise the programmers would implement features in the way they preferred them or would devote as little effort as possible to creating a good USI design.

Variability in Guideline Selection

Correlation Between Experts and Subjects. A high degree of correlation was found between the guidelines selected by the experts, and those selected by the subjects (0.68 for the ACME scenario versus 0.88 for the Widgets scenario). The high
correlation shows that the subjects tended to agree with the experts as to which guidelines applied to a scenario. Where there was certainty among experts there was also certainty among the subjects. That was shown by the fact that the cases where all the experts and all the subjects were in agreement (the "0,0" and "3,8" cells in the contingency tables) accounted for 37% of all the guidelines for the ACME scenario and 42% of all the guidelines for the Widgets scenario. Where there was disagreement among the experts, there was also disagreement among the subjects. That was shown in the dispersion of data points in the contingency tables. Another way of looking at this result is that the likelihood of the subject picking a given guideline can be predicted by the number of experts picking that guideline.

Because of the large difference in correlation between the two scenarios, further analysis was conducted to try to identify contributing factors. The guidelines that were least correlated were individually identified and analyzed (those in the upper right and the lower left corners of the scenario contingency tables), as shown in Tables 15 and 16.
### TABLE 15. Least Correlated Cells for the ACME Contingency Table

<table>
<thead>
<tr>
<th>Number of subjects selecting the same guidelines</th>
<th>Number of experts selecting any guideline</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Σ</td>
<td>1</td>
</tr>
</tbody>
</table>

### TABLE 16. Least Correlated Cells for the Widgets Contingency Table

<table>
<thead>
<tr>
<th>Number of subjects selecting the same guidelines</th>
<th>Number of experts selecting any guideline</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>1</td>
<td>14</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>6</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>Σ</td>
<td>5</td>
</tr>
</tbody>
</table>
As can be seen from the above tables, there are more guidelines which are less correlated for the Widgets scenario. Most of the guidelines that occupy those cells were from two specific functions: eighteen such guidelines were from function "2.1 Text," and nine were from function "2.2 Data Forms." For the text guidelines, it was generally the case that two experts picked the guidelines, while only two or fewer subjects selected those same guidelines. The definition provided for that function states that the text guidelines apply to data and "...[any associated] messages and other text intended for user guidance." Four of the eight subjects read the text definition during the experiment. From viewing the videotapes of the experiment, it appears as if most subjects assumed that many of text guidelines were not applicable for two reasons: 1) they were designing the interface for a data table (e.g., not displaying sentences) or 2) the text guidelines did not apply because the subjects were not designing the user guidance portion of the interface design. In fact, three of the eight subjects did not even look at any of the guidelines for this function. It appears as though there was confusion over what is considered data display and what is user guidance. Perhaps a clearer definition of that function would have helped avoided this problem.

The other function that had a large discrepancy for the Widgets scenario was "Data Forms." For the nine data forms guidelines in question, five subjects selected those guidelines, while at most only one expert selected those same guidelines. Eight of those nine guidelines were about field labels. The reason for this particular problem probably results in the organization of the guidelines. Both the Data Forms and the Data Tables functions have similar guidelines about field labels. The opposite is true for the "text" guidelines, where all the guidelines related to formatting textual material are only listed in only one function. Complicating the matter even further, there are a few non-redundant label guidelines from the data
forms section, that also apply to data tables, as was noted by one of the data table
guidelines (2.3-6).

Those subjects that went linearly through each section of the guidelines
selected the labeling guidelines in both of the functions. In fact, several of the
subjects mentioned that some of the guidelines were redundant and that they
thought they had seen similar guidelines in previous sections. Those subjects
disregarded the structure of the Guidelines by selecting the redundant data forms
guidelines. On the other hand, three subjects completely eliminated the data forms
guidelines (without looking at them) because they did not have a data form in their
designs. Actually, all of the subjects and experts had a data table in the interface
design, and no one used a data form in their interface design.

The reminder of the guidelines for which there was a large difference between
the experts and subjects were spread across each of the other Data Display functions.

Types of Guidelines Selected. It appears as though there was no single factor
that determined whether a guideline was selected for a scenario. Several factors that
did seem to contribute to the variability were (1) the interpretation of the scenario
requirements and (2) the individual preferences of how elements should be displayed
or represented. For example, subjects selected several different ways of highlighting
critical data (i.e., underlining, high intensity, and blinking). However, as can be seen
from the high recall rate and from the sketches in Appendix K, the subjects fairly
consistently selected the same elements for their designs.

Unlike the significant results reported by Reaux and Williges (1988) about
using guidelines for design evaluation, the level of specificity (concrete versus
abstract) of a guideline did not seem to influence the guideline selection. Though
there was no formal method of classifying the guidelines along this dimension, the
number of abstract and concrete guidelines in the core sets were almost evenly
divided, and subjects did well with both. It is interesting to note that only four
guidelines were chosen as core guidelines for both scenarios. Two of those guidelines
could be considered concrete and the other two abstract. Those guidelines were:

- 2.0-1 Necessary Data Displayed
- 2.0-5 Establishing Display Standards
- 2.5-10 Display Title at Top
- 2.5-11 Command Entry, Prompts, Messages at Bottom

Another question surrounding this result is the influence of having relatively
simplistic scenarios. It is possible that with more difficult scenarios there might be a
divergence in subject/expert performance. Also, the liberal response bias of the
subjects (as shown in the precision scores) may cast some doubt on how well the
subjects performed in comparison with the experts. In a real-life design situation,
there would be large penalties (in dollars) for requiring system developers to comply
with unnecessary guidelines.

However, if these results can be generalized to more complex tasks, the
relationship is beneficial and can be utilized. Currently, there is no formal training
program on how to use the Guidelines at The MITRE Corporation. Part of the
problem in developing such a program is that it is difficult to devise non-subjective
performance criteria. Tasks scenarios similar to the ones used in this thesis can be
used to measure trainee performance. The trainees' performance could then be
evaluated on the basis of setting a criterion at some specific percentage of expert
performance.
CONCLUSIONS

Effectiveness of Presentation Medium

Most likely, neither presentation medium is better for all possible tasks. Despite advances in hardware and software technology, paper is a good medium for communicating data. The quality with which data are presented in printed documents has evolved over the centuries; and it is likely that we shall not see books disappear in the near future.

Software presentation of text also has advantages. Hypertext-based design aids can provide users with additional tools and aids that go beyond "the text" and thus provide some additional value to the material. However, for software versions of printed documents (such as DRUID's implementation of Smith and Mosier's Guidelines for Designing User Interface Software) to be successful, they must be at least as usable as the original hardcopy version.

The results of this thesis have shown that in many ways, DRUID performed as well as the book for specific guideline selection tasks. Experienced as well as novice users were able to select guidelines effectively for two tasks scenarios. The only measure in which the book significantly outperformed the software was in selecting the total number of core guidelines, a very important measure. However, it should also be noted that the experiment was somewhat biased in favor of the book. When the book is used for real world tasks, there is no formatted "answer sheet" available (as there was for this study). When using the book, people must devise their own logging or checklist mechanism. A common method of logging has been to mark (such as checking or circling) the desired guidelines in the book and then to type the guideline data into a computer manually.
Fortunately, the software problems that were identified as causing the selection problem can most likely be alleviated with sufficient design revision. For example, if the next version of DRUID were to be written in a compilable programming language like C, many of the problems (speed, limited window size, font selection, etc.) could be significantly reduced or entirely eliminated. Additionally, with the tremendous rate at which hardware quality is increasing, while price is decreasing, software-related performance and display compromises can be minimized in the future.

Given those potential improvements, DRUID would clearly be the preferred medium. Not only is DRUID already subjectively preferred by the subjects in selecting guidelines, it also automates many of the more time-consuming tasks associated with selecting guidelines. DRUID's on-line checklist and hardcopy and softcopy report generation facilities significantly reduce the time it takes to perform guideline-related design and evaluation tasks.

But until those improvements to DRUID are made, potential users of the Guidelines should consider the task they need to perform before choosing which medium to use. For browsing, it may be advisable to use the book; and for creating checklists to use the software. When performing tasks that involve both browsing and selection, a user may wish use to both media to take advantage of their relative strengths.

Thus, it is possible to produce software tools that can effectively integrate large quantities of text without compromising user performance. Perhaps until the softcopy medium can always outperform the hardcopy medium for every task, it is the "value added" components (such as report generation) that will determine the success of a hypertext-based system.
Future Research

This thesis attempted to test the effects of guideline presentation media "outside the lab," by using experienced USI designers in realistic work settings. However, to encourage subject participation and maintain a high level of experimental control, the guideline selection tasks were simplified. Further studies need to be performed to evaluate the effectiveness of on-line design aids, such as DRUID, in controlled yet more realistic settings. Real world validity is of upmost importance to the end users of software tools. Longitudinal studies of how such tools are used in the workplace would provide greater insight into the usefulness and usability of those tools. With such long term studies, the effectiveness of software tools could be studied throughout the lifecycle of a system, instead of just studying subtasks in isolation.

In designing DRUID, the authors of the software tried to create an electronic version of the document. DRUID's structure and formatting closely resemble the original text. Software controls and aids were then incorporated on top of the text. As discussed earlier in the thesis, several other hypertext versions of the Guidelines also exist. Each has a unique interface and sequence control mechanism. Also, each software package provides a different level of conformance to the book's structure. Having different interfaces to the exact same text provides a unique opportunity to study how different aspects of the USI (structure, navigation aids, formatting, etc.) influence user performance in guideline-related tasks.

The results of this study demonstrate the value of added job aids to hypertexts to help users perform their tasks. Though DRUID currently provides aids to help users select guideline, it is easy to envision more sophisticated aids in applying guidelines. In the future, it may be possible to integrate a hypertext-based tool like DRUID with other forms of specification tools, such as rapid prototyping, state
diagrams, or behavioral representation techniques. Another useful area for future research would be in the development of an integrated technique that helps users decide which guidelines are relevant for a given set of requirements. Such an on-line assistant could significantly reduce the user's workload and potentially increase the user's recall and precision rates.
REFERENCES


Appendix A
Pre-test Questionnaire
Subject Number: ______

Age: ______

1. How many years of system design experience do you have?
   ____ less than 1 year
   ____ 1 year
   ____ 2 years
   ____ 3 years
   ____ 4 years
   ____ 5 years
   ____ more than 5 years

2. How many years of user-system interface (USI) design experience do you have?
   ____ less than 1 year
   ____ 1 year
   ____ 2 years
   ____ 3 years
   ____ 4 years
   ____ 5 years
   ____ more than 5 years

3. How many USIs have you been involved in designing?
   ________

4. For each interface that you helped design, briefly describe the interface and your role in the design. If you have designed more than five, list the five most recent.

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
5. How many college courses in USI design have you taken?
   ___ 0 course
   ___ 1 course
   ___ 2 courses
   ___ more than 2 courses

6. How many professional workshops, seminars, colloquia, or short courses on USI design have you attended or books you have read?
   ___ 0 course
   ___ 1 course
   ___ 2 courses
   ___ more than 2 courses
   ___ 0 books
   ___ 1 books
   ___ 2 books
   ___ more than 2 books

7. Have you ever used USI design guidelines before?
   ___ yes
   ___ no

8. If you answered yes to question number 7, please answer the following questions:

8.1. Have you ever used the Smith and Mosier Guidelines for Designing User Interface Software (book or software -- DRUID) before?
   ___ yes
   ___ no

8.2. How many times have you previously used design guidelines for each of the following tasks:

<table>
<thead>
<tr>
<th>DRUID</th>
<th>Book</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
</tr>
</tbody>
</table>


9. Do you prefer to work from a USI specification when designing USIs?
    ___ yes
    ___ no

Why or why not?

_________________________________________________________________
_________________________________________________________________
_________________________________________________________________
_________________________________________________________________

10. How many years of experience do you have using the Macintosh computer?
    ___ less than 6 months
    ___ between 6 months and a year
    ___ more than 1 year

11. How many years of experience do you have using the IBM PC (or clone) computer?
    ___ less than 6 months
    ___ between 6 months and a year
    ___ more than 1 year
Appendix B

Task Scenarios
ACME Tool Company

A software package is being developed for the ACME Tool Company. Your job is to design the USI for one component of the software for ACME's Personnel Director. You are to select those Data Display guidelines that apply to any aspect of the following description (as specified on instruction sheet).

Requirements:
For numerous tasks, the director needs to view the following detailed information about ACME's individual employees (assume employee records are to be viewed one at a time):

- name
- employee number
- office number
- telephone extension
- position (via a pre-designated code that is composed of numbers and letters)
- social security number
- salary
- date of hire
- home address (street, city, state, and zip code)
- comments (up to 50 lines of continuous text)

The director shall be able to access the information by directly entering an employee's name or employee number. The director also needs the capability to print the above employee information at a local printer.

User:
The personnel director of the company is 52 years old. He has an BA in Psychology. He is an infrequent computer user. He will be trained on the system before using it, and he will be able to get phone-based support from the MIS department when needed. A user's manual will be delivered with the system.

Implementation Considerations:
Hardware:
Because of budgetary constraints, the software which you are designing, will be running on a "dumb" terminal. The terminal can display text characters only. This particular terminal has a monochrome display.

Software:
All the software is being written in-house, but currently there are no in-house USI standards or guidelines. You may assume that any human factors specifications will be implemented as long as they are within the hardware constraints. The software cannot support windowing capabilities or graphics (can have text only).

Environment:
The director has his own office. The terminal will be placed on a computer desk in that office. The desk will be placed so that the director can talk on the phone while using the terminal. You may assume that the ambient environment conditions are satisfactory.

Simplifying Assumptions:
For this experiment, you are not concerned with how the data were entered into the system or compatibility with any other aspect of the system. Also, new data will only be entered into the system after work hours, so there will be no new updates while the director uses the software.
American Widgets Inc.

A software package is being developed for the Head of Marketing for American Widgets Inc. Your job is to design the USI for the software. You are to select those Data Display guidelines that apply to any aspect of the following description (as specified on instruction sheet).

**Requirements:**
The Head of Marketing needs to view integrated data about the company's sales. The software needs to allow the user to:

View the actual sales figures (in dollars) for the company over any previous 12 months period, thus the Head of Marketing needs to be able specify the starting month to be displayed. For reference, the predicted sales figures and the profits for those same months also need to be displayed. Any monthly actual sales figures that fall more than 20% below or 20% above the predicted sales figure should be noted.

The Head of Marketing also needs the capability to print the above sales data at a local printer.

**User**
The Head of Marketing of the company is 40 years old and has an MBA. She has used an IBM PC many times before. She will receive no special training on the system, but she will have a user's manual for the software.

**Implementation Considerations**

**Hardware:**
The software will run on a PC. The PC has a monochrome display. The PC uses DOS for an operating system.

**Software:**
All the software is being written in-house. You may assume that any human factors specifications will be implemented as long as they are within the hardware constraints. The software cannot support windowing capabilities or graphics (can have text only).

**Environment:**
The Head of Marketing has her own office. The computer will be placed on her work desk. You may assume that the ambient environment conditions are satisfactory.

**Simplifying Assumptions:**
For this experiment, you are not concerned with how the data were entered into the system or compatibility with any other aspect of the system. Also, new data will only be entered into the system after work hours, so there will be no new updates while the Head of Marketing uses the software.
Appendix C
Informed Consent Form
Informed Consent Form

This study intends to study the effects of using a computer-based aid for the selection of USI design guidelines. Please keep in mind that the system and the process are under study, not you. During the study, you will be videotaped. You have certain rights as explained below. This form will enumerate these rights and obtain your written consent to participate in the study. Your rights as a participant are:

1. You have the right to withdraw from this study at any time and for any reason by simply informing the experimenter. If you decide to terminate your participation, you will still receive a copy of the DRUID software.

2. You have the right to inspect your data and withdraw them from the experiment if you feel that you should for any reason. Data are processed and analyzed after a subject has completed the study. At that time, these data will be treated with anonymity since all identifying information will be removed from the data. No one else will have access to your data. The experiment will be videotaped. The data from the tape will be used for data analysis only, and will be erased after the experiment.

3. You have the right to be informed of the overall results of this study (if you so desire). If you wish to receive an overview of the results, include your address with your signature below.

4. If you have any questions prior to data collection please feel free to ask. Unless the outcome of the study will be influenced, the researcher will provide an answer to your satisfaction. Answers which may influence the study will be delayed until after data collection, at which time a full answer will be given.

5. There is minimal risk to you as a participant in this study. Since you are an experienced computer user, there is no risk involved in this study beyond what might be experienced in everyday computer usage.

6. You will receive a free copy of DRUID 2.0 software as payment for your time.

If you have further comments or questions about your rights as a participant, please contact Dr. Ernest Stout, chairperson of the Institutional Review Board for the Use of Human Subjects in Research. He may be contacted at:

Chairman, University Human Subjects Committee
301 Burress Hall
Virginia Polytechnic Institute and State University
Blacksburg, VA 24061
(703) 231-5283
The researcher for this study will be Jeffrey A. Fox, a graduate student in Industrial and Systems Engineering at Virginia Polytechnic Institute and State University and also a member of the Technical Staff at The MITRE Corporation.

Your signature below indicates that you have read this document in its entirety, and understand your rights as a participant as listed above, and that you consent to participate. Thank you for your cooperation.

I have read a description of this study and understand the nature of the research and my rights as a participant. I hereby consent to participate, with the understanding that I may discontinue participation at any time if I so choose.

________________________
Participant's Signature

________________________
Printed Name

________________________
Date
Appendix D
Test Tasks for Training
Hardcopy

Please answer the following questions about the Guidelines. Answer the questions aloud to the experimenter.

Using the **table of contents**, answer the following questions below:

- What page does the function on "Job Aids" start?
- What are the numbers and names of the functions listed under the functional area "DATA PROTECTION"?

Using the **glossary**, answer the following question:

- What is the definition for "Interrupt"?

Using the **index**, answer the following question:

- What is the number and title of the guideline referenced by the term "Blank Space"?

Using the appropriate list of **guideline titles**, answer the following questions:

- How many guidelines are there in function "6.1 User Identification"?
- What is the title of guideline "1.1-7"?

Using the **individual guidelines**, answer the following questions:

- What is the title of guideline "5.0-1"?
- List the numbers and titles for the cross-references to other guidelines for guideline "4.0-16"
- What is the guideline statement for guideline "3.1.3-22"?
- What is the definition for function "5.5 Receiving Messages"?
Softcopy

Please answer the following questions using DRUID. Answer the questions aloud to the experimenter.

First, start DRUID 2.0 by double-clicking on its icon and click on the title display to continue. Next select the Review Guidelines option to begin the training session.

Use the MOUSE to navigate to the appropriate displays that lists the information necessary to answer the following three questions:

What are the numbers and names of the functions listed under the functional area "DATA TRANSMISSION"?

Using display "1 DATA ENTRY", answer the following question:

What is the definition for function "1.2 Direction Designation"?

Using display "6.0-1", answer the following question:

What is the definition for function "1.0 General"?

Use the KEYBOARD and Command line only to navigate to the display that lists the information necessary to answer the following questions:

Using the glossary, answer the following question:

What is the definition for "Job"? (key in the search term "Job")

Using the index, answer the following question:

What are the numbers and titles of the guideline referenced as dealing with the term "Arrows"? (key in the search term "Arrows")

Using the appropriate list of guideline titles, answer the following questions:

How many guidelines are there in function "6.2 Data Access"?

What is the title of guideline "4.1-3"?
Using the appropriate guidelines display, answer the following questions:

What is the title of guideline "5.0-1"?

List the numbers and titles for the cross-references to other guidelines for
guideline "4.0-16"

What is the guideline statement for guideline "3.13-22"

Now please perform the following tasks with DRUID. You may use either the mouse
or command line for any of these tasks.

1. Create a new DRUID data file called "Training". From the Identify System
   Application (ISA) display (No other file information is required.)

2. Choose the Select Functions/Guidelines option on the Design Options Display.

3 From the guideline title list "1.0", select the first five guidelines by keying "y"
   into the appropriate fields.

4. From the guideline display "1.0-12", select the guideline by keying "y" into the
   appropriate field,

5. Save the file by typing "save" into the command line from the display you are
   now viewing.
Appendix E

Instructions
Instructions -- Hardcopy

Select those guidelines that you think apply to the attached system description. For this experiment, you shall only select guidelines from the Data Display section of the Guidelines (section 2). For this experiment data display shall be defined as:

"The computer output (monitor or printout) of data to a user, and the assimilation of that information from such outputs. Elements of data display are the actual task-related data, as well as labels, prompts, computer-generated advisory messages and other user guidance."

The guidelines that you select will be used to generate the design rules from which the USI will be built (as discussed in the beginning of the experiment). You are to select the guidelines as efficiently and as accurately as possible.

You will use a hardcopy of the Smith and Mosier Guidelines to select the guidelines. For each guideline that you think applies, write a "Y" into the "Selection" column on the answer sheet next to the appropriate guideline's number. For any guideline that you are unsure of (either because you don't understand its means or because you are unsure if it applies), write down a "?" If you mark a guideline with a "?", please explain why out loud to the experimenter.

You may use any strategy for selecting guidelines (i.e., using the individual guidelines, the index, title lists, cross-referencing, etc.) you desire. For this experiment, you are not required to view every guideline within the Data Display section unless you prefer to select guidelines in that manner. Remember, you should perform your work efficiently and as accurately as possible.

You have been provided several pieces of blank paper. You can use that paper in any way that may be helpful in completing the task, including making any notes or drawings.

At any time during the experiment, you may revise your list of guidelines. You can erase the guideline selections you no longer think are appropriate; and you may also go back to any place and select additional guidelines.

As you select guidelines, you are encouraged to think aloud about any decisions that you are making in the selection process. You do not need to do this for each guideline unless you so desire.

You may take as much time as you like in selecting the guidelines. We will all take a break about two hours from now. However, please feel free to take additional breaks at any time.
Instructions -- Softcopy

Select those guidelines that you think apply to the attached system description. For this experiment, you shall only select guidelines from the **Data Display** section of the **Guidelines**. (section 2). For this experiment data display shall be defined as:

"The computer output of data to a user (monitor or printout), and assimilation of information from such outputs. Elements of data display are the actual task-related data, as well as labels, prompts, computer-generated advisory messages and other user guidance."

The guidelines that you select will be used to generate the design rules from which the US1 will be built (as discussed in the beginning of the experiment). You are to select the guidelines as efficiently and as accurately as possible.

You will use DRUID to select the guidelines. For each guideline that you think applies, type a "Y" into that guideline's "Weight" column. For any guideline that you are unsure of (either because you don't understand its means or because you are unsure if it applies), type a "?" into that guideline's "Weight" column. If you mark a guideline with a "?", please explain why out loud to the experimenter.

Before selecting any guidelines, first create a new file with the same name as the company described in the scenario(s). When you are done selecting guidelines, please save the file.

You may use any strategy for selecting guidelines (i.e., using the individual guidelines, the index, title lists, cross-referencing, etc.) you desire. For this experiment, you are not required to view every guideline within the Data Display section unless you prefer to select guidelines in that manner. Remember, you should perform your work efficiently and as accurately.

You have been provided several pieces of blank paper. You can use that paper in any way that may be helpful in completing the task, including making any notes or drawings.

At any time during the experiment, you may revise your list of guidelines. You can remove the Y for those guidelines you no longer think are appropriate; and you may also go back to any place and select additional guidelines.

As you select guidelines, you are encouraged to think aloud about any decisions that you are making in the selection process. You do not need to do this for each guideline unless you so desire.

You may take as much time as you like in selecting the guidelines. We will all take a break about two hours from now. However, please feel free to take additional breaks at any time.
Appendix F
Hardcopy Answer Sheet
### ANSWER SHEET

<table>
<thead>
<tr>
<th>Guideline</th>
<th>Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.0-1</td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td></td>
</tr>
<tr>
<td>-3</td>
<td></td>
</tr>
<tr>
<td>-4</td>
<td></td>
</tr>
<tr>
<td>-5</td>
<td></td>
</tr>
<tr>
<td>-6</td>
<td></td>
</tr>
<tr>
<td>-7</td>
<td></td>
</tr>
<tr>
<td>-8</td>
<td></td>
</tr>
<tr>
<td>-9</td>
<td></td>
</tr>
<tr>
<td>-10</td>
<td></td>
</tr>
<tr>
<td>-11</td>
<td></td>
</tr>
<tr>
<td>-12</td>
<td></td>
</tr>
<tr>
<td>-13</td>
<td></td>
</tr>
<tr>
<td>-14</td>
<td></td>
</tr>
<tr>
<td>-15</td>
<td></td>
</tr>
<tr>
<td>-16</td>
<td></td>
</tr>
<tr>
<td>-17</td>
<td></td>
</tr>
<tr>
<td>-18</td>
<td></td>
</tr>
<tr>
<td>-19</td>
<td></td>
</tr>
<tr>
<td>-20</td>
<td></td>
</tr>
<tr>
<td>-21</td>
<td></td>
</tr>
<tr>
<td>2.1-1</td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td></td>
</tr>
<tr>
<td>-3</td>
<td></td>
</tr>
<tr>
<td>-4</td>
<td></td>
</tr>
<tr>
<td>-5</td>
<td></td>
</tr>
<tr>
<td>-6</td>
<td></td>
</tr>
<tr>
<td>-7</td>
<td></td>
</tr>
<tr>
<td>-8</td>
<td></td>
</tr>
<tr>
<td>-9</td>
<td></td>
</tr>
<tr>
<td>-10</td>
<td></td>
</tr>
<tr>
<td>-11</td>
<td></td>
</tr>
<tr>
<td>-12</td>
<td></td>
</tr>
<tr>
<td>-13</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Guideline</th>
<th>Selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1-14</td>
<td></td>
</tr>
<tr>
<td>-15</td>
<td></td>
</tr>
<tr>
<td>-16</td>
<td></td>
</tr>
<tr>
<td>-17</td>
<td></td>
</tr>
<tr>
<td>-18</td>
<td></td>
</tr>
<tr>
<td>-19</td>
<td></td>
</tr>
<tr>
<td>-20</td>
<td></td>
</tr>
<tr>
<td>-21</td>
<td></td>
</tr>
<tr>
<td>-22</td>
<td></td>
</tr>
<tr>
<td>-23</td>
<td></td>
</tr>
<tr>
<td>-24</td>
<td></td>
</tr>
<tr>
<td>-25</td>
<td></td>
</tr>
<tr>
<td>-26</td>
<td></td>
</tr>
<tr>
<td>-27</td>
<td></td>
</tr>
<tr>
<td>-28</td>
<td></td>
</tr>
<tr>
<td>-29</td>
<td></td>
</tr>
<tr>
<td>2.2-1</td>
<td></td>
</tr>
<tr>
<td>-2</td>
<td></td>
</tr>
<tr>
<td>-3</td>
<td></td>
</tr>
<tr>
<td>-4</td>
<td></td>
</tr>
<tr>
<td>-5</td>
<td></td>
</tr>
<tr>
<td>-6</td>
<td></td>
</tr>
<tr>
<td>-7</td>
<td></td>
</tr>
<tr>
<td>-8</td>
<td></td>
</tr>
<tr>
<td>-9</td>
<td></td>
</tr>
<tr>
<td>-10</td>
<td></td>
</tr>
<tr>
<td>-11</td>
<td></td>
</tr>
<tr>
<td>-12</td>
<td></td>
</tr>
<tr>
<td>-13</td>
<td></td>
</tr>
<tr>
<td>-14</td>
<td></td>
</tr>
<tr>
<td>-15</td>
<td></td>
</tr>
</tbody>
</table>
Appendix G
Post-test Questionnaire
Subject Number: ________

Scenario: ________

Media: ________

1. How confident are you that you found all the guidelines that applied to this scenario?

   1 2 3 4 5 6 7  Don't Know
   Very
   Unsure
   Neutral
   Very
   Confident

2. Did you lose track of where you were in the guidelines during the task?

   1 2 3 4 5 6 7  Don't Know
   Never
   Neutral
   Always

3. Did you have to remember a lot of information while performing the task?

   1 2 3 4 5 6 7  Don't Know
   Very
   to remember
   Neutral
   Too much
   to remember

4. Recording the selected guidelines was:

   1 2 3 4 5 6 7  Don't Know
   Very
   Difficult!
   Neutral
   Very
   Easy

5. How would you rate the actual guideline text (not the presentation medium) in terms of visual clarity (organization, ambiguity, and readability)?

   1 2 3 4 5 6 7  Don't Know
   Very
   Unsatisfactory
   Neutral
   Very
   Satisfactory

6. How would you rate the presentation medium (document or software) in terms of visual clarity (organization and readability)?

   1 2 3 4 5 6 7  Don't Know
   Very
   Unsatisfactory
   Neutral
   Very
   Satisfactory
7. How would you rate the document or software in terms of consistency?

1  2  3  4  5  6  7  Don't Know
 Very Inconsistent
Neutral
Very Consistent

8. How well did the document or software meet your needs?

1  2  3  4  5  6  7  Don't Know
 Very Poorly
Neutral
Very Well

9. How would you rate the document or software in terms of providing aids in performing the task?

1  2  3  4  5  6  7  Don't Know
 Very Unsatisfactory
Neutral
Very Satisfactory

10. Did the document or software provide you with all the necessary information you needed to perform your task?

____ yes
____ no

If not, what information was missing?
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

11. List three things you liked about the document or software.
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
12. List three things you did not like about the document or software.

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

13. Describe any parts of the document or software which were confusing or difficult to understand.

________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

14. Did the document or software help you in selecting guidelines.

   ___ yes
   ___ no

   If so, how?
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

15. Did the document or software hinder you in any way in selecting guidelines.

   ___ yes
   ___ no

   If so, how?
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
16. Did the presentation media (document or software) influence your design process?

   ____ yes
   ____ no

   How?

   ____________________________________________________________
   ____________________________________________________________
   ____________________________________________________________
   ____________________________________________________________
   ____________________________________________________________

17. Did you find the function definition useful in selecting guidelines?

   ____ yes
   ____ no

   Why or why not?

   ____________________________________________________________
   ____________________________________________________________
   ____________________________________________________________
   ____________________________________________________________
   ____________________________________________________________
Please sketch below the interface design for the scenario. If you have already sketched a design, you may include it with this questionnaire. Also, please briefly describe any aspects of the "look and feel" of the design that you think are necessary to help describe it.
Appendix H
Summary Questionnaire
Subject Number: __________

1. Did you "picture" the design in your mind during the tasks?
   
   ___ yes
   ___ no

   If yes, at which stage of the task did you first picture the design?
   
   ___ before you selected any guidelines
   ___ while you selected any guidelines
   ___ after you selected all guidelines

   If you answered "before" or "while" in the last question, did you refine your design as you selected more guidelines?
   
   ___ yes
   ___ no

2. Which task, if any, was more difficult to perform
   
   ___ same
   ___ personnel system
   ___ marketing system

   If one was more difficult, why?
   ________________________________________________________________
   ________________________________________________________________
   ________________________________________________________________
   ________________________________________________________________
   ________________________________________________________________

3. Did the size of the systems for which you selected guidelines affect your guideline selection? (for example: total number selected, type selected)
   
   ___ yes
   ___ no

   If so, how?
   ________________________________________________________________
   ________________________________________________________________
   ________________________________________________________________
   ________________________________________________________________
   ________________________________________________________________
4. Which presentation media did you prefer?

   ____ document (hardcopy)
   ____ software (DRUID)
   ____ equal
   ____ it depends

Why?
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

5. Which presentation media was easier to use?

   ____ document (hardcopy)
   ____ software (DRUID)

Why?
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________

6. Did the presentation media (document or software) influence your strategy for selecting guidelines?

   ____ yes
   ____ no

If so, how?
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
________________________________________________________________________
7. Did the presentation media (document or software) influence your overall design process?
   ____ yes
   ____ no

How?

8. If you had to perform a similar task again, which would you use?
   ____ document (hardcopy)
   ____ software (DRUID)
   ____ both
   ____ other

If other, what?

           ________________________________________________________________
           ________________________________________________________________
           ________________________________________________________________
           ________________________________________________________________
           ________________________________________________________________
Appendix I
List of Guidelines Selected by Experts
Guidelines for ACME Scenario

2.0 General
- 1 Necessary data displayed
- 4 Data display consistent with user conventions
- 5 Establishing display standards
- 6 Consistent display format
- 11 Context for displayed data
- 12 Familiar wording
- 16 Minimal use of abbreviation
- 17 Common abbreviations

2.1 Text
- 1 Conventional text display
- 4 Adequate display capacity
- 5 Text displayed in wide columns
- 6 Conventional use of mixed case
- 7 Separation of paragraphs
- 8 Consistent word spacing
- 9 Minimal hyphenation
- 10 Conventional punctuation

2.2 Data Forms
- 1 Forms for related data
- 2 Visually distinctive data fields
- 3 Data field labeling
- 4 Descriptive wording of labels
- 7 Consistent label location
- 8 Distinctive label format
- 9 Labels close to data fields
- 11 Consistent format across displays
- 14 Partitioning long data items

2.5 Format
- 1 Consistent format
- 4 Paging crowded displays
- 10 Display title at top
- 11 Command entry, prompts, messages at bottom

2.7.1 Display Control - Selection
- 1 User selection of data for display
- 11 Printing displays locally

2.7.2 Display Control - Framing
- 2 Easy paging

2.8 Design Change
- 1 Flexible design for data display
Guidelines for Widgets Scenario

2.0 General
- 1 Necessary data displayed
- 2 Only necessary data displayed
- 3 Data displayed in usable form
- 4 Data display consistent with user conventions
- 5 Establishing display standards

2.1 Text
- 1 Conventional text display
- 2 Printing lengthy text displays
- 3 Consistent text format
- 6 Conventional use of mixed case
- 7 Separation of paragraphs
- 8 Consistent word spacing
- 11 Clarity of wording
- 12 Sentences begin with main topic
- 13 Simple sentence structure
- 14 Concise wording
- 16 Affirmative sentences

2.5 Format
- 3 Spacing to structure displays
- 7 Integrated display
- 10 Display title at top
- 11 Command entry, prompts, messages at bottom
- 12 Logical data organization
- 13 Grouping for data comparison

2.6 Coding
- 1 Highlighting critical data
- 12 Special symbols
- 13 Consistent use of special symbols

2.7 Display Control
- 1 Flexible display control by user

2.7.2 Display Control - Framing
- 1 Integrated display
- 11 Labeling scrolling functions
Appendix J
ANOVA Summary Tables
THREE-WAY ANOVAS:

Measure: Time to complete task

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium</td>
<td>1</td>
<td>554.803</td>
<td>2.155</td>
<td>0.1803</td>
</tr>
<tr>
<td>Experience</td>
<td>1</td>
<td>46.345</td>
<td>0.180</td>
<td>0.6825</td>
</tr>
<tr>
<td>Scenario</td>
<td>1</td>
<td>16.166</td>
<td>0.063</td>
<td>0.8084</td>
</tr>
<tr>
<td>Medium x Experience</td>
<td>1</td>
<td>1.916</td>
<td>0.007</td>
<td>0.9334</td>
</tr>
<tr>
<td>Medium x Scenario</td>
<td>1</td>
<td>1.811</td>
<td>0.007</td>
<td>0.9352</td>
</tr>
<tr>
<td>Experience x Scenario</td>
<td>1</td>
<td>0.616</td>
<td>0.002</td>
<td>0.9622</td>
</tr>
<tr>
<td>Medium x Experience x</td>
<td>1</td>
<td>88.466</td>
<td>0.386</td>
<td>0.5515</td>
</tr>
<tr>
<td>Scenario</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>6</td>
<td>257.420</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Measure: Core guidelines selected - percent correct (Recall)

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium</td>
<td>1</td>
<td>0.025</td>
<td>2.105</td>
<td>0.1849</td>
</tr>
<tr>
<td>Experience</td>
<td>1</td>
<td>0.022</td>
<td>1.846</td>
<td>0.2114</td>
</tr>
<tr>
<td>Scenario</td>
<td>1</td>
<td>5.256E-5</td>
<td>0.004</td>
<td>0.9483</td>
</tr>
<tr>
<td>Medium x Experience</td>
<td>1</td>
<td>0.001</td>
<td>0.109</td>
<td>0.7500</td>
</tr>
<tr>
<td>Medium x Scenario</td>
<td>1</td>
<td>0.007</td>
<td>0.626</td>
<td>0.4517</td>
</tr>
<tr>
<td>Experience x Scenario</td>
<td>1</td>
<td>2.256E-5</td>
<td>0.002</td>
<td>0.9661</td>
</tr>
<tr>
<td>Medium x Experience x</td>
<td>1</td>
<td>0.001</td>
<td>0.109</td>
<td>0.7500</td>
</tr>
<tr>
<td>Scenario</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>8</td>
<td>0.012</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Measure: Percent picked from list

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Medium</td>
<td>1</td>
<td>961.000</td>
<td>2.337</td>
<td>0.1649</td>
</tr>
<tr>
<td>Experience</td>
<td>1</td>
<td>81.000</td>
<td>0.197</td>
<td>0.6689</td>
</tr>
<tr>
<td>Scenario</td>
<td>1</td>
<td>144.000</td>
<td>0.350</td>
<td>0.5704</td>
</tr>
<tr>
<td>Medium x Experience</td>
<td>1</td>
<td>25.000</td>
<td>0.061</td>
<td>0.8115</td>
</tr>
<tr>
<td>Medium x Scenario</td>
<td>1</td>
<td>676.000</td>
<td>1.644</td>
<td>0.2357</td>
</tr>
<tr>
<td>Experience x Scenario</td>
<td>1</td>
<td>64.900</td>
<td>0.156</td>
<td>0.7035</td>
</tr>
<tr>
<td>Medium x Experience x</td>
<td>1</td>
<td>484.000</td>
<td>1.177</td>
<td>0.3096</td>
</tr>
<tr>
<td>Scenario</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Residual</td>
<td>8</td>
<td>411.250</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## TWO-WAY ANOVAS

**Measure: Time to complete task**

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experience</td>
<td>1</td>
<td>46.345</td>
<td>0.276</td>
<td>0.6184</td>
</tr>
<tr>
<td>S/Experience</td>
<td>6</td>
<td>168.145</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>1</td>
<td>554.803</td>
<td>2.849</td>
<td>0.1424</td>
</tr>
<tr>
<td>Medium x Experience</td>
<td>1</td>
<td>1.916</td>
<td>0.010</td>
<td>0.9242</td>
</tr>
<tr>
<td>Medium x S/Experience</td>
<td>6</td>
<td>194.758</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Measure: Core guidelines selected - percent correct (Recall)**

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experience</td>
<td>1</td>
<td>0.022</td>
<td>1.344</td>
<td>0.2904</td>
</tr>
<tr>
<td>S/Experience</td>
<td>6</td>
<td>0.016</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scenario</td>
<td>1</td>
<td>0.025</td>
<td>25.126</td>
<td>0.0024</td>
</tr>
<tr>
<td>Scenario x Experience</td>
<td>1</td>
<td>0.001</td>
<td>1.299</td>
<td>0.2979</td>
</tr>
<tr>
<td>Scenario x S/Experience</td>
<td>6</td>
<td>0.001</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experience</td>
<td>1</td>
<td>0.022</td>
<td>1.344</td>
<td>0.2904</td>
</tr>
<tr>
<td>S/Experience</td>
<td>6</td>
<td>0.016</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Scenario</td>
<td>1</td>
<td>5.256E-5</td>
<td>0.010</td>
<td>0.9240</td>
</tr>
<tr>
<td>Scenario x Experience</td>
<td>1</td>
<td>2.256E-5</td>
<td>0.004</td>
<td>0.9501</td>
</tr>
<tr>
<td>Scenario x S/Experience</td>
<td>6</td>
<td>0.005</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Measure: Percent picked from list

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experience</td>
<td>1</td>
<td>0.001</td>
<td>0.014</td>
<td>0.9110</td>
</tr>
<tr>
<td>S/Experience</td>
<td>6</td>
<td>0.052</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>1</td>
<td>0.099</td>
<td>5.483</td>
<td>0.0577</td>
</tr>
<tr>
<td>Medium x Experience</td>
<td>1</td>
<td>0.014</td>
<td>0.755</td>
<td>0.4184</td>
</tr>
<tr>
<td>Medium x S/Experience</td>
<td>6</td>
<td>0.018</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Measure: Precision

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experience</td>
<td>1</td>
<td>0.001</td>
<td>0.014</td>
<td>0.9110</td>
</tr>
<tr>
<td>S/Experience</td>
<td>6</td>
<td>0.052</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>1</td>
<td>0.005</td>
<td>0.141</td>
<td>0.7201</td>
</tr>
<tr>
<td>Medium x Experience</td>
<td>1</td>
<td>0.004</td>
<td>0.106</td>
<td>0.7562</td>
</tr>
<tr>
<td>Medium x S/Experience</td>
<td>6</td>
<td>0.033</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Measure: Number of Definitions

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experience</td>
<td>1</td>
<td>20.250</td>
<td>0.771</td>
<td>0.4136</td>
</tr>
<tr>
<td>S/Experience</td>
<td>6</td>
<td>26.250</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>1</td>
<td>42.250</td>
<td>1.008</td>
<td>0.3542</td>
</tr>
<tr>
<td>Medium x Experience</td>
<td>1</td>
<td>90.250</td>
<td>2.153</td>
<td>0.1927</td>
</tr>
<tr>
<td>Medium x S/Experience</td>
<td>6</td>
<td>41.917</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Measure: Number of Times Glossary Was Used

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experience</td>
<td>1</td>
<td>5.062</td>
<td>4.765</td>
<td>0.0718</td>
</tr>
<tr>
<td>S/Experience</td>
<td>6</td>
<td>1.062</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>1</td>
<td>1.562</td>
<td>1.271</td>
<td>0.3026</td>
</tr>
<tr>
<td>Medium x Experience</td>
<td>1</td>
<td>1.562</td>
<td>1.271</td>
<td>0.3026</td>
</tr>
<tr>
<td>Medium x S/Experience</td>
<td>6</td>
<td>1.229</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
### Measure: Number of Cross References

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experience</td>
<td>1</td>
<td>0.562</td>
<td>0.184</td>
<td>0.6832</td>
</tr>
<tr>
<td>S/Experience</td>
<td>6</td>
<td>3.062</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>1</td>
<td>0.062</td>
<td>0.038</td>
<td>0.8519</td>
</tr>
<tr>
<td>Medium x Experience</td>
<td>1</td>
<td>1.562</td>
<td>0.949</td>
<td>0.3675</td>
</tr>
<tr>
<td>Medium x S/Experience</td>
<td>6</td>
<td>1.646</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Measure: Number of Times the Index Was Used

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experience</td>
<td>1</td>
<td>2.250</td>
<td>1.000</td>
<td>0.3559</td>
</tr>
<tr>
<td>S/Experience</td>
<td>6</td>
<td>2.250</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>1</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium x Experience</td>
<td>1</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium x S/Experience</td>
<td>6</td>
<td>N/A</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Measure: Number of Times Samples Were Used

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experience</td>
<td>1</td>
<td>0.562</td>
<td>1.174</td>
<td>0.3202</td>
</tr>
<tr>
<td>S/Experience</td>
<td>6</td>
<td>0.479</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>1</td>
<td>0.062</td>
<td>0.045</td>
<td>0.8394</td>
</tr>
<tr>
<td>Medium x Experience</td>
<td>1</td>
<td>0.063</td>
<td>0.045</td>
<td>0.8394</td>
</tr>
<tr>
<td>Medium x S/Experience</td>
<td>6</td>
<td>1.396</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### Measure: Guidelines per Minute

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experience</td>
<td>1</td>
<td>1.000E-4</td>
<td>0.003</td>
<td>0.9551</td>
</tr>
<tr>
<td>S/Experience</td>
<td>6</td>
<td>0.029</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>1</td>
<td>0.084</td>
<td>6.007</td>
<td>0.0497</td>
</tr>
<tr>
<td>Medium x Experience</td>
<td>1</td>
<td>0.010</td>
<td>0.714</td>
<td>0.4304</td>
</tr>
<tr>
<td>Medium x S/Experience</td>
<td>6</td>
<td>0.014</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Measure: Number of Times the Table of Contents Was Used

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experience</td>
<td>1</td>
<td>1.000</td>
<td>0.375</td>
<td>0.5628</td>
</tr>
<tr>
<td>S/Experience</td>
<td>6</td>
<td>2.667</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>1</td>
<td>0.250</td>
<td>0.073</td>
<td>0.7958</td>
</tr>
<tr>
<td>Medium x Experience</td>
<td>1</td>
<td>6.250</td>
<td>1.829</td>
<td>0.2250</td>
</tr>
<tr>
<td>Medium x S/Experience</td>
<td>6</td>
<td>3.417</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Measure: Number of Guidelines Subjects Were Unsure About

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>MS</th>
<th>F</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experience</td>
<td>1</td>
<td>10.562</td>
<td>0.661</td>
<td>0.4473</td>
</tr>
<tr>
<td>S/Experience</td>
<td>6</td>
<td>15.979</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Medium</td>
<td>1</td>
<td>18.062</td>
<td>1.724</td>
<td>0.2372</td>
</tr>
<tr>
<td>Medium x Experience</td>
<td>1</td>
<td>1.563</td>
<td>0.149</td>
<td>0.7127</td>
</tr>
<tr>
<td>Medium x S/Experience</td>
<td>6</td>
<td>10.479</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix K
Scenario Interface Drawings
Personnel Record

NAME:  
Address:  
SSN:  

Employee No:  
Position:  
Date of Hire:  
Office No:  
Telephone Ext:  
SALARY:  

Comments:  

Enter:  NAME  EMPLOYEE NO.  

Personnel Record (Continued)  

Enter:  NAME  EMPLOYEE NO.  

4 lines
N: NAME

Y: 2013
C: __________

Comments: __________________________

50 __________________________
ACME Tool Company Employee Records

Name ___________  Emp # ______  

Enter either to select

Selection Display

ACME Tool Company Employee Records

Name ___________  Emp # ______  

SSN # ______

Home address ______

Comments ______

Information Display on employee selected

NEXT EMP  PREV. EMP  RETURN  HELP
Period: 28 June 90 - 29 June 91

<table>
<thead>
<tr>
<th>Month</th>
<th>Actual ($)</th>
<th>Planned ($)</th>
<th>% Profits ($)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

6 91

> Q: Grant enters new month - P. Hunt
Sales from Jan 1990 to Date

Jan 1 Feb 1
Sales previous previous
Project project

Enter start month (MM/YY):
<table>
<thead>
<tr>
<th>Month</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Product Category</td>
</tr>
<tr>
<td></td>
<td>Actual Sales Figures</td>
</tr>
<tr>
<td></td>
<td>Predicted Sales Figures</td>
</tr>
</tbody>
</table>

**DetailedFoot**

- Entire year displayed on single screen.
- Monthly sales figures above or below 20% of predicted highlighted in some way.
XYZ CORPORATION
SALES ($1000)

Actual Sales
13.2  9.8  17.6  13.7

Predicted Sales
13.0  13.0  13.3  13.8

Profit (Loss)
5.2  2.0

User-selected starting month

Justify decimal point

code/highlight as >20% below forecast

code/highlight as >20% above forecast
<table>
<thead>
<tr>
<th>MONTH</th>
<th>JAN</th>
<th>FEB</th>
<th>MAR</th>
<th>APR</th>
<th>MAY</th>
<th>JUN</th>
<th>JUL</th>
<th>AUG</th>
<th>SEP</th>
<th>OCT</th>
<th>NOV</th>
<th>DEC</th>
</tr>
</thead>
</table>

**Actual Sales:**
- 
- 
- 
- 
- 
- 
- 
- 
- 
- 
- 

**Projected Sales:**
- 
- 
- 
- 
- 
- 
- 
- 
- 
- 
- 

**Profits:**
- 
- 
- 
- 
- 
- 
- 
- 
- 
- 
- 

**Enter Starting Month:**
<table>
<thead>
<tr>
<th>Month</th>
<th>Scheduled</th>
<th>Actual</th>
<th>% Change</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$XXX,XXX</td>
<td>$XXX,XXX</td>
<td>+32</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-10</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>-05</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>+30</td>
</tr>
</tbody>
</table>
VITA

Jeffrey A. Fox

Date of birth: June 15, 1965

EDUCATION

Virginia Polytechnic Institute and State University, Blacksburg, VA
Master of Science, January 1992
Industrial and Systems Engineering - Human Factors Engineering

University of Massachusetts, Amherst, MA
Bachelor of Science, Cum Laude, May 1987
Industrial Engineering/Operations Research - Human Factors Engineering

HONORS AND AWARDS

Tau Beta Pi
Alpha Pi Mu
National Engineering Award Winner
Engineering Alumni Research Scholarship - University of Massachusetts

RELEVANT EXPERIENCE

1987-1991 THE MITRE CORPORATION - Member of the Technical Staff
Developed DRUID (Dynamic Rules for User Interface Design), a hypertext tool for accessing and applying HCI design guidelines using HyperCard software.
- Wrote most of the software, performed usability tests, and wrote user documentation.
- Publicized this work through publications and professional presentations.

Led human factors efforts in the definition and acquisition of an automated mission planning system for the USAF.
- Wrote human factors specifications, including defining new ways of applying rapid prototyping and usability testing to the acquisition process.
- Monitored and evaluated contractor's technical work.

Ensured good HCI design for internally developed MIS applications for IBM mainframes.
- Created and applied HCI design rules for new applications.
- Evaluated existing applications and prototyped new designs.

1986-1987 Technical Assistant
As a coop, designed and programmed the user interface for C2 decision aids software.

1990 INDUSTRIAL AND SYSTEMS ENGINEERING DEPT. (VPI & SU) - Teaching Assistant
Taught six recitation sections (per week) of Engineering Economics.

1985-1987 INDUSTRIAL ENGINEERING DEPT. (UNIV. OF MASSACHUSETTS ) - Research Assistant
Developed and evaluated strategies for implementing computer systems and office automation.
PROFESSIONAL PUBLICATIONS AND PRESENTATIONS


Demonstrated DRUID and discussed the use of on-line guidelines at:
- DOD HFE TG (Human Factors Engineering Technical Group) Meeting (May 89, 90, 91)
- 33rd Annual Meeting of the Human Factors Society (October 89)
- Hypertext II - York, England (June 89)
- ACM SIGCHI ’89 Conference (May 89)
- Tufts University user interface design class - Guest Lecturer (March 89)
- New England Chapter of the Human Factors Society (March 89)
- 32nd Annual Meeting of the Human Factors Society (October 88)

PROFESSIONAL ACTIVITIES

Human Factors Society:
- National Chapter: Student member
- Virginia Tech Chapter: Chair of Guest Speaker Committee 1990-1991
- New England Chapter: Director 1988, Secretary 1989, Treasurer 1990
- University of Massachusetts: Started student Chapter 1987

ACM SIGCHI: Student member