1

**PM-1 3½"x4" PHOTOGRAPHIC MICROCOPY TARGET
NBS 1010a ANSI/ISO #2 EQUIVALENT**

PRECISIONSM RESOLUTION TARGETS

PIONEERS IN METHYLENE BLUE TESTING SINCE 1974

MicroD
RESOLUTION TARGETS

3000 TEXAS HIGHWAY 5 BURRIDGEVILLE, OHIO 44603 U.S.A.
TEL: (419) 432-7801 FAX: (419) 432-7801 TOLL FREE: 1-800-541-0000

National Library
of Canada

Bibliothèque nationale
du Canada

Canadian Theses Service

Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

NOTICE

The quality of this microform is heavily dependent upon the quality of the original thesis submitted for microfilming. Every effort has been made to ensure the highest quality of reproduction possible.

If pages are missing, contact the university which granted the degree.

Some pages may have indistinct print especially if the original pages were typed with a poor typewriter ribbon or if the university sent us an inferior photocopy.

Reproduction in full or in part of this microform is governed by the Canadian Copyright Act, R.S.C. 1970, c. C-30, and subsequent amendments.

AVIS

La qualité de cette microforme dépend grandement de la qualité de la thèse soumise au microfilmage. Nous avons tout fait pour assurer une qualité supérieure de reproduction.

S'il manque des pages, veuillez communiquer avec l'université qui a conféré le grade.

La qualité d'impression de certaines pages peut laisser à désirer, surtout si les pages originales ont été dactylographiées à l'aide d'un ruban usé ou si l'université nous a fait parvenir une photocopie de qualité inférieure.

La reproduction, même partielle, de cette microforme est soumise à la Loi canadienne sur le droit d'auteur, SRC 1970, c. C-30, et ses amendements subséquents.

Observer Accuracy in Usability Testing: The Effects of
Obviousness of Usability Problems, Prior Knowledge of Problems,
and Training

by

Adriane Martina Donkers, B.A.

A thesis submitted to
the Faculty of Graduate Studies and Research
in partial fulfilment of
the requirements for the degree of
Master of Arts

Department of Psychology

Carleton University

Ottawa, Ontario

April, 1992

© copyright

1992, Adriane Donkers

National Library
of Canada

Bibliothèque nationale
du Canada

Canadian Theses Service Service des thèses canadiennes

Ottawa, Canada
K1A 0N4

The author has granted an irrevocable non-exclusive licence allowing the National Library of Canada to reproduce, loan, distribute or sell copies of his/her thesis by any means and in any form or format, making this thesis available to interested persons.

The author retains ownership of the copyright in his/her thesis. Neither the thesis nor substantial extracts from it may be printed or otherwise reproduced without his/her permission.

L'auteur a accordé une licence irrévocable et non exclusive permettant à la Bibliothèque nationale du Canada de reproduire, prêter, distribuer ou vendre des copies de sa thèse de quelque manière et sous quelque forme que ce soit pour mettre des exemplaires de cette thèse à la disposition des personnes intéressées.

L'auteur conserve la propriété du droit d'auteur qui protège sa thèse. Ni la thèse ni des extraits substantiels de celle-ci ne doivent être imprimés ou autrement reproduits sans son autorisation.

ISBN 0-315-75906-2

Canada

The undersigned recommend to the Faculty of Graduate Studies
and Research acceptance of the thesis

Observer Accuracy in Usability Testing: The Effects of
Obviousness of Usability Problems, Prior Knowledge of Problems,
and Training

submitted by
Adriane Martina Donkers, B.A.
in partial fulfilment of the requirements for
the degree of Master of Arts

Thesis Supervisor

Chair, Department of Psychology

Carleton University
April, 1992

Abstract

This study examined performance of trainee observers in a usability test. Accuracy at identifying usability problems and performance at rating the importance of problems was assessed as a function of four independent variables: training, obviousness of usability problems, prior knowledge of potential problems, and trials. All subjects received procedural training. In addition, one-half of the subjects viewed a training videotape that provided instruction on the role of interface specialists in product design.

Profile analyses revealed that ability to identify problems improved with an increase in obviousness of the problems, and with prior knowledge of problems. Subjects rated low obvious problems of which they did not have prior knowledge as less important than low obvious problems of which they did have prior knowledge. Also, ratings of importance by subjects who viewed the training videotape varied from those of two experts less than the ratings of subjects who received only procedural training.

Acknowledgements

I wish to extend a very special thank you to Dr. Jo Tombaugh and Dr. Richard Dillon for their guidance, their support, and their patience throughout this project.

I also wish to thank members of the Computer User Research and Evaluation Laboratory for their assistance with the development of the test materials and their feedback on the experimental design.

To my family and friends, I thank you for encouraging me to continue, and offering me support and understanding during the entire process.

Table of Contents

	Page
Acceptance Sheet	ii
Abstract	iii
Acknowledgements	iv
Table of Contents	v
List of Tables	vii
List of Figures	viii
 Introduction	 1
Usability Testing	1
Purpose of Study	3
Factors that Affect Observer Performance	5
Observer Training	9
The Conditions and Hypotheses of the Current Study.....	13
 Method	 16
Subjects.....	16
Experimental Design	16
Apparatus and Materials	18
Experimental Procedure	26
 Results	 30
Dependent Variables.....	30
Problem Identification	32
Importance ratings	40
Qualitative Analyses on Other Problem Reports	48
Other Analyses	52
Interrater Agreement.....	53
Opinions of the Observational Task and Training.....	55
Opinions of the Training Videotape	58
 Discussion	 60
Design Summary	60
Effects of Obviousness and Prior Knowledge on Problem Identification	61
Effects of Obviousness and Prior Knowledge on Importance Ratings	64
Effects of Training	66
Effect of Practice.....	68
Interrater Agreement versus Accuracy	68
Limitations of the Current Study	69
Conclusions.....	70
 References	 72

Appendices.....	Page 75
Appendix A: Script of Videotape for Extended Training Group.....	76
Appendix B.1: Task Script.....	86
Appendix B.2: Steps for Completing Tasks	88
Appendix C: Criteria Used to Rate Importance of Problems.....	90
Appendix D: Problems Presented to Observers	91
Appendix E.1: Coding Form	93
Appendix E.2: User Summary Form	97
Appendix E.3: General Summary Form	102
Appendix E.4: Rating Summary Form	107
Appendix F: Microwave Manual.....	110
Appendix G.1: Task and Training questionnaire.....	114
Appendix G.2: Training Videotape questionnaire.....	119
Appendix G.3: Personal Data Questionnaire.....	121
Appendix H: Verbal Instructions to Observers	123
Appendix I: Informed Consent Form	131
Appendix J: Debriefing.....	132
Appendix K: Analyses on the Videotape Summaries.....	133
Appendix L: Rules for Scoring	140
Appendix M: Analysis on All Prior Knowledge Sets.....	142
Appendix N: Pearson correlation on importance ratings.....	145
Appendix O: Qualitative Analysis on Problems Identified.....	146
Appendix P: Analyses on Gender and Microwave Usage.....	154
Appendix Q: Analysis at Identifying the Number of Users Who Experienced Usability Problems	157

List of Tables

Table	Description	Page
Table 1.	Description of the levels of obviousness of problems	21
Table 2.	Problems observers were informed about prior to observing	23
Table 3.	Frequency and percent of importance ratings by experts for each set of problems of which subjects were informed	41
Table 4.	Frequency and percent of low, moderate, and high importance ratings by experts for each of the obviousness conditions	47
Table 5.	Description of invalid problems and frequency of invalid problems for each training group.....	49
Table 6.	Description of possibly valid problems and frequency of possibly valid reports for each training group.....	50
Table 7.	Description of valid problems and frequency of valid problems for each training group.....	51
Table 8.	2 x 2 table to represent agreements and disagreements amongst observers	53
Table 9.	List of opinions on the task and training	56
Table 10.	Frequency of subjects in each training group who thought components of the training could be omitted from the training program.....	58
Table 11.	Opinions of the training videotape.....	58

List of Figures

Figure	Description	Page
Figure 1.	Design of the microwave prototype and supporting diagrams.....	19
Figure 2.	Accuracy scores for problem identification by obviousness and prior knowledge	36
Figure 3.	Accuracy scores for problem identification by obviousness.....	37
Figure 4.	Accuracy scores for problem identification by prior knowledge.....	38
Figure 5.	Accuracy scores for problem identification by training group	39
Figure 6.	Accuracy scores for problem identification by trials	40
Figure 7.	Mean ratings of importance by obviousness and prior knowledge.....	43
Figure 8.	Mean ratings of importance by obviousness	44
Figure 9.	Variability of subjects' importance ratings from experts' importance ratings by obviousness.....	46
Figure 10.	Variability of subjects' importance ratings from experts' importance ratings by training group	48
Figure 11.	Opinions of the training program by training group	57
Figure 12.	Opinions of the training videotape.....	59

This study examined the accuracy and biases of observers reporting problems in a usability test as a function of the obviousness of usability problems, prior knowledge of potential problems, and observer training. As background for this study, the introduction describes the objectives of a usability test, the main purpose of the study, and a review of factors that may affect observer performance at identifying and interpreting usability problems. Previous relevant research is described.

Usability Testing

A usability test is an observational technique to evaluate a product under development to determine if it will meet the needs of users. The test determines if the product provides the functionality that users require, and if the product is easy to learn, and is easy to use. Products for which usability tests are performed may include computers, software packages, telephones, video-recorders (VCR), answering machines, or microwave ovens.

In a usability test, potential users of a product complete several tasks with the product or with a prototype of the product. For example, if an office telephone designed for receptionists is tested, then receptionists (the users) would perform a series of tasks such as make a telephone call, forward a call, transfer a call, store telephone numbers for quick-key access, and set up a conference call. An observer records performance of the receptionists as they attempt to complete each task. Performance measures might include time to complete tasks, correct and incorrect key presses, and ability to recover from errors. In addition, users' comments are recorded.

Subsequent to the usability test, the observer summarizes the data for all users, analyzes it, and proposes recommendations for modifications to the design of the product.

Usability tests are used to help identify and correct many problems with a product while it is being developed. Products that have extensive functionality and complexity are particularly important to test because these products, if designed incorrectly, can confuse users and lead to customer dissatisfaction. Consequently, usability testing in product design is gaining acceptance among companies that are concerned about user satisfaction. Also, researchers working in the area of product design have advocated its use because of the high value of user feedback for determining if a system is usable (Bergfeld-Mills, 1986; Gould & Lewis, 1983; Lund, 1985).

To successfully conduct usability tests, observers must have special knowledge and skills. Rosenbaum (1989) stated that usability testers require two kinds of knowledge. The first requirement is an in-depth understanding of experimental design and methodology, including data analysis. The second requirement is familiarity with relevant literature on human factors and usability. While there are many people who meet the first criteria, particularly those with training in psychology, there are few people who have combined their experimental training with human factors training. Because of the shortage of individuals with appropriate experience in both areas of expertise, psychologists with no human factors background are often hired to conduct usability evaluations.

Purpose of Study

In addition to knowledge of experimental design methodology and human factors issues identified by Rosenbaum (1989), the usability tester will require special observational skills. The observational skills focused on in this study are the ability to accurately identify usability problems, and the ability to rate the importance of fixing the problems.

Accuracy is a measure of agreement with a standard. It assesses the correctness of responding by examining how well observers record data in comparison to a previously established criterion. Accuracy is similar to reliability, but reliability refers to how well two or more observers record data in comparison to each other. Reliability is often measured when no criterion exists, and it is taken by most investigators to be an estimate of observer accuracy when a direct measure of accuracy is not available (Reid, 1970).

This study was designed to assess observer accuracy at identifying usability problems, and at assessing the importance of fixing the problems. Accuracy at identifying usability problems is the ability of observers, in comparison to a criterion, to record and summarize all problems that users have with a product. In a usability test, a measure of accuracy can be obtained if usability test experts establish a criterion by developing a "key" with the usability problems they identified. The number of agreements between the experts and observers on the occurrence of problems determines the accuracy of the observers. Accurate identification is important in usability testing because the problems identified help determine which recommendations will be made for redesign of the product.

While observers must accurately identify usability problems, they must also be capable of interpreting the observed problems and evaluate the importance of fixing each usability problem. Identifying importance of problems involves prioritizing the problems so that problems that are serious and may lead to user errors or dissatisfaction are given a higher priority for making design modifications than those that are less serious and may not lead to user errors or dissatisfaction. For example, if several users of an alarm clock have difficulty understanding how to set the alarm, then observers should identify this as an important problem since setting the alarm is a frequently used feature. However, if a button to set a second alarm is difficult to locate, then observers should give this problem a lower priority since a second alarm is a rarely used feature. Because observers make recommendations for product design, they must be able to recognize the critical problems that could seriously impact acceptance of the product.

Although accurate reporting is essential for improving product design, no reported research has examined observer performance at identifying usability problems or rating the importance of problems. The primary goal of this study was to provide such data on a number of novice usability test observers. Factors which may influence observer performance were considered.

In addition to assessing accuracy of observers, a second goal of this study was to examine biases when observers identify problems or when they assess the importance of problems. In general, a bias involves making decisions or judgments because of some inappropriate, but influential factor. In the context of a usability test, observers may identify problems because they have an expectancy that the problem might occur. Similarly,

an anticipation of problems may lead to an underestimation or an overestimation in rating the importance of fixing the problem.

A third goal of this study was to examine the effectiveness of a brief usability testing training program on observer performance. Programs designed to train usability testers are rare. Therefore, it is common for individuals who have no usability testing experience, and have no or little knowledge about user testing and human factors, to conduct such tests. To date, an evaluation of observer training in a usability test has not been reported in the literature.

Observer performance in the area of usability testing has not been addressed before. Consequently, to address the three goals of this study, and to describe the factors to be examined in the current study, research in other areas that is relevant to this study are described below.

Factors that Affect Observer Performance

Studies in clinical and developmental psychology investigated factors that affect accuracy and lead to biases in observations (Foster & Cone, 1986; Johnson & Bolstad, 1973; Kazdin, 1977; Repp, Nieminen, Olinger, & Brusca, 1988). Two of these factors may affect observer performance in a usability test. The two factors are characteristics of what is observed and expectancy of what may occur.

Characteristics of Observed Events

The characteristics of observed events include dimensions such as frequency, duration, latency, spatial separation, and intensity of individual events, and sequence of multiple events. Examining the effect of such dimensions on observer performance is of interest to researchers because it

is important to design an observational system that is sensitive to the values on the dimension of interest. For example, as explained by Foster and Cone (1986), if frequency of occurrence is the dimension of interest and an event can occur at frequencies ranging from zero to 400 times per minute, then the technique for detecting frequency must be sensitive to this range. The observational system will be less than adequate if it is sensitive within a range of only zero to 200 times per minute. Identifying the accuracy of the observational technique at a range of dimensions is important since it will have an impact on the confidence in the data and the conclusions that may be drawn. For example, if rarely occurring events are not reported, it is important to know if the events did not occur or if the events occurred but were not detected.

Although it is important to identify dimensions that may affect observer performance, few researchers have investigated the impact of characteristics of events on observer accuracy. A few studies have evaluated characteristics such as frequency of events, but the results of these studies were contradictory (Dorsey, Nelson, & Hayes, 1986; Jacob, Grounds, & Haley, 1982; House & House, 1979; Kapust & Nelson, 1984). The shortage of studies and the contradictory results on the effects of dimensions on observer performance indicate that there is a need for further research in this area.

Identifying the sensitivity of observers to the range of user problems that occur in usability testing is of concern in the current study. One dimension that may affect identification of problems is the obviousness of the problems, as determined by the users' behaviour. For example, when a user struggles for a long time to perform a task, the usability problem is more obvious than when the user makes a single error. Observers may

detect problems that are highly obvious, but miss problems that are not obvious. To examine the effect of obviousness on observer performance, the usability problems in this study were rated for their obviousness.

Expectations of Observations

If observers expect certain events to occur, this may influence what they record or what they conclude (Kazdin, 1977). The probability that usability test observers will have prior knowledge about potential problems with a product is high because decisions to test aspects of the product are usually based on anticipated problems. Expectations may have two effects on observer performance in a usability test. First, expectations may lead to improved accuracy at identifying problems that were anticipated. Second, it may lead to a biasing effect. A biasing effect involves influential factors, such as prior knowledge, inappropriately affecting the performance of the observers. For example, if observers report problems that are expected to occur, but in fact do not occur, the observations are biased by the expectancy.

A few clinical and developmental studies examined biasing effects on observer performance. The results indicate a tendency for biasing effects to occur more often when observers make subjective evaluations such as rating behaviour than if they record objective data (Foster & Cone, 1986; Harris & Lahey, 1982). For example, when observers expected to find emotional disturbances in a child, this was reflected when they rated a child's emotional disturbance (Shuller & McNamara, 1980). However, when observers recorded the rates of deviant behaviour of children (Skindrud, 1972), or when they recorded the frequency of events such as face-touching and eye contact in an interview (Harris & Ciminero, 1978)

there was no biasing effect. Further, one group of researchers examined both behavioural recordings and subjective evaluations and found that subjective evaluations of disruptive children were related to observers knowledge of expectations, but actual event recordings were unrelated to expectations (Kent, O'Leary, Diament, & Dietz, 1974). However, another researcher found biasing effects even when observers recorded events and when they evaluated the social adjustment of children (Michelson, Mannarino, Marchione, Kazdin, & Costello, 1985). Previous research, therefore, indicates that biasing effects may occur when observers are recording events or when they are making evaluations. Although the research is contradictory, it does tend to show biases more often when observers evaluate behaviour than when they record the occurrences of behaviours.

In the current study, the effect of expectancies on observer performance was examined. One-half of the observers were told to watch for some of the problems that were known to occur in the videotapes from which they collected data. The potential problems were presented to observers prior to conducting the observations. The other one-half of observers were informed of a different set of potential problems. It was predicted that observers would be more attentive to problems that are anticipated, and so detect them more easily than problems that were not identified in advance.

Although performance was expected to improve with prior knowledge, biasing effects were also anticipated. Observers were instructed to watch for some problems that did not occur in any of the videotapes. The frequency at which subjects reported these non-existent problems was examined. Biasing effects were also expected when subjects

rated the importance of problems. Performance at rating the importance of problems was examined to determine if observers underestimated or overestimated the importance of problems based on prior knowledge of some problems. Because previous research has tended to show more biasing effects when making subjective judgments than when recording behaviour, more biasing effects were anticipated when observers rated the importance of problems than when they identified the occurrence of usability problems.

Summary of Factors

Two factors, problem obviousness and prior knowledge of problems, were expected to affect observer performance in a usability test. This study examined the effect of these factors on observer performance as a function of observer training.

Observer Training

There is ample evidence in the observational literature that errors in data recording are inevitable when using observers. To deal with this problem, researchers developed a variety of training programs for teaching observers how to observe and record data. It has generally been shown that observers who receive training perform considerably better than observers in control groups who do not receive training (Spool, 1978). This positive effect of training on levels of accuracy led researchers to assume that training is a necessary component for collecting observational data. Researchers, however, have not paid a great deal of attention to the technology of training (Foster & Cone, 1986).

Since training is generally effective in other observational research, it is reasonable to assume that training will also be effective for observers of a usability test. However, a training approach that is effective in one area of research may not be effective in a different area. Previous studies involved psychiatric, medical, or employment interviewing while others involved making ratings of students and teachers, recording verbal interactions or motor behaviour, and assessing employer behaviour. These studies differ from usability testing in the events that are observed, who is observed, and the use of the data. The current study, therefore, examined the effectiveness and appropriateness of a training program designed for usability testing.

The Training Program

In examining the role of observers in usability testing and the setting in which observations will be made, it is apparent that there are certain requirements for the observer training program. Since observers will be employed by companies who want to collect data as early as possible and want observers to take little time or money to learn their task, a first requirement is that the training program be efficient. This implies that observers quickly and independently become familiar with the observational technique. Independent training simply means learning how to observe primarily through self-training, employing little time of a trainer. A training program designed to be efficient has been referred to as minimalist training (Carroll, 1984). Some of the principles of this training may be applied to observer training.

Carroll (1984, 1987) examined minimalist training for documentation that accompanies software packages. As outlined by Carroll, minimalist

documentation is task-oriented so users can work with the product immediately. While working with the product, they can learn the necessary procedural and declarative knowledge required for understanding a task domain, such as text editing. There is no rote learning involved and the learner is encouraged to practice. Also, there is the elimination of verbiage such as found in overviews, repetitions, reviews, and summaries. Therefore, procedures, rather than concepts, are emphasized. The documentation does not overwhelm the users and it answers users' questions as simply as possible.

In applying the minimalist training principles to observer training, observers would be provided with a task-oriented approach to learning such that instructions are given in a succinct fashion and the observers have an opportunity to conduct some observations. There is no rote-oriented approach to learning.

Although procedural knowledge is necessary for training observers, this training may not be sufficient for observers in a usability test. Usability testing requires more than simply observing and recording what users do. It also involves complex post test analyses such as prioritizing problems when making recommendations for design modifications. Skill at prioritizing may require more than just procedural knowledge. It may also require an understanding of the context of usability testing, including knowing the role of usability testing in product design, knowing the user testing process, and understanding the implications of the observations made. This information may be especially important for observers who have no experience in the area of usability testing or human factors.

The current study examined the effect of a training program that instructed observers on the context of usability testing. This instruction,

presented to one-half of the subjects, may be referred to as preinstructional training. In the training, subjects learned about the role of user interface specialists in product design, and they were introduced to the use of evaluation techniques such as usability testing for assessing product usability. The implications and the effect of usability testing on product design were emphasized. Performance of subjects who received preinstructional training in addition to procedural training was compared to performance of subjects who received only procedural training.

Research on Preinstructional Training

There are many forms of preinstructional training. Some of the forms include pretests, behavioural objectives, overviews, and advance organizers (Hartley & Davies, 1976). The preinstructional training provided to one-half of the subjects in this study is most closely related to overviews and advance organizers.

In general, the purpose of preinstructional training, including overviews and advance organizers, is to provide a "bird's-eye-view" of what lies ahead for the learner (Hartley & Davies, 1976). Overviews are generally thought of as a means to introduce learners to new material by familiarizing them with the central argument. They emphasize key concepts, principles, and technical terms. Advance organizers have a similar purpose, but the organizer is usually described as being at a higher level of abstraction than an overview. The main purpose of the organizer is to provide a framework that learners can use to clarify the task ahead, and to provide a context for a general understanding and comprehension of the upcoming material.

The preinstructional training provided to subjects in this study combines elements consistent with both an overview and an advance organizer. These forms of training have generally been helpful, or they have shown small facilitative effects in a number of learning situations (Hartley & Davies, 1976; Luiten, Ames, & Ackerman, 1980; Mayer, 1979; Stone, 1983). For example, advance organizers have positively affected comprehension of a second language (Tudor, 1986), performance at solving mathematical problems (Doyle, 1986); and problem solving (Bayman & Mayer, 1988; Mayer, Dyck, & Cook, 1984). Some researchers have reported that the effects are greater for lower ability subjects than for higher ability subjects (Bayman & Mayer, 1988; Tudor, 1986). This study examined whether or not the use of preinstructional training which has shown to be effective in other research areas, is effective in usability testing, or if a minimalist approach to training that emphasizes procedures is adequate.

Summary of Training

Previous research has shown preinstructional training to be of value in many areas of learning. This type of training may also be useful for trainee observers of usability tests, particularly if the observers have no experience with usability testing.

The Conditions and Hypotheses of the Current Study

To determine the effectiveness of a training program on observer performance, two groups were compared. Both groups received some basic training that involved learning about the purpose of usability testing and learning how to observe. One of the groups also viewed a training videotape that described the role of user interface specialists in product design and

the use of usability testing for evaluating product usability. This group is referred to as the extended training group. The other experimental group which did not view the training videotape is referred to as the basic training group.

After training, each observer viewed three videotapes, each with a user performing tasks on a microwave oven. Observers identified usability problems and rated the importance of fixing the problems. The actions of the users were scripted to include a range of usability problems that varied on the dimension of obviousness. Two usability test experts rated all problems for their level of obviousness. Problems were rated as either low, medium or high in obviousness.

As a further manipulation, all of the observers were informed, prior to the observational sessions, of some of the usability problems that occurred in the videotapes. One-half of the observers in each experimental training group were informed of one set of problems and the other one-half of observers in each training group received a different set of potential problems. Observers were told that designers were concerned about the problems, and so they should look for the problems and record them if they occur.

The study was designed to evaluate observer performance at identifying problems and at rating the importance of problems as a function of the three independent variables (training, obviousness, and prior knowledge). Observer performance was compared to the performance of two experts in usability testing.

The following hypotheses were formulated about the performance of observers at identifying usability problems. First, it was hypothesized that as obviousness of the problems increased, accuracy would also increase.

Second, observers would record data more accurately for the problems of which they had prior knowledge than for those of which they had no prior knowledge. Third, subjects in the extended training group that received instruction on the context of usability testing would identify more of the usability problems than the subjects who received only procedural training (basic training group), particularly for low obvious problems and problems of which observers had no prior knowledge. Fourth, performance of observers would improve with practice.

The following hypotheses were formulated about the performance of observers at rating importance. First, it was expected that observers would rate problems of which they had prior knowledge as more important than problems of which they had no prior knowledge, particularly for low obvious problems. Second, ratings of importance by subjects in the extended training group would be less variable from the experts' ratings of importance than would ratings by subjects in the basic training group.

Method

Subjects

Ten male and 14 female psychology students participated in the study as observers. Eight of the subjects were undergraduates in their fourth year and 16 were graduate students. The average age of the subjects was 26.1 and ranged from 21 to 38. Each subject received \$50 for their participation. Subjects were run individually in two three-hour sessions.

Experimental Design

To examine the effects of several variables on observer performance, a mixed design with one between-subjects factor and three within-subjects factors was used. The between subjects-factor was the form of training that observers received. Twelve of the 24 subjects viewed a training videotape (extended training group) and the other 12 did not see the videotape (basic training group). The videotape taught observers about the role of user interface specialists in product design and techniques they use to improve product designs. The videotape focused on the usability testing approach to evaluate products.

The within-subjects factors included trials, prior knowledge of expected user problems, and obviousness of usability problems. Subjects completed three trials, each consisting of observing a person perform usability tasks on a prototype of a microwave oven. Subjects observed the individuals using the microwave oven, and recorded the users' performance. The order of the three videotapes was completely counterbalanced with six orders of presentation (A B C, A C B, B A C, B C

A, C A B, C B A). Observer performance across the three videotapes was examined to see if there were improvements with practice.

Knowledge of expected problems is the second within-subjects factor. All subjects were informed of some usability problems. One-half of the subjects in the basic training group and one-half in the extended training group received one set of problems. The other subjects in each training group received a separate set of usability problems. All subjects were told that the problems of which they were informed were potential problems that are of special concern to the designers, and they should look for the problems and record them if they occur. They were also instructed that there would be problems that the designers did not anticipate, and so they should look for other problems that occur. Informing subjects of potential problems was used to examine the effect of prior knowledge on accuracy and to examine biasing effects.

The third within-subjects variable was obviousness of each of the problems encountered by users. Two usability test experts viewed each of the videotapes, recorded the usability problems, and rated each problem for its level of obviousness. Each problem was rated as either low, medium, or high obvious. Two factors determined the ratings for level of obviousness: 1) the likelihood of seeing the problem occur, and 2) the likelihood of recognizing it as a usability problem and recording it. Observer performance was examined at the three levels of obviousness to determine if obviousness of problems affects accuracy and ratings of importance of problems.

Apparatus and Materials

Training Videotape

One-half of the observers (the extended training group) viewed a 20-minute training videotape. A script of the videotape is shown in Appendix A. The videotape described the role of user interface specialists in product design, and described techniques the specialists use for assessing user needs, including task analysis, usability testing, and field trials. The videotape emphasized the rationale and general procedure for conducting a usability test. Observers saw enactments of various phases of designing a product, and conducting a usability test. For example, some of the acts included a user performing tasks, an observer watching a user and summarizing the data, and a specialist negotiating with the designers.

Producing the training videotape

To produce the training videotape, an Hitachi VHS colour video camera was used. Various clips of a narrator or enactments were videotaped and edited to produce a single videotape approximately 20 minutes long. When the narrator was speaking, either the narrator or a chart or list was shown. For the enactments, various actors acted out scenes that typically occur with user interface specialists, usability testers and designers.

User Test Videotapes

Observers viewed three videotapes, each with a user performing tasks on a prototype of a microwave oven. The design of the microwave prototype is shown on the right side of Figure 1. On the left side of the figure are pictures of a miniature microwave oven and a probe

(thermometer) beside a bowl. These diagrams supported the context of operations involving a microwave oven, that is, opening and closing the microwave door and inserting a temperature probe into some food.

Figure 1. Design of the microwave prototype and supporting diagrams.

To create the videotapes, three individuals acted as users of the microwave oven. The users followed a script while performing various tasks with the microwave oven. Each user performed a total of 11 tasks (see Appendix B.1). The actors, portraying as users, talked out loud to voice any confusions they were experiencing, and the reasons for their behaviours. The confusions and difficulties experienced by a user were scripted to have a variety of problems occur in the videotapes. The majority of the scripted problems were based on errors made by three individuals who tested the prototype. Each videotape was approximately 30 minutes long.

Producing the user test videotapes

To produce the user test videotapes, a mouse-driven Macintosh IIX computer, a VHS video cassette recorder (VCR), and a television monitor were used. The microwave prototype was designed in Hypercard (version 1.2.5) and run on the computer. The computer, equipped with an eight-bit video card, Quick-Draw software, and Video Card Utility software, was connected to the VCR via an NTSC video cable. The VCR was connected to the television monitor. With this configuration, the television monitor replaced the computer monitor. While the user performed tasks on the microwave prototype shown on the television monitor, actions performed by the user on the prototype were recorded on videotape. Any mouse clicks to select buttons on the microwave, and any confusions voiced by the user were audiotaped. The user was never seen on videotape.

Problem characteristics

Obviousness of usability problems

Two usability test experts observed the videotapes, identified the usability problems, and rated each problem in each videotape for its level of obviousness. The operational definition of obviousness was the ease of detecting the problem and ease of identifying the problem as a usability problem. Obviousness is dependent on the users' behaviour. For example, if a user quickly makes an error and recovers, then the problem would not be obvious. However, if the error is repeated for a long duration or the user comments on the error then the problem would be obvious. Each problem in the videotapes was rated as one of three levels of obviousness (low, medium, or high). The definitions of each of these levels of obviousness are shown in Table 1.

Table 1

Description of the levels of obviousness of problems.

Level of obviousness	Description
Low obvious	Problems are difficult to detect and observers are unlikely to recognize and record them as usability problems.
Medium obvious	Problems may or may not be detected and observers might recognize and record them as usability problems.
High obvious	Problems are easy to detect and observers are likely to recognize and record them as usability problems.

The usability test experts summarized the problems identified in the three individual videotapes into a single general summary so that similar problems experienced by two or more users were described as a single problem. To determine the level of obviousness for a single problem that occurred in two or more of the videotapes, the highest level of obviousness given to one of these problems was used. For example, if two users had the same problem, but the portrayal for one user was of high obviousness and for the other of low obviousness, then the high obvious rating determined the obviousness of the problem.

Importance of problems

The usability problems varied in importance, where importance refers to how necessary it is to fix a problem. The two usability test experts who rated the obviousness of each problem also rated the importance of each problem. Ratings were made on a five-point scale ranging from not important to fix (1) to very important to fix (5) with the mid-point labelled as moderately important to fix. The criteria used by experts for making the ratings are presented in Appendix C.

Prior Knowledge of Problems

Prior to conducting the observations, one-half of the observers in the extended training group, and one-half in the basic training group were informed of some problems that, they were told, might occur in the videotapes. The other subjects, one-half from each training group, were informed of a different set of problems. In each set there were three low obvious, three medium obvious, three high obvious, and three non-existent problems. The non-existent problems did not occur in any of the videotapes. The non-existent problems were included to determine if subjects reported the problems because they were instructed that they might occur.

The low, medium, and high obvious problems were selected so that the number of users who experienced a problem were matched in the two sets. For example, if one set contained a medium obvious problem encountered by two users, then the other set also contained a medium obvious problem encountered by two users. Subjects received the set of problems on a piece of paper (see Appendix D). The problems were pointed out to subjects when they were taught how to use the prototype microwave oven. Instructions emphasized that the designers were concerned about the potential problems in the set. Subjects were told to look for the problems and record them if they occur. They were also told that there will inevitably be other problems that they should record. The problems presented to subjects, arranged in order by level of obviousness, are shown in Table 2.

Table 2

Problems observers were informed about prior to observing.

Set A Problems given to 1/2 of the subjects in the extended training group and 1/2 of the subjects in the basic training group	Set B Problems given to 1/2 of the subjects in the extended training group and 1/2 of the subjects in the basic training group
LOW OBVIOUS: <ol style="list-style-type: none"> 1) Do not know what the labels on the display mean (e.g., 1stStg).--- (1 user) 2) Do not like it that there is a difference between the way temperature is entered and way time is entered.--- (1 user) 3) Do not put the probe in when using temperature.--- (1 user) 	LOW OBVIOUS: <ol style="list-style-type: none"> 1) Try to set the POWER before setting cooking time or temperature.--- (1 user) 2) When using DEFROST, do not like it that entering a single number (e.g., 1) indicates only a fraction of a pound.--- (1 user) 3) Expect that have to enter 2 digits when setting power level (e.g., 4 & 0 for 40%).--- (1 user)
MEDIUM OBVIOUS: <ol style="list-style-type: none"> 1) Do not know how to enter time for the DELAY feature. May put 100 for one minute from current time.--- (1 user) 2) Want the oven to stop between stages when doing multi-stage cooking.--- (1 user) 3) For QUICK-ON feature, try to put in 00's after entering 1.--- (2 users) 	MEDIUM OBVIOUS: <ol style="list-style-type: none"> 1) Want to see the power level during cooking.--- (1 user) 2) Do not know to use the MEMORY button for multi-stage cooking.--- (1 user) 3) Confused by the word COOL when using temperature.--- (2 users)
HIGH OBVIOUS: <ol style="list-style-type: none"> 1) Use the QUICK-ON button for speed instead of using POWER.--- (1 user) 2) When using DEFROST, think that time should be entered as well as weight.--- (2 users) 3) Do not know how to set power to a different level than HI%.--- (3 users) 	HIGH OBVIOUS: <ol style="list-style-type: none"> 1) Try to use MEMORY instead of HOLD WARM to set warming temperature.--- (1 user) 2) Do not know where to use the MEMORY button when doing multi-stage cooking.--- (3 users) 3) Do not know to touch the DELAY button after entering the delay time.--- (3 users)
NON-EXISTENT: <ol style="list-style-type: none"> 1) Expect to touch START after entering the time for QUICK-ON. 2) Think the clock time has to be removed before using another feature. 3) Try to set the HOLD WARM temperature before setting POWER level. 	NON-EXISTENT: <ol style="list-style-type: none"> 1) Expect to touch START after entering the time for QUICK-ON. 2) Think the clock time has to be removed before using another feature. 3) Try to set the HOLD WARM temperature before setting POWER level.

Forms

To identify problems and report the importance of fixing problems, observers completed a number of forms, including coding forms, user summary forms, a general summary form, and a rating summary form. Oral and written instructions for completing the forms were given and subjects were shown sample forms with hypothetical data from a telephone usability test. The forms are described below and the complete set of forms with sample forms is shown in Appendices E.1 to E.4.

Coding form

The observer completed the coding form while viewing the user test videotape. In the form, the subject recorded the users' actions, comments, and/or usability problems encountered. Subjects completed separate forms for the 11 tasks performed in each of the three videotapes.

User summary form

Subjects completed a user summary form after observing a user-test videotape. There was a separate form for each of the 11 tasks conducted by a user. In the form, the observer summarized, for each task, the usability problems encountered by a user. This form was used to assess observer ability at identifying problems within a videotape.

General summary form

After viewing all three videotapes, the observer summarized the problems on a per task basis for all three videotapes. In each summary, observers recorded the problem encountered by users and the number of the three users who experienced the problem.

After summarizing the problems for all tasks, observers rated the

importance of fixing each of the problems. Ratings of importance were made on a scale ranging from not important to fix (1) to very important to fix (5), with the mid-point (3) labelled as moderately important to fix. This scale is the same scale the experts used when they rated the importance of fixing each problem. Also, the same criteria for making the ratings was used (see Appendix C). The general summary form was used to assess the accuracy of observers' reports of problems and ability to rate the importance of fixing each of the problems.

Rating summary form

All of the problems identified by the two usability test experts were listed in the rating summary form. This form included all of the problems that subjects should have reported in the general summary form. Using the same criteria and 5-point rating scale as in the general summary, subjects rated the importance of fixing each of the problems listed in the form. This form was used to assess subjects' perceptions of the importance of all problems, including those they identified and those they missed.

Microwave Oven Manual

Subjects had a microwave oven manual that they could consult at any time. The manual described the functions of the controls, and how to use them. The manual is shown in Appendix F.

Questionnaires

Observers completed several short questionnaires at the end of the experiment. The questionnaires, described below, are shown in Appendix G.1 to G.3.

Opinion of the task and the training questionnaire

Observers completed a questionnaire designed to assess their opinions on the difficulty of the task and the quality and usefulness of the training. There were two versions of the questionnaire, one for each of the training groups. The differences between the questionnaires were very small.

Opinion of the training videotape questionnaire

Observers who received the training videotape completed a questionnaire on the quality and usefulness of the videotape.

Personal data questionnaire

Subjects completed a personal data questionnaire to collect demographic information, assess subjects' use of microwave ovens, their involvement in product design, and their experience with formal observation.

Experimental Procedure

Setting

Subjects were seated at a table in an experimental room equipped with a Macintosh II computer, a television monitor and a VCR for viewing the videotapes. The video equipment was placed approximately four feet in front of the subject.

The experimenter was not present in the room while observers watched the videotapes or completed the forms. The experimenter entered the room to change videotapes, to provide instructions, and to provide

materials to observers. Instructions given to subjects are presented in Appendix H. The observer was permitted to ask questions at any time throughout the experiment.

Introduction to the Study

At the beginning of the experiment, subjects completed the informed consent form (see Appendix I). The consent form explained that the purpose of the study was to assess performance at recording observational data.

Training

Initial training involved a brief explanation of what usability testing is and its purpose. Subjects in the extended training group then viewed the training videotape on the role of user interface specialists, and the use of techniques for evaluating product designs.

All observers received training on their observational task. This training involved informing subjects of all the forms they would be completing (i.e., coding forms, user summary forms, and general summary form) and instructing them how to complete each of the forms. Subjects were also shown the sample forms at this time. In addition, various types of usability problems were described (e.g., performing steps in the incorrect order, using incorrect buttons, omitting steps, and making negative comments).

Subjects were instructed that they could record the data in two different formats or use a combination of the two approaches. One approach is to record every button press or comment the user makes, and then subsequently identify the usability problems when summarizing the data. The second approach is to identify and summarize problems as they

occur.

For the user summary forms and the general summary, subjects were given a rationale for completing the forms, i.e., putting the data into a useful and meaningful package and to look for patterns in the data. The method for completing each of the forms was then explained and subjects reviewed the sample forms.

Both groups of subjects then spent about 20 minutes learning how to use the microwave oven prototype which ran in HyperCard on the Macintosh computer. Observers were told that it is necessary to learn the microwave oven so they can identify problems, but that they did not have to be experts on the oven since they would have aids with them while recording. To learn the microwave oven, the experimenter read a task in the task script and then reviewed the correct steps for completing the task (see Appendix B.1). The subject then performed the task on the prototype. Any difficulties encountered by the subjects or any comments they made about the design of the microwave were recorded by the experimenter. Further, to manipulate prior knowledge of problems, subjects were presented with 12 potential problems (see Appendix D). Each problem was presented one at a time appropriate to the task being learned on the microwave oven. For example, a problem with setting power level was described when the subject learned about the power function.

Observation Sessions

Prior to observing the three videotapes, subjects were instructed to be as accurate as possible and to get what information they needed to identify the problems, no matter how minor the problems might seem. Subjects were not permitted to rewind the videotape or to stop it during a task, but

they were allowed to stop the videotape between tasks to complete their recording for the previous task and to review the steps for the next task. Stopping the videotape between tasks simulates a real usability test where the observer can ask the user to wait while writing notes or preparing for the next task. Rewinding and stopping the videotape during a task was not permitted because, if a usability test is live, i.e., not videotaped, then users could not be asked to redo their steps.

After viewing the first videotape, the experimenter reviewed the instructions for completing the user summary form. Subjects then completed this form for the first videotape watched. They also completed user summary forms after viewing the second and third videotapes.

After viewing all three videotapes, observers summarized, on the general summary form, the problems that occurred for all three users. Again, the experimenter explained the procedure for completing the form. In the form, subjects recorded each problem that occurred, the number of users who experienced each problem, and they rated the importance of fixing each problem.

Finally, observers completed the rating summary form that included all problems identified by the experts. In this form, subjects rated the importance of fixing each of the problems in the videotapes, including those they did not identify.

At the end of the experiment, subjects completed the questionnaires on the task and training, and the personal data questionnaire. They were then debriefed on the purpose of the study (see Appendix J).

Results

Observer performance at identifying usability problems, and at rating the importance of problems were assessed. Prior to the experiment, two usability testing experts watched the usability test videotapes to report the usability problems that they detected, and to rate the importance of fixing each problem. The experts completed two types of summary forms, a videotape summary form for each of the three videotapes observed, and a general summary which is a composite form based on all problems recorded in the videotape summaries. In the videotape summaries, the experts identified 28 problems in the first videotape, 26 in the second, and 25 in the third for a total of 79 problems. A single problem sometimes occurred in more than one of the videotapes. These 79 problems were therefore collapsed into a total of 54 distinct problems in the general summary.

Subjects also watched the videotapes and completed the same summary forms as the experts. In the summaries, subjects recorded the problems they identified in the videotapes, and they rated the importance of fixing each problem. Results obtained from the experts provided a scoring standard for correct problem identification, and for importance ratings.

Dependent Variables

Three main types of dependent variables were derived from the summaries. The first dependent variable was accuracy at identifying problems. Accuracy scores were computed separately for the videotape summaries (79 problems) and the general summary (54 problems). The

second dependent variable was mean ratings of importance of fixing each of the problems in the general summary. This variable was analyzed to examine biases in ratings of importance. The third dependent variable measured how much observers' ratings of importance varied from the experts' ratings of importance.

Scoring the Data for the Experimental Conditions

There were two within-subjects variables, obviousness and prior knowledge of the problems. The two usability test experts rated each problem as low, medium or high obvious. Of the 79 problems in the videotape summaries, 18 were in the low obviousness category, 21 in the medium obviousness category, and 40 in the high obviousness category. In the general summary with 54 problems, the number of low, medium, and high obvious problems was 13, 14, and 27, respectively.

To manipulate prior knowledge of problems, subjects were informed of a set of problems that occurred in at least one of the videotapes. There were two sets of problems, one set was presented to one-half of the subjects, and the other set was presented to the other one-half of subjects. In addition, there were problems that did occur in the videotapes but were not included in either of the balanced sets. Each problem in the videotapes, therefore, belonged to one of three prior knowledge sets. The two balanced sets of problems are referred to as problems of which subjects were informed (informed condition), and problems of which subjects were not informed, but other subjects were informed (not informed condition). The remaining problems are referred to as the problems of which no one was informed (no one informed condition).

The dependent variables were summarized for each subject by averaging scores in a 3 x 3 table representing the combination of the obviousness and prior knowledge independent variables. For example, an accuracy score for identifying problems was derived for low obvious problems of which a subject had prior knowledge, low obvious problems of which other subjects had prior knowledge, and low obvious problems of which no one had prior knowledge. The same approach was used for medium and high obvious problems. Described below are the methods used to obtain scores for each of the three dependent variables, and the results of statistical tests conducted on these variables.

Problem Identification

This section presents the results on accuracy at identifying problems. Separate analyses were conducted for the videotape summaries and the general summary to determine if the factors affected problem identification similarly on the two types of summaries. Since the general summary involves the skill of summarizing problems across the three videotapes, the results in the two types of summary forms could have been different. The results, however, were very similar, and so the general summary results are described below with those for the videotape summaries only mentioned where different results exist between the two summaries. The complete results for the videotape summaries are described in Appendix K.

Calculation of Accuracy at Identifying Problems

In the scoring scheme for problem identification, problems identified by the subjects were compared to those reported by the experts. A subject received a score of 1 for reporting a problem that was also reported by the

experts. A subject who failed to report a problem (a miss), received a score of 0. Problems identified by subjects that experts did not identify were recorded and analyzed separately. Specific rules for scoring are presented in Appendix L.

The experimenter, who was unaware of the training condition of each subject, scored the data. To determine reliability of scoring, a second rater separately coded the data for six randomly chosen subjects.¹ The reliability check between the experimenter and the second rater revealed a Cohen Kappa coefficient of .85 ($Z = 25.0, p < .001$), and a total agreement percentage of 93 percent. This coefficient indicates that the two raters were highly consistent at scoring the data.

Based on the coding, accuracy scores at identifying problems were obtained for each subject by computing the percent agreement between the subjects and the experts.² An accuracy score (percent agreement) was computed for each subject for each level of obviousness by each level of the prior knowledge conditions. The percent agreement formula is as follows:

$$\% \text{ agreement} = \frac{\text{number of agreements}}{\text{number of agreements} + \text{number of disagreements}} \times 100$$

¹ The rater was one of the usability test experts who previously viewed the videotapes and identified usability problems. It was important to have two individuals who were familiar with the usability problems in the videotapes, and who could interpret subjects' descriptions of the problems.

² A reliability measure such as the Pearson product-moment correlation was not calculated because a high reliability estimate could be obtained even if the observers and experts did not agree. For example, for low obvious problems the reliability could be high if the observers' scores are predictable from the experts' scores because of a lot of disagreement.

In the denominator, the sum of the number of agreements and number of disagreements is equivalent to the total number of problems that the experts identified for any combination of the obviousness and prior knowledge conditions. If a subject identified 10 problems in a condition that had 20 problems then the subject received an accuracy score of 50 for correctly identifying 50% of the problems.

Preliminary Analysis

Before conducting the main analyses, an analysis was conducted to ensure that the two prior knowledge sets were "balanced" so that performance at identifying problems was equivalent for the two sets. A profile analysis appropriate for split-plot designs was conducted. The repeated measures independent variables included prior knowledge and obviousness, and the between-subjects independent variable was prior knowledge set. The analysis revealed no significant effects on accuracy at identifying problems due to the set of problems about which subjects were informed ($F < 1$). Furthermore, set did not interact with obviousness ($F(2,44) = 1.59, p = .22$) or with prior knowledge ($F < 1$). These results indicate that the problems within one set were not easier to identify than the problems in the other set.

Primary Analyses

The main analyses involved profile analyses appropriate for split-plot designs. The repeated measures factors included prior knowledge and obviousness and the between-subjects factor was training. These analyses examined the effects of obviousness, prior knowledge of problems, and training on accuracy at identifying problems. Whenever the test involved the effect of prior knowledge, only the 18 problems in the balanced sets were

used. Other analyses, such as main effects for obviousness and training, included all 54 problems in the general summary.

An additional analysis was performed with obviousness and trials as the repeated measures, and training as the between-subjects factor. The trials independent variable assessed performance at identifying problems in each of the three videotapes. The analysis examined the main effects of trials and interaction effects involving trials.

Interaction effects on problem identification

Analyses were conducted to examine the predicted interactions between obviousness, prior knowledge, and training on accuracy at identifying problems. Since it was predicted that prior knowledge would affect performance, the initial analysis involved only the two balanced sets of problems of which subjects had prior knowledge. By including only these sets of problems it was possible to determine whether or not prior knowledge of a set of problems lead to improvements in performance over a similar set of problems of which subjects did not have prior knowledge.

The analysis revealed no 3-way interaction between obviousness, prior knowledge, and training ($F < 1$). Further, the 2-way interactions were non-significant between obviousness and prior knowledge ($F(2,44) = 1.49, p = .24$), between obviousness and training ($F < 1$), and between prior knowledge and training ($F(1,22) = 1.71, p = .21$).

Although the interaction between prior knowledge and obviousness was not significant, Figure 2 shows that there was a tendency for the observers to identify more of the problems of which they had prior knowledge than those of which they did not have prior knowledge for the low and medium obvious problems. A difference in the scores by prior

knowledge did not occur for the high obvious problems. Because this pattern was predicted, simple-effect tests were conducted to examine the differences in accuracy at identifying problems on the low and medium obvious problems. The differences were non-significant for both the low obvious problems ($F(1,22) = 1.64, p = .21$) and the medium obvious problems ($F(1,22) = 3.01, p = .10$).

Figure 2. Accuracy scores for problem identification by obviousness and prior knowledge.

Effect of obviousness on problem identification

A main effect analysis was conducted to test the hypothesis that accuracy would increase as obviousness increased. The analysis included all 54 problems in the general summary. The predictions were confirmed. There was a significant linear effect of obviousness on problem identification ($F(1,44) = 135.80, p < .001, \eta^2 = .75$)³. As shown in Figure 3,

³ η^2 is a measure of the proportion of variance in accuracy accounted for by the independent variable.

⁴ The effect was also significant when the analysis included only the two balanced sets of problems of which subjects were informed ($F(1,44) = 79.72, p < .001, \eta^2 = .63$).

performance was poorest for the low obvious problems, and improved as the obviousness of the problems increased. The mean accuracy was 44% for low obvious problems, 66% for medium obvious problems, and 88% for high obvious problems.

Figure 3. Accuracy scores for problem identification by obviousness.

Effect of prior knowledge on problem identification

A profile analysis involving the two balanced sets of prior knowledge problems was conducted. Performance on problems of which subjects had prior knowledge was compared to performance on problems of which subjects did not have prior knowledge but other subjects did have prior knowledge. There was a marginally significant effect of prior knowledge on problem identification ($F(1,22) = 3.51, p = .07, \eta^2 = .13$)⁵. This effect was significant for the videotape summaries ($F(1,22) = 4.13, p = .05, \eta^2 = .15$) (see Appendix K). As predicted, the means in Figure 4 indicate that subjects performed better on problems of which they had prior knowledge

⁵ The η^2 is fairly high. Although the probability does not exceed a traditional alpha level of .05, because of the low number of degrees of freedom, the results are interpreted.

(mean accuracy = 75%) than on problems of which they did not have prior knowledge (mean accuracy = 69%).

Figure 4. Accuracy scores for problem identification by prior knowledge.

An analysis was also conducted on the prior knowledge variable to examine performance at identifying problems in the set of problems of which no one was informed. The results of this analysis are presented in Appendix M.

Effect of training on problem identification

An analysis was conducted to examine the impact of the two training programs on problem identification. A significant main effect was not found ($F < 1$). As shown in Figure 5, subjects in the extended training condition who saw the videotape did not identify problems better than subjects in the basic training condition. The mean accuracy was 64% for the extended training group and 68% for the basic training group.

Figure 5. Accuracy scores for problem identification by training group.

Effect of trials on problem identification

To examine the effect of trials, subjects' accuracy at identifying problems was calculated for each of the videotapes (three trials) and for each level of the obviousness conditions. A profile analysis with obviousness and trials as repeated measures and training as a between-subjects factor was conducted. There was no significant trials main effect ($F(2,44) = 1.32, p = .28$), no interaction of trials with obviousness ($F < 1$), and no interaction of trials with training ($F(2,44) = 1.32, p = .34$). As shown in Figure 6, performance at identifying problems did not improve over the three videotapes observed. The mean accuracy for the first two videotapes was 66%, and for the third videotape was 70%.

Figure 6. Accuracy scores for problem identification by trials.

Importance ratings

The previous analyses examined subjects' accuracy at identifying problems. The following analyses examined their ability to rate the importance of fixing the usability problems. The analyses on importance ratings examined biases in rating importance and variability of observers' ratings from the experts' ratings.

After summarizing all problems in the general summary, subjects rated the importance of each problem. Each subject did this task twice. The first rating was for only the problems they identified and summarized in their general summary. The second rating was for all 54 problems that the experts identified in the general summary. For this second rating, all problems identified by the experts were presented to subjects in a list which the subjects then rated.

Mean importance ratings, and reliability coefficients were calculated to determine if subjects gave similar ratings of importance to each of the problems on the two separate rating sets. The results, presented in

Appendix N, revealed that subjects did give similar ratings. Consequently, the second rating of importance was chosen for the analyses because it included summaries of all problems, thereby avoiding missing data problems.

Preliminary Analysis

Prior to examining biases in importance ratings, it was necessary to determine if the problems in the prior knowledge sets varied in importance because importance was not used in developing the matched sets. If the problems were all of high importance, then there would be no variability in the ratings to examine performance of observers at rating importance as a function of prior knowledge. Table 3 shows the distribution of the experts' ratings for the two prior knowledge sets. Some of the problems in each set were highly important, moderately important, and of low importance. With a range in importance ratings, it was possible to examine biases.

Table 3

Frequency and percent of importance ratings by experts for each set of problems of which subjects were informed.

Importance Rating by Experts	<u>Problems in Set 1</u>		<u>Problems in Set 2</u>	
	Frequency	Percent	Frequency	Percent
>=4	3	33%	3	33%
=3	2	22%	1	11%
<=2	4	45%	5	56%

Calculation of Mean Ratings of Importance

To examine biases, for each subject a mean importance rating for each combination of the obviousness by prior knowledge conditions was calculated. For example, a mean of 3.2 for low obvious problems of which a

subject had prior knowledge indicated that the average importance rating for the problems in that condition was 3.2.

Analysis Conducted

A profile analysis with prior knowledge and obviousness as repeated measures factors and training as a between-subjects factor was conducted on average ratings of importance. This analysis examined biasing effects of obviousness, prior knowledge, and training on importance ratings.

Interaction effects on mean ratings of importance

The initial analysis included only the two balanced prior knowledge sets to determine if prior knowledge, obviousness, and training, interactively affected mean ratings of importance. A 3-way interaction was not found, and the 2-way interactions were non-significant between obviousness and training, and between prior knowledge and training. All F values for these interactions were less than 1. The analysis did, however, reveal a significant interaction between obviousness and prior knowledge ($F(2,44) = 3.56, p < .04, \eta^2 = .14$). This interaction, along with the ratings of the experts, is shown in Figure 7. A simple effects analysis for the low obvious problems revealed that subjects rated those problems of which they had prior knowledge as more important than those problems of which they had no prior knowledge ($F(1,22) = 6.98, p = .015, \eta^2 = .24$). The simple effects were not significant for the medium and high obvious conditions (both F values < 1). The mean ratings on the low obvious problems were examined. It was found that the mean rating for problems of which subjects had prior knowledge was 2.7, and the mean rating for problems of which they did not have prior knowledge was 2.2. These results are consistent with the hypothesis that subjects would give higher

importance ratings to those problems of which they were informed than those of which they were not informed, but only for the low obvious, subtle problems.

To determine if subjects were overestimating importance of informed problems or underestimating importance of non-informed problems, ratings assigned by the subjects were compared to ratings assigned by the experts. Figure 7 shows that, for low obvious problems, the bias was due to subjects underestimating the importance of problems of which they were not informed. The experts' mean importance rating for low obvious problems was 2.7. This mean rating is higher than the mean rating assigned by subjects to problems of which they were not informed, and identical to the mean rating for problems of which they were informed.

Figure 7. Mean ratings of importance by obviousness and prior knowledge.

Effect of obviousness on mean ratings of importance

Simple-effect tests for each prior knowledge condition (informed and not informed) in the above analysis revealed that there was a significant linear increase in the rating of importance with an increase in obviousness. Since the same effect occurred for both prior knowledge conditions, the analysis was simplified by examining the main effect of obviousness on mean ratings of importance. The analysis included all problems that occurred in the general summary. Statistical tests revealed a significant linear effect for obviousness ($F(1,44) = 121.02, p < .001, \eta^2 = .73$)⁶. Rated importance increased as obviousness of the problem increased. The mean rating of importance was 2.4 for the low obvious problems, 2.9 for the medium obvious problems, and 3.5 for the high obvious problems. As shown in Figure 8, these mean ratings are similar to experts' ratings in each of the obviousness conditions.

Figure 8. Mean ratings of importance by obviousness.

⁶ The effect was also significant when the analysis included only the two balanced sets of problems of which subjects were informed ($F(1,44) = 85.38, p < .001, \eta^2 = .65$).

Variability of subjects' ratings from the experts' ratings

While the mean ratings of the subjects are similar to the mean ratings of the experts, mean ratings do not indicate whether the observers gave exactly the same ratings as the experts. For example, a mean rating by the observers could be identical to the mean rating of the experts if one-half of the observers gave ratings two units higher than the experts and one-half gave ratings two units lower than the experts. The mean rating does not reflect the two-point differences in the ratings. It is therefore important to consider the variability of the subjects' ratings of importance around the experts' ratings of importance. To examine variability, the absolute value of the difference between the importance rating given by the experts and the importance ratings given by subjects was calculated.

$$\text{Absolute (Subject rating - Expert rating)}$$

The formula included the absolute value so that positive and negative differences between the experts' and a subject's ratings would not cancel out and sum to 0 to indicate no variability. By taking the absolute value, a score of 0 indicates no variability from the experts' rating of importance while a score greater than 0 indicates that a subject's ratings varied from that of the experts.

As shown in Figure 9, the mean variability on ratings of importance from the experts was greatest for the medium obvious problems than for low and high obvious problems. A significant quadratic effect was found ($F(1,44) = 15.50, p < .001, \eta^2 = .25$). The mean deviation from the experts was 1.3 for medium problems, 1.0 for low obvious problems, and 0.9 for high obvious problems.

Figure 9. Variability of subjects' importance ratings from experts' importance ratings by obviousness.

Further analyses revealed a correlation between obviousness and importance. The relationship between importance and obviousness is presented in Table 4. Based on ratings by the experts, the majority of low obvious problems were of low importance (54%), and few were of high importance (23%). Further, the majority of high obvious problems tended to be of high importance (52%), and few were of low importance (30%). For low and high obvious problems, subjects may be using the level of obviousness as a basis for making ratings of importance. For medium obvious problems, however, there is no systematic relationship between obviousness and importance. Again, based on the experts' ratings, 43% of the problems were of high importance and 43% were of low importance. Consequently, it may have been more difficult to make ratings of importance for medium obvious problems. This difficulty may account for the greater variability of observers' ratings from the experts' ratings.

Table 4

Frequency and percent of low, moderate, and high importance ratings by experts for each of the obviousness conditions.

Importance Rating by Experts	<u>Low obvious</u>		<u>Medium obvious</u>		<u>High obvious</u>	
	Frequency	Percent	Frequency	Percent	Frequency	Percent
>=4	3	23%	6	43%	14	52%
=3	3	23%	2	14%	5	18%
<=2	7	54%	6	43%	8	30%

Effect of training on mean ratings of importance

In addition to examining mean ratings of importance for the obviousness and prior knowledge conditions, an analysis on the effect of training on mean ratings of importance was also conducted. A significant main effect of training was not found ($F < 1$). Therefore, mean ratings of importance were not affected by the type of training that subjects received. The mean rating of importance was 2.2 for the extended training group, and 2.3 for the basic training group.

The variability of the importance ratings for the two training groups was also examined. Again, variability was calculated by examining the mean absolute difference between the subjects' ratings of importance and the experts' ratings of importance. There was a significant effect of training on the variability of observers' ratings from the experts ($F(1,22) = 7.17, p < .02, \eta^2 = .25$)⁷. Figure 10 shows that observers who viewed the training videotape were less variable from the experts in their ratings of importance (mean variability = 1.0) than subjects who did not see the videotape (mean variability = 1.2).

⁷ The effect was also significant when the analysis included only the two balanced sets of problems of which subjects were informed ($F(1,22) = 4.26, p = .05, \eta^2 = .16$).

Figure 10. Variability of subjects' importance ratings from experts' importance ratings by training group.

Qualitative Analyses on Other Problem Reports

The previous analyses examined the effects of various factors on the ability of observers to identify problems and rate importance in comparison to two experts. The following analyses examined (1) additional problems identified by observers that were not identified by experts, and (2) problems that subjects were informed about as potential problems but did not occur.

Additional Problems

Subjects not only missed some of the 54 problems in the general summary, they also recorded problems that experts did not report. There was a total of 56 additional problems reported in the general summary. These additional problems are described under the categories invalid, possibly valid, and valid.

Invalid problems identified by subjects

An invalid problem occurred when subjects gave incorrect descriptions of problems, or they reported problems that never occurred. These problems are false alarms. Invalid problems accounted for 82% of all additional problem reports. Examples of the two types of invalid problems are shown in Table 5 which also shows the frequency of invalid reports for each training group. Subjects in the basic training group reported 33 invalid problems while subjects in the extended training group reported only 13 problems. The mean number of invalid problem reports was 1.1 for subjects in the extended training group, and 2.8 for subjects in the basic training group. A t-test to examine this difference was significant ($t(22) = 2.45, p = .02$).⁸

Table 5

Description of invalid problems and frequency of invalid reports for each training group.

Category	Description and Example of Category	Frequency of Reports	
		Extended Training	Basic Training
Incorrectly stated	The subject incorrectly described the problem. (e.g., subject stated that the user used two buttons together because both buttons had to be used. The problem was that the user alternated between two buttons because there was no feedback on the display.)	5	16
Never happened	The subject described a behaviour or problem that never occurred. (e.g., subject reported that the user pressed POWER before pressing TEMP. This behaviour never took place.)	8	17
TOTAL		13	33

⁸ The t-test for the videotape summaries was not significant (see Appendix K).

It was also found that invalid problem reports are widespread. Sixty-seven percent of the subjects recorded one or more of the invalid problems. Further, subjects rated a large proportion of the invalid problems as highly important. Ratings of 4 or 5, to indicate high importance, were assigned to 43% of the problems that were stated incorrectly, and 56% of the problems that never happened. Considering that approximately 50% of invalid problem reports received high importance ratings, the presence of false alarms should be a concern in usability tests.

Possibly valid problems identified by subjects

Possibly valid problems occurred when subjects did not give sufficient information to determine if the problem was correctly identified. This problem type is shown in Table 6. The frequency of this problem type was very low; only one occurrence for the extended training group, and three for the basic training group.

Table 6

Description of possibly valid problems and frequency of possibly valid reports for each training group.

Category	Description and Example of Category	<u>Frequency of Reports</u>	
		Extended Training	Basic Training
Vague	The subjects' description was too vague to determine if they correctly saw the behaviour or problem. (e.g., subject described problem as "Use of unnecessary functions - MEMORY". It was not clear if the subject understood that the user thought MEMORY was supposed to be used in the sequence.)	1	3

Valid problems identified by subjects

There were no cases where subjects reported valid problems that the experts missed. Instead, valid problems are problems that involved different interpretations than the experts' interpretations, or problems that were really part of another problem. Examples of the two types of problems are shown in Table 7. There were very few valid additional problems for each training group.

Table 7

Description of valid problems and frequency of valid reports for each training group.

Category	Description and Example of Category	Frequency of Reports	
		Extended Training	Basic Training
Possible interpretation	The subjects' interpretation of the problem was different from the experts. The subject or experts may have interpreted more from the problem. (e.g., subject specified that user did not like the lack of a colon when time is entered. Experts indicated that user was confused by lack of feedback and so had difficulty performing task.)	2	2
Sub-problem	The problem was part of another problem and did not have to be specified separately. The subject separated one aspect of the problem as a separate type of problem. (e.g., subject specified that additional buttons (TEMP and HOLD WARM) were not necessary but were used. This was part of the problem of not knowing how to enter the temperature (which the subject already identified).)	0	2
TOTAL		2	4

Prior Knowledge of Non-Existent Problems

In each of the sets of problems of which subjects had prior knowledge, there were three potential usability problems that did not occur in any of the videotapes. Subjects' reports of problems were examined to determine if subjects identified these as usability problems. One subject reported one of the non-existent problems, five subjects reported another problem, and none of the subjects reported the third non-existent problem. Therefore, prior knowledge of potential, but non-existent problems did not lead to a bias. Very few subjects reported problems because they were told the problems might occur.

Other Analyses

In addition to the analyses reported above on performance at identifying problems and rating importance, other analyses were conducted. The results were generally idiosyncratic to the specific usability test in this study or they showed no statistically significant difference. The results of the analyses are reported in the appendices. Appendix O describes the characteristics of problems which subjects had difficulty identifying and those they identified easily. It also describes problems for which subjects' ratings of importance were similar to experts' ratings of importance, and problems for which ratings were not similar to experts. Appendix P presents results on the effects of gender and microwave usage on problem identification and rating importance. Appendix Q is an analysis on observer accuracy at identifying the number of users who experienced each usability problem.

Interrater Agreement

Interrater agreement was calculated to determine consistency among subjects, independent of the experts. Agreement was calculated for problem identification and importance ratings.

Problem Identification

Interrater agreement was based on occurrence agreement percentage. The formula can best be described by presenting a 2 x 2 table to represent the incidence of agreements and disagreements (see Table 8).

Table 8

2 x 2 table to represent agreements and disagreements among observers.

		Observer 2	
		Occurrences	Nonoccurrences
Observer 1	Occurrences	A	B
	Nonoccurrences	C	D

Cells A and D are agreements on occurrences and nonoccurrences, respectively, while cells B and C are disagreements. Using this table, the formula⁹ for occurrence agreement percentage is as follows:

$$\% \text{ agreement} = \frac{\text{number of agreements on occurrences (A)}}{\text{number of agreements (A) + number of disagreements (B + C)}} \times 100$$

⁹ Occurrence agreement was calculated instead of total agreement $(A+D)/(A+B+C+D)$ because this study was directed toward occurrences of a problem, that is, observers did not have to record nonoccurrences of problems. Therefore, occurrence agreement represents the focus of the study better than total agreement.

First, percentage agreement was calculated for each pair of subjects (e.g., subject 1 with subject 2, subject 1 with subject 3). For each subject, mean agreement with all other subjects was calculated. Finally, to obtain the overall percentage agreement, the mean of the interrater agreement for all subjects was taken.

The overall interrater agreement was 62% for identifying problems. The accuracy measurement (ratings compared to the experts) was slightly higher (66%)¹⁰.

Importance Ratings

Lawlis and Lu's (1972) formula for ordinal and interval scales was used to calculate interrater agreement on importance ratings. In this formula, agreement is not calculated in an all-or-none fashion because ratings do not have to be identical to be considered as agreement. Instead, criteria are set to identify the seriousness of disagreements among raters. Ratings of importance differing by 1 point or less indicated agreements and differences greater than 1 denoted disagreements. Using this criteria, the percentage agreement was calculated as follows:

$$\% \text{ agreement} = \frac{\text{number of agreements}}{\text{number of agreements} + \text{number of disagreements}} \times 100$$

¹⁰ The results were similar on the videotape summaries. The interrater agreement (67.2%) was slightly lower than the accuracy (70.4%).

Since interrater agreement was based on Lawlis and Lu's (1972) formula, agreement of subjects with experts was also calculated using this formula. The mean interrater agreement for importance ratings was 64%, and the mean agreement with experts was slightly higher at 67%.

Opinions of the Observational Task and Training

Several measures were taken to examine subjects' opinions of the observational task and training. Opinions were collected in a questionnaire that consisted of 7-point rating scales. Of those included in the following analyses, 25 were bipolar scales with a negative and positive opinion on the extremes of the scale. Four of the scales were unipolar scales. Also, four of the scales had negative opinions on the end-points of the scales and the mid-point was labelled as a positive opinion. Table 9 shows the opinion measures derived from the rating scales. Some of the opinions were based on two or more individual ratings with responses to related items averaged together. The table shows the number of ratings that were related and combined into single indices. Also, for the opinions based on multiple items, Cronbach's alpha, a measure of inter-item reliability is shown. The alphas indicate relatively high inter-item reliability for the multiple-item indices.

Table 9

List of opinions on the task and training.

Dependent Variable	Number of Items	Cronbach's Alpha
Ease of observational task	7	.69
Appropriateness of duration and detail of training	4	.68
Ease/organization of training	9	.85
Effectiveness and importance of training	9	.92
Usefulness of training for learning about usability testing and observing	3	.84
Overall quality of the training	1	-

A multivariate analysis of variance was conducted to examine opinions on the task and training. The independent variable was exposure to or no exposure to the training videotape. The dependent variables were the opinions listed in Table 9. The analysis revealed a significant multivariate main effect ($F(6,17) = 3.92, p < .02, \eta^2 = .58$). Using Bonferroni's post hoc multiple comparison test, the alpha for the tests on the six dependent variables was set to .002 (.05/6). At this alpha level, all six univariate tests were non-significant.

Average ratings for the two training groups on five of the opinion measures from Table 9 are presented in Figure 11. The figure shows that subjects in the extended training group gave less positive ratings on all of the five measures than subjects in the basic training group. The opinion rating on "Appropriateness of duration and detail of training" is not included in Figure 10 because the rating scale was not directly comparable to the other rating scales. The scale ranged from too much (1) to not enough (7) with the mid-point (4) labelled as about right. The mean rating on this scale was 4.1 for the basic training group and 4.0 for the extended training group.

Figure 11. Opinions of the training program by training group.

Subjects' open-ended comments about the training were also examined. The most frequent comments were that the training should incorporate more practice (reported by 8 subjects), more examples of how already trained persons record (reported by 3 subjects), and more detail on the observational task (reported by 4 subjects). Subjects in both training groups made these comments. It is interesting to note, however, that subjects rated the adequacy of details and examples as about right. Even though the training was rated as acceptable on most training measures, subjects stated, in open-ended questions, a need for more detail about the task at hand.

Subjects were also asked if they thought any of the components of the training program could be omitted. Table 10 shows, for each training group, the number of subjects who thought each component could be omitted. Very few of the subjects indicated that any of the components could be eliminated. However, three subjects (25%) who saw the training videotape indicated that it could be omitted.

Table 10

Frequency of subjects in each training group who thought components of the training could be omitted from the training program.

Components of Training	Extended training	Basic training
Introduction to usability testing	0	1
Training videotape	3	N/A
Presentation on how to use microwave	1	1
List of problems which may occur	1	2
Oral instructions for completing coding forms	1	0
Written instructions for completing coding forms	1	1
Sample forms	0	0

Opinions of the Training Videotape

Subjects in the extended training condition were asked their opinions about the training videotape. Opinions were rated on 7-point bipolar scales. Table 11 shows results for all of the videotape training questions. Two opinions, usefulness of the videotape and its organization, were based on several items that were averaged together. The number of ratings combined into a single measure, and Cronbach's alpha for multiple-item measures are shown in Table 11. Cronbach's alphas indicate that the items had high internal consistency.

Table 11

Opinions of the training videotape.

Opinion	Number of Items	Cronbach's Alpha
Usefulness of the videotape	5	.93
Organization of the videotape	3	.85
Helpful for learning role of usability testing in product design	1	-
Helpful for learning observer's role in product design	1	-
Helpful for learning how to observe	1	-

The average ratings for each opinion score on the videotape are presented in Figure 12. The average ratings on all measures are above 4 on the 7-point scale, indicating that subjects regarded the videotape as valuable. The highest ratings were for organization of the videotape and for learning the roles of the observer and usability testing in product design. The lowest ratings were on overall helpfulness for observing and usefulness for learning how to observe. These ratings are consistent with subjects' comments that the training requires more practice and more specific examples on how to observe.

Figure 12. Opinions of the training videotape.

Discussion

Design Summary

This study examined the performance of novice observers in a usability test. Three factors that may affect observer performance were considered; obviousness of usability problems, prior knowledge of potential problems, and training. The effects of these variables on three performance measures were examined. Performance measures included accuracy at identifying usability problems, biases in rating the importance of fixing the problems, and variability in rating the importance of the problems relative to two usability experts.

Training provided to observers was manipulated by providing two forms of training. One-half of the subjects received procedural training on how to observe and record data (basic training group), and one-half of the subjects received a training videotape in addition to procedural training (extended training group). The training videotape gave an overview of the role of usability specialists in product design, and described the use of assessment techniques for evaluating products.

Prior knowledge of problems was manipulated by instructing subjects to watch for selected potential usability problems. Subjects were told about one of two sets of potential problems.

Obviousness of each problem identified in the usability test was determined by two usability test experts. The experts rated each problem as either low, medium, or high in obviousness.

The results for the effects of obviousness and prior knowledge are discussed first. Training did not interact with obviousness or prior knowledge, so the results concerning training are discussed separately.

Effects of Obviousness and Prior Knowledge on Problem Identification

It was hypothesized that obviousness and prior knowledge would interactively affect accuracy at identifying problems. The interaction was not significant, but the results showed the predicted pattern. Awareness of potential usability problems improved observer accuracy at identifying problems for the low and medium obvious problems. For the high obvious problems, there was no difference in observers' ability to identify problems for which they had prior knowledge or did not have prior knowledge. Although the interaction was non-significant, obviousness and prior knowledge may be important variables to consider when assessing observer performance at identifying problems in a usability test.

A main effect of obviousness was hypothesized. It was predicted that subjects would identify the high obvious problems more easily than the medium or low obvious problems. This hypothesis was confirmed. Subjects identified approximately one-half of the low obvious problems, and identified slightly more than one-half of the medium obvious problems. Accuracy scores were nearly perfect for the high obvious problems.

The effect of obviousness of problems on the ability of observers to identify those problems is consistent with previous research on other characteristics of what is being observed. Factors such as frequency of occurrence of events have been related to performance of observers at detecting the events (House & House, 1979; Kapust & Nelson, 1984). The effect of obviousness has not been previously examined, but in the context of a usability test, as demonstrated in the current study, it is an important variable.

The effect of obviousness on problem identification has important implications for the use of novice observers in usability testing. The results clearly indicate that observers are more sensitive to some problem types than others. Problems that are of low or medium obviousness, as defined in the current study, may not be detected. This result is of particular interest when the undetected problems are of high importance. Critical recommendations that should be made for a product could be overlooked. Companies who conduct usability testing should be concerned about these results and their implications, particularly if they employ observers who have no or little experience with observing in a usability test. Observer training that emphasizes detection of problems that vary in obviousness should be provided.

Further, it is important to note that obviousness of problems is multidimensional. For example, frequency and duration of a behaviour, and comments made by the user could affect the obviousness of a problem. In the current study, obviousness of each problem was established by two experts. It would be of interest to attempt to relate experts' ratings to behavioural dimensions to determine which dimensions are important to the obviousness of problems. Observational training could then sensitize the observers to characteristics of user behaviours which are associated with low obvious problems. For example, if duration affects problem obviousness then observers could be instructed to identify problems that are of short duration.

The results also revealed a marginally significant main effect of prior knowledge on accuracy at identifying problems. Subjects identified more of the problems of which they were informed than those of which they were not informed. Therefore, informing subjects of potential problems

and instructing them to look for the problems, and record them if they occur, was beneficial and heightened attention to these problems. It is important to note, however, that this result also suggests that potential problems that are not identified in advance of the usability test, are less likely to be detected. This finding is of concern because problems that are unanticipated may be the most important to detect precisely because they are not expected. Further, sometimes anticipated problems with products may be fixed even if a usability test is not conducted. With unanticipated problems, however, recommendations for changes would not occur unless the problem was detected.

Although hypothesizing and looking for potential problems did improve observer performance, it is possible that there may be disadvantages to predicting usability problems. Expectations of problems may lead to biases. For example, trainee observers may report usability problems that are anticipated to occur, but in actuality, do not occur. In this study, the possibility of this effect was examined. Subjects were instructed to look for some problems that did not occur. Only six of 24 subjects incorrectly reported at least one of the hypothesized, but non-existent problems. Reporting of non-existent problems of which subjects were informed was rare in this study, and so detrimental effects were not found. This result, however, could be a function of the non-existent problems selected for this study. Further, although this study did not address the effect of hypothesizing problems on the detection of unanticipated problems, it is possible that a focus on selected potential problems could cause observers to miss problems of which they have no prior knowledge. The possibility of this effect should be investigated further. This effect could be examined by including a condition in which

none of the observers are given prior knowledge of some problems. Performance at identifying a set of problems of which no one was informed could be compared for the different conditions to determine if observers who did have prior knowledge of some problems identified problems differently than observers who did not have prior knowledge of any problems. If prior knowledge does cause observers to miss unanticipated problems, it may be necessary to train observers to watch for problems that are not expected.

Effects of Obviousness and Prior Knowledge on Importance Ratings

The effects of obviousness and prior knowledge on ability to rate importance were examined. Analyses on the mean ratings of importance revealed an interaction between obviousness and prior knowledge. Observers' ratings were biased for low obvious problems because subjects underestimated the importance of low obvious problems of which they were not informed. A biasing effect did not occur for medium and high obvious problems. These results are consistent with the predicted hypotheses. In addition, the results are in agreement with previous research which has reported biases on subjective judgments (Kent et al., 1974; Shuller & McNamara, 1980). In this case, the bias occurred for only subtle (low obvious) problems.

The bias effects that occurred in this study are consistent with previous research. In this study, there was not a bias effect for identifying usability problems, but there was a bias for rating the importance of problems. Previous research has shown that bias effects tend to be stronger when making subjective judgments than when recording objective data (Foster & Cone, 1986; Harris & Lahey, 1982). These results suggest that

training programs that are designed for usability test observers, should focus on the type of bias that occurred in the current study.

The results also revealed that subjects' ratings of importance varied from the experts' ratings more for medium obvious problems than for low and high obvious problems. This result may be specific to the particular set of problems used in this study; obviousness predicted importance in the high and low obvious conditions, but not the medium obvious conditions. High obvious problems tended to be of high importance, and low obvious problems tended to be of low importance. Because observers ratings were less variable from the experts for low and high obvious problems than on medium obvious problems, it seems that observers were using obviousness as a basis for their importance ratings.

Information is required to the extent to which naturally occurring usability problems show a correlation between obviousness and importance. The usability problems in this study were scripted, and may or may not be representative of other usability problems. Low obviousness problems may be important, and high obvious problems may not be important. If the relationship between obviousness and importance is not high, performance at rating importance of problems might be lower than found in this study. Failing to identify important problems, and misclassifying the importance of problems can lead to inappropriate or even incorrect product design recommendations. Therefore, it may be essential to train observers to identify problems of all levels of obviousness, and to teach them to make judgments of importance independent of the obviousness of the problems.

Effects of Training

Although training did not interact with obviousness or prior knowledge, it did have limited, but interesting effects. Analyses showed no statistically significant differences between training groups on accuracy at identifying problems, and on mean ratings of importance for fixing the problems. Training did, however, affect the variability of subjects' ratings of importance from the experts' ratings of importance. Subjects in the extended training condition rated importance of problems with less variability from the experts than subjects in the basic training condition.

There are possible explanations for better performance of the extended training group over the basic training group when rating importance but not when identifying problems. One explanation relates to the contents of the training videotape. The videotape, which was presented prior to procedural training, gave an overview of interface specialists and their use of techniques, such as usability testing, for assessing product designs. A few points presented in the videotape may have played a key role in the performance of observers. First, it described the impact of usability testing, and other evaluation techniques, on product acceptance. Second, and possibly more importantly, it showed that usability test observers may have to defend their recommendations for product design. Therefore, subjects who viewed the training videotape knew that observers must account for their design choices. By providing this knowledge, the training videotape probably helped subjects understand the implications of their usability test results. They, therefore, may have been more careful when rating importance of problems than subjects who did not see the videotape. However, both groups of observers performed equally well on the objective

task of identifying problems. This result likely occurred because both groups received the same procedural training on how to observe and record usability problems. Because procedural training was not included in the videotape, subjects who saw the videotape would not have had an advantage on this task.

A second and striking difference between the two training groups was that subjects in the basic training group recorded many more invalid problems than subjects in the extended training group. The problems were invalid because they never occurred, or they were not stated correctly. This result supports the notion that subjects in the extended training condition, in comparison to subjects in the basic training condition, were more careful when assessing user performance because they better understood the goals and role of usability testing in product design. Subjects in the basic training condition probably did not understand the implications of their decisions as well as subjects who did see the videotape, and tended to look for problems in all user behaviours and verbalizations.

A third difference between the training groups was that subjects in the extended training group gave lower scores for quality of the training than subjects in the basic training group. This pattern suggests that because subjects in the extended training group understood the implications of usability testing better than subjects in the basic training group, they also recognized a need for more, better, or different training. Subjects in both training groups made verbal comments about the weaknesses of the training. Some subjects felt that the training did not provide practice at identifying problems, did not provide an opportunity to compare their data to that of an expert, and did not provide enough detail on the observational task. Although the training videotape was of some value

for rating the importance of fixing usability problems, it was not sufficient, along with the procedural training, for novice usability test observers. An appropriate format for the training is unknown, but guidelines for an effective training program can be provided after further research.

Effect of Practice

Ability to identify problems was comparable on the three videotape trials. A failure to improve performance in this study does not suggest that performance would not improve with practice. It only indicates that three videotapes were not sufficient for a noticeable improvement to occur. Subjects may need more time to feel comfortable with the task of identifying usability problems, and becoming familiar with the types of usability problems that might occur. Further, the nature of the training given to observers may have affected the rate of improvements. More training may be required, or training programs may need to incorporate learning techniques such as feedback or modelling. A practical solution will have to be investigated. The videotapes developed for the current study could be used in a second study to examine the effects of feedback whereby observers are allowed to compare their results to those of the experts after viewing each videotape.

Interrater Agreement versus Accuracy

Both accuracy and interrater agreement were examined. Accuracy measures how well subjects agree with the experts, and interrater agreement measures how well subjects agree amongst themselves. Accuracy and interrater agreement scores were similar for performance at

identifying problems, and for performance at rating importance. Interrater agreement was slightly lower than accuracy. From these results, it may be inferred that interrater agreement may be used as an estimate of accuracy. Reid (1970) has noted that researchers have estimated accuracy on the basis of interrater agreement scores. However, there is not a direct relationship between the two measures, and so further studies should report both measures of observer performance. Also, when choosing a measure for assessing observer performance, the tradeoffs between agreement and accuracy should be considered. While interrater agreement is more practical to assess because a criterion (scoring key) does not have to be established, a weakness of interrater agreement is that the correct observer cannot be identified when there are discrepancies. With accuracy, there is more confidence that the criterion is correct if experts establish the criterion. Therefore, when assessing performance in a usability test, accuracy is the superior measure.

Limitations of the Current Study

Several dimensions important to observing in a usability test were not explored in this study. In addition to ability to identify problems and rate the importance of problems, there are other skills that usability testers require. One skill is the ability to interpret the data and make recommendations to improve the design of the product. A second skill is ability to interact with users. This interaction involves knowing when to prompt users for more information, knowing when to give hints to users, and establishing a rapport with users so they are comfortable in the testing

situation. It is unknown from this study how well observers perform these various skills.

There are other issues that this study did not address. It did not examine other characteristics besides obviousness of problems, that may influence performance. There may be other important characteristics to consider when identifying problems or when rating the importance of fixing the problems. Further, while this study examined effects of prior knowledge as a function of being told about possible problems, it did not examine the impact of observers' own predictions on performance.

Finally, this study did not examine the type of experience required of usability testers. Rosenbaum (1989) suggested that usability testers should have knowledge of experimental design and human factors literature. Subjects in this study were senior undergraduate students and graduate students familiar only with experimental design and methodological issues. It is unknown whether or not lack of knowledge in the human factors domain was a factor in performance. The knowledge and level of expertise necessary for successful usability testing should be investigated.

Conclusions

Many interesting and useful results were found in this study. First, both obviousness of usability problems and prior knowledge of problems can have a significant impact on performance at identifying problems and rating the importance of problems. More obvious problems are likely to be identified, and prior knowledge can improve ability to identify problems. Second, not having prior knowledge of low obvious problems can lead to an underestimation of the importance of those problems. Third, observers'

ratings of importance are more variable from the experts' ratings of importance for medium obvious problems than for low and high obvious problems. Fourth, preinstructional training on the goals of user interface specialists and the use of techniques such as usability testing, can reduce variability when rating the importance of problems, and reduce the number of invalid, non-existent problem reports.

All of these results suggest that training programs need to be developed for novice observers of a usability test. There needs to be an emphasis on instructing observers how to identify problems, and how to rate the importance of problems, which vary in obviousness and which may be anticipated or unanticipated prior to the usability test. Training must also reduce reporting of invalid problems. Because inaccurate data collection can lead to inappropriate or incorrect design recommendations, the performance of observers must be seriously considered.

References

- Bayman, P., & Mayer, R. E. (1988). Using conceptual models to teach BASIC computer programming. Journal of Educational Psychology, 80, 291-297.
- Bergfeld-Mills, C. (1986). Usability testing in the real world. In M. Mantei & P. Orbeton (Eds.), Human Factors in Computing Systems - CHI '86 Proceedings, (Boston, April 13-17), pp. 212-215.
- Carroll, J. (1984). Minimalist training. Datamation, Nov., 125-136.
- Carroll, J. (1987). Minimalist design for active users. In R. M. Baecker & W.A.S. Buxton (Eds.), Readings in Human-Computer Interaction: A Multidisciplinary Approach, Los Altos, California: Morgan Kaufmann Publishers.
- Cohen, J. (1960). A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 20, 37-45.
- Dorsey, B. L., Nelson, R. O., & Hayes, S. C. (1986). The effects of code complexity and of behavioral frequency on observer accuracy and interobserver agreement. Behavioral Assessment, 8, 349-363.
- Doyle, W. H. (1986). Using an advance organizer to establish a subsuming function concept for facilitating achievement in remedial college mathematics. American Educational Research Journal, 23, 507-516.
- Foster, S. L., & Cone, J. D. (1986). Design and use of direct observation procedures. In A. R. Ciminero, K. S. Calhoun, & H. E. Adams (Eds.), Handbook of Behavioral Assessment, second edition. New York: John Wiley & Sons.
- Gould, J. D., & Lewis, C. (1983). Designing for usability - key principles and what designers think. In A. Jandra (Ed.), Human Factors in Computing Systems - CHI '83 Proceedings, (Boston, December 12-15), pp. 50-53.
- Harris, F. C., & Ciminero, A. R. (1978). The effect of witnessing consequences on the behavioral recordings of experimental observers. Journal of Applied Behavior Analysis, 11, 513-521.
- Harris, F. C., & Lahey, B. B. (1982). Recording system bias in direct observational methodology: A review and critical analysis of factors causing inaccurate coding behavior. Clinical Psychology Review, 2, 539-556.

- Hartley, J., & Davies, I. K. (1976). Preinstructional strategies: The role of pretests, behavioral objectives, overviews and advance organizers. Review of Educational Research, 46, 239-265.
- House, B. J., & House, A. E. (1979). Frequency, complexity, and clarity as covariates of observer reliability. Journal of Behavioral Assessment, 1, 149-165.
- Jacob, T., Grounds, L., & Haley, R. (1982). Correspondence between parents' reports on the Behavior Problem Checklist. Journal of Abnormal Child Psychology, 10, 593-608.
- Johnson, S. M., & Bolstad, O. D. (1973). Methodological issues in naturalistic observation: Some problems and solutions for field research. In M. A. Hamerlynck, L. C. Handy, & E. J. Mash (Eds.), Behavior Change: Methodology, Concepts, and Practice. Champaign, Illinois: Research Press Company.
- Kapust, J. A., & Nelson, R. O. (1984). Effects of the rate and spatial separation of target behaviors on observer accuracy and interobserver agreement. Behavioral Assessment, 6, 253-262.
- Kazdin, A. E. (1977). Artifact, bias, and complexity of assessment: The ABC's of reliability. Journal of Applied Behavior Analysis, 10, 141-150.
- Kent, R. N., O'Leary, K. D., Diament, C. & Dietz, A. (1974). Expectation biases in observational evaluation of therapeutic change. Journal of Consulting and Clinical Psychology, 42, 774-780.
- Lawlis, G. F., & Lu, E. (1972). Judgment of counseling process: Reliability, agreement, and error. Psychological Bulletin, 78, 17-20.
- Luiten, J., Ames, W., & Ackerson, G. (1980). A meta-analysis of the effects of advance organizers on learning and retention. American Educational Research Journal, 17, 211-218.
- Lund, M. A. (1985). Evaluating the user interface: The candid camera approach. In L. Borman & B. Curtis (Eds.), Human Factors in Computing Systems - CHI '85 Proceedings, (San Francisco, April 14-18), pp. 107-113.
- Mayer, R. E. (1979). Can advance organizers influence meaningful learning. Review of Educational Research, 49, 371-383.
- Mayer, R. E., Dyck, J. L., & Cook, L. K. (1984). Techniques that help readers build mental models from scientific text: Definitions pretraining and signaling. Journal of Educational Psychology, 76, 1089-1105.

- Michelson, L., Mannarino, A., Marchione, K., Kazdin, A. E., & Costello, A. (1985). Expectancy bias in behavioral observations of therapeutic outcome: An experimental analysis of treatment and halo effects. Behavior Research and Therapy, 23, 407-414.
- Reid, J. B. (1970). Reliability assessment of observation data: A possible methodological problem. Child Development, 41, 1143-1150.
- Repp, A. C., Nieminen, G. S., Olinger, E., & Brusca, R. (1988). Direct observation: Factors affecting the accuracy of observers. Exceptional Children, 55, 29-36.
- Rosenbaum, S. (1989). Usability evaluations versus usability testing: When and why?. IEEE Transactions on Professional Communication, 32, 210-216.
- Shuller, D. Y., & McNamara, J. R. (1980). The use of information derived from norms and from a credible source to counter expectancy effects in behavioral assessment. Behavioral Assessment, 2, 183-196.
- Skindrud, K. (1972). An evaluation of observer bias in experimental-field studies of social interaction. Unpublished doctoral dissertation, University of Oregon, Eugene.
- Spool, M. D. (1978). Training programs for observers of behavior: A review. Personnel Psychology, 31, 853-888.
- Stone, C. L. (1983). A meta-analysis of advance organizer studies. Journal of Experimental Education, 51, 194-199.
- Tudor, I. (1986). Advance organizers as adjuncts to L2 reading comprehension. Journal of Research in Reading, 9, 103-115.

Appendices

Appendix A

Script of Videotape for Extended Training Group

Note: The videotape consisted of a narrator and several acts. Bold italic text is used to indicate what is shown on the videotape (narrator, a diagram or an act). Plain text is the words spoken by either the narrator or those who acted. Unbold italic text is a description of what was occurring in the act and who was speaking.

THE ROLE OF USER INTERFACE SPECIALISTS IN PRODUCT DESIGN

While showing narrator, he says:

In this videotape I will tell you about the role of user interface specialists in the design of products and describe techniques that specialists use to improve product design.

Computerized products we use everyday such as telephones, VCR's, microwaves, and even our home and office computers, have become increasingly complex with the advance of technology. We're required to learn, on our own, the purpose of each feature and how to use the features. This is not always easy to do.

Act 1:

There is a user using a telephone trying to set up a conference call between two other individuals.

While on the phone, user says:

OK. I'm going to try to set up a conference call with you and Jill. Let me see if I can get Jill on the phone at the same time.

User hesitates and looks confused, gets manual to look up conference feature. Touches flash button and says:

FLASH!

User dials the number of Jill and says:

Hi Jill, I've got Tim on the phone. I want to set up a conference call with the three of us so if you can hang on for a second, I'll try to get us all connected.

User looks at manual again trying to decide what to do next. Touches Conference button and says:

Jill are you there? How about you Tim?

While showing narrator, he says:

Because of the increase in complexity of products and an increasing difficulty in using them, designers are aware that users should be consulted during the development of new products. Feedback from users helps designers determine which features are required in a product. It also indicates if it is easy to learn, and is easy to use.

Getting user feedback during product development allows designers to modify features which are unsatisfactory or difficult to use before the product is put on the market.

While showing a chart of a product design team, narrator says:

While all members of the design team, including User Interface Specialists, System Developers, Engineers, and Industrial Designers work together to design the product, it is the Interface Specialist who collects the feedback from users.

While showing a chart of a product, user, and interface specialist, narrator says:

The specialist consults with users to determine their needs and use this information to design a better product for the users.

While showing a chart of user input techniques, narrator says:

User interface specialists use three techniques to get user input. These techniques are task analysis, usability testing and, for some products, a field trial. I will describe each of these techniques with an emphasis on the usability testing approach.

While showing the chart of user input techniques with a check mark beside the task analysis item, narrator says:

The task analysis, the first user input technique, is conducted before the product is designed.

While showing narrator, he says:

It is used to understand users' tasks, examine the skill level of users, obtain the frequency with which tasks are performed, and determine the technological requirements for completing the tasks. This is done by examining users complete tasks in their workplace or home and by

conducting structured interviews with users to ask them about what they want the product to do -- that, is, its functionality.

Act 2:

There is a user interface specialist interviewing a user.

Interface Specialist: I would just like to ask you a few questions about your use of the phone on your desk. Do you use all of the features that are on the phone?

User: Well no. I can't remember how to use the majority of the features so I really only use a few features such as forward, redial, and storing numbers.

Interface Specialist: Is there anything that you dislike about the features you do use?

User: No, not really. I find them to be pretty straightforward unlike some of the others. I especially like being able to store numbers and using the redial button. I often call the same people back in a day and so those features are really handy.

Interface Specialist: Are there any features you would like to see added to your phone which would be useful to you in your current work situation?

User: I guess it would be nice to have a log kept of the numbers I've called but were busy. With this log then I would know who I've already tried but still have to get in touch with.

While showing narrator, he says:

Based on the information collected in the task analysis, the user interface and the functionality of the product are designed.

Act 3:

A designer and interface specialist are discussing how to design a telephone.

Designer: What did you find out from the users?

Interface Specialist: Well in the interviews I conducted, a number of people wanted a log of the calls they made that were busy so I think this feature should be added.

Designer: What do you mean? I don't understand what a log is.

Interface Specialist: Well, when they call a number that is busy they wanted that number placed in a file of numbers so that when they select that number they get an automatic redial of it.

Designer: That sounds interesting. Why don't we try it. We have some memory limitations. If we can get enough items in there, say more than 5, we'll be OK. If we can't then we'll be in a little bit of trouble.

Interface Specialist: Well, I think if you can do it with 5 calls then this would be very useful to them.

Designer: You might be right. Lets check it. I'll see how much we can do with the memory we have available. My argument is that if we can get at least 5, that's fine, but if we can't get 5, then I don't think we should do it.

Interface Specialist: That sounds reasonable to me.

While showing the chart of user input techniques with a check mark beside the usability test item, narrator says:

Once initial designs are available, the features undergo usability testing to determine whether or not the features are suited to the users' needs, are easy to learn and are easy to use.

While showing narrator, he says:

In a usability test a few typical users perform specific tasks with a product. While the user performs the tasks, an observer records problems the user has and identifies what the user likes or dislikes about the product.

While showing chart of stages of usability testing, narrator says:

There are 3 major stages to a usability test. They are design the test, conduct the test, and then summarize the data and make recommendations. I will describe each of these stages in more detail.

While showing chart of stages of usability testing with check mark beside first stage (Design the test), narrator says:

To design a usability test, the user interface specialist must create a task script, and make available a prototype of the product.

While showing narrator, he says:

A task script is a list of the tasks the user is asked to perform with the product. This list includes typical tasks for which the product is designed.

While showing chart of a task, narrator says:

For example, to test a telephone, one task might be "Assume you have received a call intended for someone else in the office. Transfer the call to this person at extension 3825."

While showing chart of another task, narrator says:

A second type of task might be "Store the telephone number 351-9046 in a 'speed dial' button so that in the future you can press that button and the number will be automatically dialed for you."

While showing narrator, he says:

Thus the script is a list of tasks the person is going to follow while using a prototype.

The prototype is a mock-up of the product. It has working features to simulate what the real product will do. For instance, a prototype of a telephone might look like this.

While showing prototype of a telephone, narrator says:

This prototype is on a computer. It has number buttons and special feature buttons. The users performs tasks on the prototype by selecting the buttons with a pointer.

Act 4:

User is using the prototype of the telephone, performing a task.

User lifts the receiver and says:

I'll lift the receiver to dial a number.

User enters numbers on the phone. The phone dials, user hangs up and says:

Then to dial that number again I use the redial button.

User lifts the receiver, touches the redial button, and hangs up.

While showing chart of stages of usability testing with check mark beside second stage (Conduct the test), narrator says:

The user test is conducted when the prototype and task script are available.

While showing narrator, he says:

Users try to do the activities listed in the task script on the prototype. They are instructed to verbalize their thoughts as they perform the tasks.

Act 5:

User is using the prototype of the telephone performing a task and verbalizing his thoughts. Users says:

To transfer a call, I guess I put the caller on hold first so I'll touch hold.

Users touches hold and says:

Then I would use transfer, dial the number.

User touches transfer and dials the number and then says:

Then I guess I can just hang up to complete the transfer.

User hangs up.

While showing user using the prototype of the telephone, narrator says:

Typically, users are instructed to do tasks without using manuals because consumers often fail to use manuals when trying to use a product. Also, by observing how the users work with the product without a manual, the user interface specialist can more easily identify where the product is not intuitive or obvious.

While showing user using the prototype with observer watching the user, narrator says:

While the user performs each of the tasks, the observer records the user's actions on a coding form.

Act 6:

User is using the prototype of the telephone performing a task and observer is watching. Users says:

Now I have to store the number 3519046 in a speed dial button.

User touches buttons on the phone, and says:

Leave the receiver down. Then I think I'll dial the number.

User dials the number, and then says:

Then I think I touch store and another number. I guess its also possible that I do store and the number before dialing, but since I've already gotten this far, I'll go ahead and try it. So I touch store and 3.

User touches store and 3, lifts receiver and says:

I'll test that now by touching code and 3.

User touches code and 3 and then says:

Great. I got it. That wasn't too difficult.

While showing the user and observer, narrator says:

The observer records problems the user has, tasks they perform well, and the essence of any important comments which indicate satisfaction or dissatisfaction with the product.

It is not always necessary to have the observer in the room with the user. Sometimes the usability test is videotaped and the user's actions and comments analysed later.

Also, in this film, the observer is the same person as the interface specialist, but this is not always the case.

While showing chart of stages of usability testing with check mark beside third stage (Summarize the data and make recommendations), narrator says:

After a few typical users are observed, the data are summarized and recommendations for improving the product design are made.

While showing narrator, he says:

For the summary, the performance of all users is compared to determine the extent to which users had similar problems and to identify what the users liked or were satisfied with. Here is an example of an observer summarizing data collected in a usability test.

Act 7:

An interface specialist is looking through the data collected and summarizing it onto another form. The specialist is saying what they are thinking as they do this task.

The first task is going to be call transferring. This user couldn't transfer the call to the other extension because she put the caller on hold before touching the transfer button. I wonder if any of the others users had that same problem.

Specialist looks through papers one by one and says:

This user did the same thing. This user also had that problem. This user did it correctly.

Specialist writes on a paper and says:

So three of the users were unable to transfer the call. They touched the hold button before the transfer button. This left the caller on hold waiting for a connection that was never made.

Specialist looks through papers one by one again and says:

Now for the second task, storing numbers. This user did it correctly and commented that she liked the feature. This user also did it correctly. This user lifted the receiver before starting to store, but then quickly realized the receiver needed to be down and corrected it, and did it all right. This user did it correctly and also commented on liking the feature.

Specialist writes on a paper and says:

So, three of the users were able to store the number without problems and the fourth had a minor problem. Two of the 4 users also made a comment that they like the feature.

While showing narrator, he says:

The summary is then used to make recommendations for product improvement. In making recommendations, it is important to consider the seriousness of problems and their impact on ability to use the product and satisfaction with the product. If the results indicate that users had a lot of difficulty with a particular feature then a recommended change for that feature may be given a higher priority over other recommendations which are of a less serious nature. For instance, the observer shown in the next clip is prioritizing problems users had with the telephone.

Act 8:

An interface specialist is making recommendations for how to improve the product. While writing on paper, the specialist is saying what they are thinking as they do this task.

Since three of the 4 users had difficulty transferring a call and left the caller on hold, I think this transfer feature will definitely have to be changed and I'll recommend that to the designers.

But storing telephone numbers went quite well. There were few difficulties and some of the users even indicated they like the feature. So I don't think any changes are needed in the procedure for storing telephone numbers.

While showing narrator, he says:

The summary of data and recommendations for changes are passed onto other members of the design team; the engineers, system developers and industrial designers. The recommendations are not always understood or accepted by these members so any disagreements can be dealt with in a discussion group. Further, the data collected in the usability test is valuable in instances of disagreement because it provides evidence that changes are needed.

Act 9:

Two designers and the interface specialist are discussing the improvements required in the product.

Interface Specialist: That takes care of the conference feature, lets talk about the transfer feature now. The users tried putting the caller on hold when doing a transfer so we need to change this feature. I think the way to change it is so that pressing the transfer button will release the hold. That would make it seem as if they hadn't pressed hold in the first place and then they wouldn't get into trouble.

Designer 1: That's not a bad idea, but that means new functionality for the transfer button. And that means time and we just don't have the time in this project.

Interface Specialist: Yes, but 3 of 4 users had that difficulty so I'm convinced if a change isn't made, users won't be able to use transfer at all. So, its an important change to make.

Designer 2: Why don't we check our resources and see how much time it takes to do the programming. If it is feasible, then we'll do it, otherwise we'll just leave it for the next version.

Interface Specialist: But I think this problem has to take priority over some of the other problems we talked about and I really want to see it done in this version.

While showing chart of iterative design process, narrator says:

After the product is tested in a usability test and problems are identified, the product is redesigned, a prototype is developed with the new design, and it is user tested again. Hence, product design is an iterative process.

It continues until the user interface specialists and designers are confident they have a well-designed product.

While showing the chart of user input techniques with a check mark beside the field trial item, narrator says:

In some cases, the finished product is tested further in a field trial - the third type of user input technique.

While showing narrator, he says:

In a field trial, the product is implemented in an environment in which it would actually be used, such as a home or office, so a larger group of users can perform tasks with it on a regular basis. For example, a newly designed telephone may be distributed to a few office groups who use it daily. The trial usually runs for a few weeks to a few months. During the trial, the users are asked to answer questionnaires or participate in interviews. A small sample of questions that might be asked on the questionnaire is as follows:

While showing items in a questionnaire, narrator says:

Rate how easy the following telephone features were to learn.

The users would then rate on a 5-point scale how easy it was to learn each of the four features listed: call transfer, storing numbers, call forward, and conference.

The users would also be asked to write any comments about learning the features.

While showing narrator, he says:

The information collected in the questionnaires and interviews will allow the user interface specialists to assess satisfaction with the product in its implemented form.

In summary, as a member of the design team, the user interface specialist plays an important role in product design.

While showing chart of user input techniques, narrator says:

By employing user input techniques such as task analysis, usability testing, and field trials, the specialist ensures that the product meets the needs of the users.

While showing narrator, he says:

The involvement of users in product design will lead to success of the product in the market place. This will in turn increase the product sales and, hence, revenue for the company developing the product.

Appendix B.1

Task Script

In this session, you will be requested to conduct a number of tasks with a prototype of a microwave. The length of time the microwave cooks for in the tasks is generally short. This is to reduce the amount of waiting time in the exercises.

1. Set the current time.
2. Heat a cup of water for 10 seconds.

The microwave you are using comes equipped with a setting for determining the speed at which food is cooked. Speed is not associated with the length of time (e.g., 3 mins.) that food cooks, but is associated with how much the microwave is actually cooking in a given time period. For instance, for vegetables you would set the microwave to cook for 3 minutes at a fast speed. This means the oven is cooking for most or all of the 3 minutes. You could also melt butter for 3 minutes, but on a slow speed. This means the oven is cooking only a small portion of the 3 minutes, melting the butter slowly.

3. Have the microwave heat spaghetti sauce for 1 minute and 10 seconds at the fastest speed.
4. Heat a cup of milk for 15 seconds at half the speed.

The microwave you are using is equipped with a temperature probe. The probe is used to judge how well the food is done. The probe must be used when you want to set the temperature the food is to reach. For example, if you want to cook a casserole until it reaches 150°F, then the probe is inserted into the casserole and you set the temperature you want the food to reach. For tasks involving heating food to a particular temperature, insert the probe and take it out again when cooking is complete.

5. Warm a bowl of soup to 130°F using 70% of the maximum speed.
6. Heat a cup of hot chocolate to 140°F using 40% of the maximum speed.

7. Assume that at the last minute you decided to make hamburgers for dinner. Defrost 1 pound of hamburger.

When you cook food in the oven and the oven shuts off, the food begins to cool immediately. You can have the oven maintain the temperature of the food after it has finished cooking.

8. Have the microwave heat a cup of coffee to 110°F at 60% of the maximum speed and maintain this temperature of 110°F until you turn the microwave off.

You may set the microwave so that it automatically comes on at a later time. For instance, if it is 4:00 you can have the microwave come on at 5:00.

9. Have the microwave come on 1 minute from the current time to warm a muffin for 10 seconds.

The microwave may be programmed to cook in stages for different time lengths or temperature levels and at varying speeds.

10. Program the oven to cook a cold casserole to 120°F using the fastest speed and then cook to 140°F using only 30% of the maximum speed.

The oven comes equipped with a pre-programmed button such that it always cooks at the highest level of speed regardless of the length of time specified by you. This feature is used to have the oven come on without having to use the TIME and speed (POWER) buttons.

11. Without using the TIME and POWER buttons, have the microwave cook for 1 minute using the fastest speed.

2

of/de

2

PM-1 3½"x4" PHOTOGRAPHIC MICROCOPY TARGET
NBS 1010a ANSI/ISO #2 EQUIVALENT

PRECISIONSM RESOLUTION TARGETS

PRINTED IN METHYLLINE BLUE TESTING SINCE 1974

MicroD

3000 N. W. 10TH AVE. SUITE 100
FORT LAUDERDALE, FL 33309

Appendix B.2

Steps for Completing Tasks

Task 1. Set the current time.

- Step 1: touch TIME OF DAY
- Step 2: touch numeral pads to set time
- Step 3: touch TIME OF DAY

Task 2. Heat cup of water for 10 seconds.

- Step 1: touch TIME
- Step 2: touch 1 0
- Step 3: touch START

Task 3. Heat spaghetti sauce for 1 minute and 10 seconds at fastest speed.

- Step 1: touch TIME
- Step 2: touch 1 1 0
- Step 3: touch POWER
- Step 4: touch START

Task 4. Heat cup of milk for 15 seconds at half the speed.

- Step 1: touch TIME
- Step 2: touch 1 5
- Step 3: touch POWER
- Step 4: touch 5
- Step 5: touch START

Task 5. Warm soup to 130°F using 70% power.

- Step 1: insert probe
- Step 2: touch TEMP
- Step 3: touch 3
- Step 4: touch POWER
- Step 5: touch 7
- Step 6: touch START

Task 6. Heat hot chocolate to 140°F using 40% power.

- Step 1: insert probe
- Step 2: touch TEMP
- Step 3: touch 4
- Step 4: touch POWER
- Step 5: touch 4
- Step 6: touch START

Task 7. Defrost 1 pound of hamburger.

- Step 1: touch JET DEFROST
- Step 2: touch 1 0
- Step 3: touch START

Task 8. Heat coffee to 110°F at 60% power and hold warm at 110°F.

- Step 1: insert probe
- Step 2: touch TEMP
- Step 3: touch 1
- Step 4: touch POWER
- Step 5: touch 6
- Step 6: touch HOLD WARM
- Step 7: touch 1
- Step 8: touch START

Task 9. Delay heating a muffin for 10 seconds by 1 minute from current time.

- Step 1: touch DELAY
- Step 2: touch numbers for desired start time
- Step 3: touch DELAY
- Step 4: touch TIME
- Step 5: touch 1 0
- Step 6: touch START

Task 10. Cook casserole to 120°F at HI power and then at 140°F at 30% power.

- Step 1: insert probe
- Step 2: touch TEMP
- Step 3: touch 2
- Step 4: touch POWER
- Step 5: touch MEMORY
- Step 6: touch TEMP
- Step 7: touch 4
- Step 8: touch POWER
- Step 9: touch 3
- Step 10: touch START

Task 11. Cook for 1 minute using the fastest speed.

- Step 1: touch QUICK-ON
- Step 2: touch 1

Appendix C

Criteria Used to Rate Importance of Problems

To assess importance of fixing a problem consider the following factors:

- a) The consequences of the problem, i.e, were the users able to proceed with the task.
- b) Were the users dissatisfied with the design of the microwave? Dissatisfaction could be a function of either not liking the current design or wanting more functionality in the microwave.
- c) The frequency of the problem, i.e., the number of users who have the problem and/or are dissatisfied with the microwave.
- d) Was the problem general or task specific? A general problem is one that could affect overall use of the microwave and a task specific problem is one that pertains to a particular task.

Appendix D

Problems Presented to Observers

Note: Each list of problems on this page and the following page was presented to observers. One-half of the observers in each training condition received one of the lists. The other one-half of the observers in each training condition received the other list. The lists were written in point form but items were explained to the subjects. The problems were described as problems that the users may have.

Concerns of Designers

- Think the clock time has to be removed before using another feature.
- Do not know what the labels on the display mean (e.g., 1stStg).
- Use QUICK-ON for speed instead of POWER.
- Do not know how to set the power to a different level than HI%.
- Do not put the probe in when using temperature.
- Do not like it that entering temperature is different from the method for entering time.
- When using DEFROST, think the time has to be entered as well as the weight.
- Set the HOLD WARM temperature before setting POWER level.
- For DELAY, put 100 for 1 minute from the current time.
- Want the oven to stop between stages when doing multi-stage cooking.
- For QUICK-ON, try to put in 00's after entering the 1.
- Expect to touch START after entering the time for QUICK-ON.

Concerns of Designers

- Think the clock time has to be removed before using another feature.
- Want to see the power level during cooking.
- Initially expect to put two digits in when entering the power level (e.g., 4 and 0 for 40%).
- Do not know what COOL means.
- Try to set the POWER level before setting the TEMP or the TIME.
- Do not like it that for DEFROST entering one digit puts in only fraction of a pound (.1) and not one pound.
- Use MEMORY instead of HOLD WARM to set the warming temperature.
- Set HOLD WARM temperature before setting POWER level.
- Do not know to touch the DELAY button after entering the delay time.
- Do not know to use the MEMORY button for multi-stage cooking.
- Do not know where to use the MEMORY button for setting stages.
- Expect to touch START after entering the time for QUICK-ON.

Appendix E.1

Coding Form

Coding Form Instructions

Following are guidelines for completing the Coding form.

- For each task, write the problems the users have with the microwave or write the steps the user does to perform a task. Problems may be identified by actions or by comments.
- Try to record as many of the problems or steps that you can. If you think you may have missed something, record what you can but do not rewind the videotape.
- Stop the videotape between each task to review the steps of the next upcoming task. Do not stop the videotape during a task.

Task 1. Set the current time.

Step 1: touch TIME OF DAY

Step 2: touch numeral pads to set time

Step 3: touch TIME OF DAY

[illegible]

Task 1. Set the current time. (CONT.)

Step 1: touch TIME OF DAY
Step 2: touch numeral pads to set time
Step 3: touch TIME OF DAY

This image shows a single sheet of white paper with horizontal ruling lines. The lines are evenly spaced and run across the width of the page. There are no margins, text, or other markings on the paper.

Sample Coding Form

1. Forward calls to extension 4032.

- Step 1: lift the receiver
 Step 2: press Forward (F)
 Step 3: dial the four digit-number to which the calls should be routed
 Step 4: hang up

_hesitate_____
 _move cursor to Transfer, then numbers (don't press anything)_____
 _transfer_____
 _4032_____
 _4032_____
 _transfer_____
 _see Forward button_____
 _used Forward, then #'s_____
 _Not work because not pick up receiver_____
 _try line before F_____
 _try release at end_____
 _"Wonder if pick up receiver. Don't understand."_____
 _Should work when down."_____
 _lift receiver_____
 _restart - get it_____

Appendix E.2

User Summary Form

User Summary Form Instructions

Following are guidelines for completing the User Summary form.

- Summarize the problems the user had for each task. Be explicit and concise in your descriptions of the problems.
- If difficulties are a result of a greater more general problem then identify these difficulties as sub-problems. List the sub-problems with the more general problem.

Task 1. Set the current time.**Problems**

1. _____

2. _____

3. _____

4. _____

5. _____

6. _____

7. _____

First Sample User Summary Form**Task 1. Forward calls to extension 4032.**Problems

1. ___ Not know how to start sequence - hesitated_____
___ - tried Transfer then #'s_____
___ - tried #'s then Transfer_____
2. ___ Button confusion - used Transfer instead of Forward_____

3. ___ Did not know to lift receiver to Forward call_____
___ a) - tried line before Forward_____
___ b) - tried using Release button after entering the numbers_____
4. ___ Did not like it that have to pick up receiver. Think it should work____
___ when Receiver is down._____

5. _____

6. _____

Second Sample User Summary Form**Task 1. Forward calls to extension 4032.****Problems**

1. ___ Did not know had to lift receiver to Forward call. _____
___ a) - tried line before Forward _____

2. ___ Sequence error - touched numbers, then Forward. _____

3. ___ Touched two numbers at the same time when entering 4032. _____

4. _____

5. _____

6. _____

Third Sample User Summary Form**Task 1. Forward calls to extension 4032.**Problems

1. ___ I'd not know had to lift receiver to Forward call - realized quickly. ___

2. ___ Button confusion - used Pick-Up instead of Forward. Corrected self
___ quickly. Thought could have the call "Picked Up" elsewhere. _____

3. _____

4. _____

5. _____

6. _____

Appendix E.3

General Summary Form

General Summary Form Instructions

Following are guidelines for completing the General Summary form.

- For each task summarize all of the problems you have identified for the 3 users. In the summary of each item, include:
 - a) the number of users who had the problem, and
 - b) what the problem was.

If there are sub-problems then list these under the general problem. Be explicit and concise in your descriptions of the problems.
- Rate the importance of fixing each of the problems. To assess importance of fixing a problem consider the following factors:
 - a) The consequences of the problem, i.e., were the users able to proceed with the task.
 - b) Were the users dissatisfied with the design of the microwave? Dissatisfaction could be a function of either not liking the current design or wanting more functionality in the microwave.
 - c) The frequency of the problem, i.e., the number of users who have the problem and/or are dissatisfied with the microwave.
 - d) Was the problem general or task specific? A general problem is one that could affect overall use of the microwave and a task specific problem is one that pertains to a particular task.

The ratings are made on a 5 point scale ranging from 1-5. The scale is presented below.

Not Important to Fix		Moderately Important to Fix		Very Important to Fix
1	2	3	4	5

A rating of 2 should be given if a problem is not important to moderately important to fix and a rating of 4 if it is moderately to very important to fix.

Task 1. Set the current time.

# Users		Not Important to Fix	2	Moderately Important to Fix	4	Very Important to Fix
1.	_____	1	2	3	4	5

2.	_____	1	2	3	4	5

3.	_____	1	2	3	4	5

4.	_____	1	2	3	4	5

Task 1. Set the current time. (CONT.)

# Users		Not Important to Fix		Moderately Important to Fix		Very Important to Fix
_____ 5.	_____	1	2	3	4	5

_____ 6.	_____	1	2	3	4	5

_____ 7.	_____	1	2	3	4	5

_____ 8.	_____	1	2	3	4	5

Sample General Summary Form

Task 1. Forward calls to extension 4032.

# Users		Not Important to Fix		Moderately Important to Fix		Very Important to Fix
<u>2</u>	1. _Not know how to start sequence._____	1	2	3	4	5
	_ - 1 user hesitated - tried Transfer then _					
	_ #'s and #'s, then Transfer_____					
	_ - 1 user used #'s, then Forward_____					

<u>2</u>	2. _Button confusion. Did not know to use_	1	2	3	4	5
	_Forward._____					
	_ - 1 user used Transfer._____					
	_ - 1 user used Pick-Up - corrected quickly.					

<u>3</u>	3. _Not know have to pick up receiver._____	1	2	3	4	5
	_ - 2 users tried number of things, e.g.,__					
	_a) touched Release after entering ext.__					
	_b) touched Line before Forward_____					
	_ - 1 user corrected self quickly_____					
<u>1</u>	4. _Not like it that have to pick up_____	1	2	3	4	5
	_receiver. Thinks should work when____					
	_down._____					

Task 1. Forward calls to extension 4032. (CONT.)

# Users		Not Important to Fix		Moderately Important to Fix		Very Important to Fix
<u>1</u>	5. Touched 2 numbers at same time _____ _____when entering 4-digit extension. _____ _____ _____ _____	1	2	3	4	5
_____	6. _____ _____ _____ _____ _____	1	2	3	4	5
_____	7. _____ _____ _____ _____ _____	1	2	3	4	5
_____	8. _____ _____ _____ _____ _____	1	2	3	4	5

Appendix E.4

Rating Summary Form

Note: Sample pages from the first two tasks are included.

- Rate the importance of fixing each of the problems listed. To reiterate, to assess importance of fixing a problem consider the following factors:
 - a) The consequences of the problem, i.e., were the users able to proceed with the task.
 - b) Were the users dissatisfied with the design of the microwave? Dissatisfaction could be a function of either not liking the current design or wanting more functionality in the microwave.
 - c) The frequency of the problem, i.e., the number of users who have the problem and/or are dissatisfied with the microwave.
 - d) Was the problem general or task specific? A general problem is one that could affect overall use of the microwave and a task specific problem is one that pertains to a particular task.
- The ratings are made on the same 5-point rating scale as you used in the form you previously completed. The scale is shown below.

Not Important to Fix		Moderately Important to Fix		Very Important to Fix
1	2	3	4	5

- If you are unclear about what any of the problems are then ask for assistance.

Task 1. Set the current time.

# Users		Not Important to Fix		Moderately Important to Fix		Very Important to Fix
<u>3</u>	1. _Confused by names TIME OF DAY____ _and TIME. _____ _Did not know which of these buttons____ _to use to set clock time. _____	1	2	3	4	5
<u>3</u>	2. _Did not know TIME OF DAY is correct _button because of lack of feedback _____ _after touching it (e.g., 0 on display _____ _not change). _____	1	2	3	4	5
<u>3</u>	3. _Did not know what to do after enter ____ _numbers for clock time. _____ _Users touched START instead of _____ _TIME OF DAY. _____	1	2	3	4	5
<u>1</u>	4. _Did not know if clock was 12 hour or ____ _24 hour clock. _____	1	2	3	4	5
<u>2</u>	5. _Did not know what information to put_ _ in because no colon shown in display._ _One user tried to enter secs. Other user _thought clock not working when _____ _entered time based on 24-hour clock. ____	1	2	3	4	5
<u>1</u>	6. _Sequence error. Touched numbers ____ _before touching TIME OF DAY. _____	1	2	3	4	5

Task 2. Heat water for 10 seconds.

# Users		Not Important to Fix		Moderately Important to Fix		Very Important to Fix
<u>1</u>	1. _Poor feedback on display._____	1	2	3	4	5
	_Not sure if 10 = 10 secs. _____					
<u>1</u>	2. _Did not know what 1stStg means._____	1	2	3	4	5
<u>1</u>	3. _Did not know how to start sequence to____	1	2	3	4	5
	_ set time. Put 10 for 10 secs. without_____					
	_touching TIME first._____					
<u>1</u>	4. _Did not know how to end sequence_____	1	2	3	4	5
	_ when setting time. _____					
	_Thought oven start after entering 10.____					
<u>1</u>	5. _Pressed TIME before touching START_	1	2	3	4	5
	_(i.e., used TIME, 10, TIME, START).____					
	_Did this because clock was set by_____					
	_using TIME OF DAY twice._____					

Appendix F

Microwave Manual

The Display and Oven Controls

DISPLAY

- Shows the current clock time, the cooking time or the defrost time.
 - When the oven controls are selected a corresponding name is highlighted in the display.
- The oven control names and the corresponding display names are as follows:

<u>Oven Control Name</u>	<u>Display Name</u>
Delay	Delay
Hold Warm	HldWrm
Quick-On	Auto
Jet Defrost	JetDf
time or temperature	1stStg
time or temperature	2ndStg

NUMERAL PADS (1, 2, 3,....9, 0)

- Use to enter the numbers when setting the time, power level, temperature, etc.

CLEAR

- Touch to cancel the selected power level, cook or defrost times or temperatures before touching the START button.

START

- Touch to start the oven cooking or defrosting.

Note: At end of operation, three beeps will sound, and the time of day will return to the display window.

STOP

- Use to interrupt operation of oven while cooking or defrosting.
- To continue cooking touch the START pad, and the oven will restart. If no further cooking is required touch the CLEAR pad.

CLOCK

- Use to set the time of day.
- Time of day must be set before any other oven controls are used.
- Time of day will be displayed at all times except when the oven is operating.

POWER

- Use to set the speed at which the food is cooked. The default is the fastest speed (HI%).

TIME

- Use to set the cooking time.

TEMP

- Use to set the desired food temperature. To use temperature, the probe must be used and inserted into the food.
- The desired temperature can be set in 10°F increments between 100°F and 200°F. The internal temperature of the food will be shown in 5°F increments in the display until it reaches the preset temperature.

HOLD WARM

- Use to hold the internal temperature of food after cooking.
- HOLD WARM may only be used with the temperature probe and after the oven has been set to cook to a desired temperature using the TEMP feature. The HOLD WARM temperature and the cooking temperature are specified separately.
- The oven will maintain the HOLD WARM temperature selected until the oven is shut off.

MEMORY

- Use to set a first and second cooking stage.
- Stages are used to have the oven cook food for varying lengths of time or to different temperature levels using different levels of power.

DELAY

- Use to set the oven to come on at a later time.
- The DELAY time must be set before the cooking time or the temperature level is specified.

JET DEFROST

- Use to defrost frozen foods, by weight.
- The defrost cycle starts at medium-high power, then automatically reduces to medium-low and then to low as the food thaws.

QUICK-ON

- Use to turn the oven on quickly and cook foods at the highest power level.
- The oven can be set at any number of minutes to a maximum of 9 minutes.

Using the Oven Controls

Setting the clock

- Touch TIME OF DAY
- Touch numbers for correct time of day (e.g., 615 for 15 minutes after 6 o'clock)
- Touch TIME OF DAY

Time cooking

- Touch TIME
- Touch numbers for desired cooking time (e.g., 300 for 3 minutes)
- Touch POWER
- Touch number for desired cooking power level (e.g., 6 for 60%)
- Touch START

Temperature cooking

- Insert probe
- Touch TEMP
- Using the chart below, touch number for desired temperature (e.g., 5 for 150°F)
- Touch POWER
- Touch number for desired cooking power level (e.g., 6 for 60%)
- Touch START

<u>Number</u>	<u>Temperature</u>
1	110°F
2	120°F
3	130°F
4	140°F
5	150°F
6	160°F
7	170°F
8	180°F
9	190°F
0	200°F

Heat and hold warm

- Insert probe
- Touch TEMP
- Touch number for desired temperature (e.g., 3 for 130°F)
- Touch POWER
- Touch number for desired cooking power level (e.g., 2 for 20%)
- Touch HOLD WARM
- Touch number for desired temperature to hold warm at (e.g., 3 for 130°F)
- Touch START

Multi-Stage cooking

- Touch TIME or TEMP (with TEMP probe)
- Touch number(s) for desired cooking time or temperature
- Touch POWER
- Touch number for desired cooking power level
- Touch MEMORY
- Repeat steps for setting TIME or TEMP and POWER to set second cooking stage. Do not touch MEMORY again.
- Touch START

Delay cooking

- Touch DELAY
- Touch numbers for desired start time (e.g., 500 for 5 o'clock)
- Touch DELAY
- Touch TIME or TEMP (with TEMP probe)
- Touch number(s) for desired cooking time or temperature
- Touch POWER
- Touch number for desired cooking power level
- Touch START

Defrosting

- Touch JET DEFROST
- Touch numbers for weight of food (e.g., 5 for one-half pound)
- Touch START

Automatic cooking

- Touch QUICK-ON
- Touch number for desired minutes to cook (e.g., 3 for 3 minutes)

Appendix G.1

Task and Training questionnaire

Note: Text in italics was presented to the subjects in the extended training group who saw the videotape.

The following questions are designed to assess your opinion of the observational task you performed and the training you received to do this task.

Rate the observational task (i.e., observing, recording, summarizing and rating the importance of problems) on the following items.

Uninteresting	1	2	3	4	5	6	7	Interesting
Simple	1	2	3	4	5	6	7	Complex
Intensive	1	2	3	4	5	6	7	Unintensive
Stressful	1	2	3	4	5	6	7	Not Stressful
Easy	1	2	3	4	5	6	7	Difficult
Uncomplicated	1	2	3	4	5	6	7	Complicated
Demanding	1	2	3	4	5	6	7	Effortless
Straight forward	1	2	3	4	5	6	7	Confusing

In this study you received training for performing your task. The training included instructions on what usability testing is, a videotape on *User Interface Specialists*, a demonstration of how to use the microwave, examples of the types of problems users may have with the microwave, and oral and written instructions with examples on how to record and summarize data. Think about the instructions you received and then complete the questions on this page and the following 3 pages.

Rate the training you received on the following items.

Simple	1	2	3	4	5	6	7	Complex
Intensive	1	2	3	4	5	6	7	Unintensive
Stressful	1	2	3	4	5	6	7	Not Stressful
Easy	1	2	3	4	5	6	7	Difficult
Uncomplicated	1	2	3	4	5	6	7	Complicated
Demanding	1	2	3	4	5	6	7	Effortless
Suitable	1	2	3	4	5	6	7	Unsuitable
Helpful	1	2	3	4	5	6	7	Unhelpful
Inadequate	1	2	3	4	5	6	7	Adequate
Beneficial	1	2	3	4	5	6	7	Useless
Insufficient	1	2	3	4	5	6	7	Sufficient
Appropriate	1	2	3	4	5	6	7	Inappropriate
Ineffective	1	2	3	4	5	6	7	Effective
Disorganized	1	2	3	4	5	6	7	Organized
Straight forward	1	2	3	4	5	6	7	Confusing
Muddled	1	2	3	4	5	6	7	Clear
Unnecessary	1	2	3	4	5	6	7	Necessary
Important	1	2	3	4	5	6	7	Unimportant

Rate the training on the following items.

Too long			Appropriate length			Too short
1	2	3	4	5	6	7
Too slow			Appropriate speed			Too quick
1	2	3	4	5	6	7
Too detailed			Appropriate level of detail			Not enough detail
1	2	3	4	5	6	7
Too few examples			Appropriate # of examples			Too many examples
1	2	3	4	5	6	7

Indicate the extent to which the training was useful to you for the following.

	Not useful						Very useful
1. Learning about the role of usability testing in product design?	1	2	3	4	5	6	7
2. Learning about the observer's role in product design?	1	2	3	4	5	6	7
3. Learning how to observe and record user performance?	1	2	3	4	5	6	7

Did you learn something from the training that you did not know already?

Learned nothing new						Learned a lot
1	2	3	4	5	6	7

What information, if any, do you think could have been omitted from the training for learning about and conducting the observational task? Check in the right-hand column those items below which you think could have been omitted.

	Could be omitted
The introduction to what usability testing is.	-----
<i>The videotape entitled "The Role of User Interface Specialists in Product Design".</i>	-----
The presentation on how to use the microwave.	-----
The presentation and list of problems which may occur in the videotapes.	-----
The oral instructions for completing the coding forms and summary forms.	-----
The written instructions for completing the coding forms and summary forms.	-----
The sample coding and summary forms.	-----

Other (please specify):

What information, if any, do you think should be included in the training for learning about and conducting the observational task.

Overall, how would you rate the quality of the training.

Poor
quality

Excellent
quality

1

2

3

4

5

6

7

Additional comments about the training.

Appendix G.2

Training Videotape questionnaire

Now that you have answered questions about the training in general, I would like you to answer the following questions on only the training videotape, i.e., the videotape entitled "The Role of User Interface Specialists in Product Design". Think about the videotape and then answer the questions on this page and the following page.

Indicate on the items below the extent to which the training videotape was of assistance to you in learning about and carrying out your observational task (i.e., observing, recording, summarizing and rating the importance of problems).

Suitable	1	2	3	4	5	6	7	Unsuitable
Unhelpful	1	2	3	4	5	6	7	Helpful
Useless	1	2	3	4	5	6	7	Beneficial
Appropriate	1	2	3	4	5	6	7	Inappropriate
Ineffective	1	2	3	4	5	6	7	Effective

Rate the presentation of the training videotape on the following items.

Organized	1	2	3	4	5	6	7	Disorganized
Straight forward	1	2	3	4	5	6	7	Confusing
Muddled	1	2	3	4	5	6	7	Clear

Indicate the extent to which the training videotape was useful to you for the following.

	Not useful					Very useful	
1. Learning about the role of usability testing in product design?	1	2	3	4	5	6	7
2. Learning about the observer's role in product design?	1	2	3	4	5	6	7
3. Learning how to observe and record user performance?	1	2	3	4	5	6	7

Did you learn something from the training videotape that you did not know already?

Learned
nothing new

Learned
a lot

1 2 3 4 5 6 7

Additional comments about the training videotape.

Appendix G.3
Personal Data Questionnaire

Gender: _____ Age: _____

Education: Undergraduate _____

Graduate _____

Major area of study: _____

Microwave usage

Do you use a microwave on a regular basis (at least once a week)?

YES _____ NO _____

If yes, does the model of microwave you use have most of the features you saw on the microwave in this study, some of the features or only a few of the features? Note: Answer this question on the basis of the function of the features and not on the basis of the names of the features.

Most features _____

Some features _____

Few features _____

Product Design Experience

Have you ever been involved in the design of a product?

YES _____ NO _____

If yes, approximately how many products have you helped to design? _____

For each instance in which you helped to design a product, briefly describe the product and your role in the design (e.g., product designer, prototype developer, coder, system tester, user test designer, usability tester, user)?

Formal Observation Knowledge and Experience

Have you ever received any instruction on techniques for conducting formal observation?

YES _____ NO _____

Have you ever received any instruction on the goals of usability testing and/or the stages involved in usability testing?

YES _____ NO _____

Have you ever participated as an observer in a formal situation, including psychology experiments?

YES _____ NO _____

If yes, approximately how many experiments or observational situations have you been involved in as an observer? _____

For each instance in which you have acted as an observer, briefly describe the conditions under which you have done this (e.g., type of study, who did you observe, what did you observe, how long did you observe)?

Appendix H

Verbal Instructions to Observers

a) Introduction to the task

In this study, you will be doing some observations in a usability test. Before you do this, I want to tell you a bit about what usability testing is. If you have any questions along the way, then please ask. Also, when I go through the instructions, I want you to assume you have been hired to do a usability test. Think of this as a real situation where I'm training you to do the test and then you have to actually go ahead and do it, produce some results and pass these on to other members of the company you're working for.

A usability test involves having users perform tasks on a product and an observer who is watching the users use the product. When I say a product, I'm referring to products such as televisions, telephones, VCR's, computers, or any specialized systems such as Air Traffic Control Systems. A task the users might be required to do in a usability test could be program a VCR to tape a movie at 9:00 p.m.. The observer would watch the users perform this task and record the performance of the users. You'll be the observer in a usability test and you'll be watching users perform tasks on a microwave and so you'll be recording the performance of the users that you watch.

The reason you're going to observe users is so that you can identify what types of problems or difficulties the users have with the microwave. By doing this, you'll be able to identify how the design of the product needs to be improved so that its more suited to the needs of the users. The information you collect in the usability test would be used to make decisions about how to best improve the product before it is released to the consumers.

So that's what usability testing is about. In this study, you will watch 3 users perform tasks on a microwave. The users aren't going to come in here, but rather you'll watch them on videotape. Further, they won't be performing tasks on a real microwave. They'll do their tasks on a prototype of a microwave. A prototype is simply a mock-up of a product which simulates what the real product will do. I'll show you what I mean. (*show alarm clock prototype on Macintosh*). This prototype which is an alarm clock allows users to perform tasks as though this were a real clock. I'll operate the mouse as though I'm a user and you can watch me do a few things. As a user I would use the mouse to move the hand pointer and touch buttons on the clock. (*touch a few buttons*) This is just a simple prototype but its just to show you that you can use prototypes to test products and they are especially useful for testing products early in the design of a product.

So you are going to be watching 3 videotapes of users using a prototype of a microwave. You'll watch them to identify the problems the users have, and things they dislike about the prototype of the microwave. I will be giving you specific instructions on how to record the behaviours you observe and how to summarize the data you collect.

b) Viewing the training videotape (extended training group only)

Before I give you instructions on observing, I would like you to watch this videotape which describes the role of user interface specialists and techniques they use for assessing user needs. One of the techniques is usability testing. The videotape is about 20 minutes long. I simply want you to watch and listen to the videotape.

c) Learning the microwave features

Now you have to learn how to use the microwave oven used in the videotapes you will be viewing. It is necessary for you to learn the features of the microwave so that when you observe users, you are able to record their performance and identify where they have difficulty. You don't have to be an expert on the microwave because you will have aids with you when you record to let you know how each feature is used. One of the aids is a manual (*show manual to observer*) which you may consult at any time. In the manual is a description of each of the oven controls and how to use them.

Before you learn the microwave, I would like to explain some things about the prototype of the microwave. (*show prototype of microwave*) The display and oven control buttons are shown on the right-hand side of the screen. The users press the buttons to perform the tasks and the display shows the numbers pressed or shows information about the buttons pressed. The user uses the mouse to select the buttons they want to press. On the left-hand side of the screen are two pictures, one of a miniature microwave and one of a probe placed beside a dish. A probe is an instrument which is put inside your food to measure how well done the food is, i.e., to measure the internal temperature of the food. When I click with the mouse on the box beside the words "Insert Probe", an X will be placed in the box and the picture of the dish will change so that the probe is inside the dish. To take the probe out, I select the box again and the X disappears and the picture changes back to its original state. The users had to use the probe for some of the tasks they completed. The picture of the microwave is used to show an open and closed door. If I press on the oven control button "Open Door", the picture of the microwave will change so that the door is open. Then when I press on the box beside the words "Close Door" the picture will change back to a closed-door microwave. The users weren't asked to open and close the door between tasks. Also, as you've probably noticed, anytime I press with the mouse, there is an audible sound and the

cursor changes from a hand to an arrow. These cues will help you as an observer to know when the users perform some action.

Now, you can learn how the microwave works. *(show task script)* To learn the oven controls, I will first read each of the tasks that the users completed in the videotapes that you will watch. *(show list of correct steps)* Then I'll read through this other list which gives the correct steps to complete each of the tasks. Finally, using this list of steps, you'll complete each of the tasks on the prototype. I'll have you press the buttons so you can learn how the microwave works.

I should point out that users aren't given the list of correct steps for doing the tasks. The reason is you want to see how well they can do the tasks without being told how to do them. This will give you an indication of how easy it is to use the microwave. You will have this information in this list when you record the data so you don't have to memorize it.

To start, we'll go through the first task, then I'll show you the coding forms you will use when you're observing the users, and then we'll come back and go through the rest of the tasks. While you're learning the microwave, keep in mind that this is the way in which the microwave has been designed and so this is the way in which the prototype is designed. *(go through first task with observer performing task on microwave)*

d) Learning the coding form

Now I'll show you the forms you'll complete when you observe users. Here is the coding form you will be using for recording data for the three videotapes. Here is a sample form and here is a list of points which summarizes my explanation. You can refer back to this list at any time. You'll be using the coding form for writing down the difficulties the users have while performing each task. There are 2 sheets per task. Use the second one if you run out of room on the first page. The steps the user was to take are written at the top of each form. I want you to write data for each task on the lines below.

The sample form is from a usability test of a telephone and the task is to forward calls to extension 4032. Below this are the correct steps the user was supposed to perform. Further below this is what the user actually did. *(read through what is written on the form)*. There are 2 ways to record the data. First, you can write everything the user does or says. If you write everything down, you will have a chance later to use the microwave manual or look at the steps at the top to determine what the problems were. The second approach is to summarize the problems or difficulties as they occur. You can use either method or a combination of both -- whatever you need to record in order to identify the problems or difficulties users have with the product.

I'll go through the form again and this time I'll point out some of the types of problems that might occur. Problems could be a result of doing steps in the incorrect order or not knowing how to start a sequence (e.g., pressing numbers and then Transfer) using the wrong buttons (e.g., pressing Transfer instead of Forward), omitting steps (e.g., Not picking up the receiver), or using additional buttons (e.g., using Line before Forward), or it could be hesitations or comments about the design of the product. For example, the comment in the sample "Don't understand why have to pick up receiver. Should work when down" indicates the user doesn't like the way the feature is designed. This is a separate problem from the one above where the user did not know to lift the receiver because in the earlier problem they didn't know what to do while this comment indicates they don't like the way they have to use the feature.

Everything the user does or says is listed in chronological order and every data point is written on a separate line. You'll find you'll develop your own method for recording the data. You might want to use X's or checkmarks or underlining or indenting to indicate any problems. Further, you can use abbreviations such as the one written on this sample, F for Forward. If you use abbreviations indicate what the abbreviation means at the top of the page so I know what everything means on the form.

So that's what you would record. You'll be recording problems for each of the 11 tasks each user performs in the 3 videotapes. Before I go back to teaching you the microwave, I want to show you what you'll do with the data you collect. If the data is going to be of any value, then the data must be summarized. An observer is typically writing data in a hurry as they watch a user. Also, you cannot pass your data on to other members of your company in this format so the information has to be gathered together. After you view the first videotape and observe the user's performance, I'll have you summarize the data for that user. I'll show you briefly how you do that but give you more instructions later. Here is the summary form you'll use and here is a sample summary form. You'll complete this form for each of the 3 users for each task. You just have to summarize all of the problems, dislikes, or difficulties you identify when observing the videotape. So looking at the example of forwarding a call, the summary may be as follows (*go through points on summary*). I should also mention that sometimes users will have problems which occur as a result of another problem. For instance, as shown on the summary form, in using the telephone the user had problems of using Line before Call Forward and trying Release at the end of the sequence. These problems actually occurred because the user did not lift the receiver, the user was trying a number of sequences. You'll likely come across these when observing and so should keep it in mind. I just want you to realize that these are not individual problems but are actually sub-problems of a much larger problem.

Finally, after you view each videotape and summarize the data for each user, I'll have you summarize data for all 3 users combined. Again this is important to do because it is necessary to put the data in a meaningful

package and to look for similarities and differences in how the users performed, i.e., look for patterns in the data. The form you'll use for that is the General Summary form. You'll summarize the problem and then you'll rate the importance of fixing each problem. I'll be giving you criteria for how to make your ratings. *(read through 2 summaries on the sample form)*

I'll give you more specific instructions later on the summaries but I thought I should tell you about them now to optimize how you record data as you view the videotapes.

e) Inform observers of expected problems

Before you view the first videotape, we have to go through the rest of the tasks the users performed to show you how the microwave is used. I'll also point out some problems that the designers expressed some concern about. Here is a list of those problems. *(show list)* I want you to watch for these problems when you're viewing the videotapes and record them if you see them occur. Also, there will inevitably be other problems so record them as well. I'm just pointing out ones that some of the designers mentioned they thought might happen. *(proceed through rest of tasks with observer performing them on the microwave prototype and point out the problems in the list at an appropriate time)*

f) Viewing the first videotape

Now you get to view the first videotape. In the videotape there will be a user using the prototype microwave which you have learned about. You will not see the user, but will see the computer screen with the prototype and see the pointer which the user uses to select buttons. You will also hear the beeps when the user touches anything. Throughout the videotape, you will hear the users voice as they perform the tasks. The reason for this is that in a usability test, users are asked to verbalize their thoughts. This will help you as an observer to identify the problems users have. Also, on occasion, you will hear the voice of an assistant. The assistant was available in case users had difficulty with a task and needed some help.

When you are watching and recording, I want you to stop the videotape between tasks. The reason for this is so that you can review the steps of the next upcoming task. *(show how observer is to do this)* You'll know the task is done because the user reads their task before starting the next one. For example, for the first task, they'll say "My first task is", and for the second one say "My second task is" and then read the task. Once you have reviewed the steps of the next upcoming task, you can start the videotape again. Do you have any questions at this point?

I realize this is going to seem difficult at first, but I want you to try the best that you can. Remember, you can either record only the problems you see

or try to write everything down the user does -- whatever you need to record to be able to identify the problems. You'll find that over the 3 tapes, you'll develop your own method for how to record the data. If you're recording the users steps, you might want to use techniques to highlight problems such as marking problems with X's or underlining.

If you miss something, don't worry about it just go ahead with watching the tape and try to get what you can. I ask that you do not stop the tape during a task and that you do not rewind. This is important for two reasons. First, if you were watching the user live, you wouldn't be able to have them redo their steps. The second reason I don't want you to stop the videotape is for the purpose of my study. I'm interested in watching performance from one videotape to another so I need all of my subjects to have the same amount of viewing time. So, the main things to remember are to stop the videotape between each task. Remember that your goal is to identify the problems users are having no matter how minor the problem may seem. Also, try to be as accurate as possible when you are recording data. Do you have any questions?

g) Summarizing data for first videotape

I would now like you to summarize the data for the first user. As I mentioned earlier this is important to do so that your observations are put into a meaningful package. Here is the form I told you about earlier. Here is the sample summary form (*show form with telephone data summarized for one task*) and a brief list of instructions which you can refer back to at any time. I'll read through the sample again even though we went through it earlier. (*explain how to summarize the problems and remind subjects of sub-problems*).

When summarizing be sure to stick to the data. Even though you may be aware of problems with the microwave that are not encountered by the users, just record what happens with the users.

There is one form per task with enough room for 7 problems. If you find more than 7 then write them on the back of the page. Do you have any questions?

h) Viewing 2nd and 3rd videotape

Now I'll have you watch the 2nd (3rd) videotape. Just do the same thing as you did for the last coding form. (*give forms to observer*) Again, watch the videotape and record the users performance. Do you have any questions? (*answer questions, if any*)

i) Summarizing 2nd and 3rd videotape

Now you have to summarize the problems for this videotape. (*give forms to observer*) Again, summarize the data for each task the user conducted. Do you have any questions?

j) Writing the general summary

Now that you have viewed the three videotapes, I would like you to summarize the data for all 3 users combined. Again this is important because it is necessary to put the data together for all users combined and to look for patterns and identify which problems are important to fix. Lets go over the sample form. First, the number of users must be written, then the problem experienced by users along with any sub-problems, and the rating for the importance of fixing each problem. Going over the problems, the first point is that two users were confused about which button to use, one used Transfer, and another used Pick-Up. Although users did different things both users had the same problem of not knowing which button to use. Hence, these two problems would go together as one problem. When you join the problems of two or more users, then I'd like you to specify below the problem, what each user did. Also, make sure the problems really belong together as one problem. (*proceed through the rest of the summaries on the sample form*) So everything you have on the individual summaries should appear somewhere on the General Summary form.

When summarizing be sure to stick to the data. Even though you may be aware of problems with the microwave that are not encountered by the users, just record what happens with the users.

On your form, there is enough room for 8 problems per task. If you run out of room continue on the back of the page.

To do the ratings of importance, I want you to consider a number of factors. (*show the list of factors and read through them with the observer*) You're making your ratings from the perspective of the users and not how difficult the problem is to fix. Also, don't use the ratings on the sample form as a guide for how you should make your ratings. Also, summarize all of the data first and then do the ratings. This is a better approach because you may find a problem occur several times throughout the videotapes. This may affect how you rate certain problems. Do you have any questions?

k) Completing the Rating Summary form

Now I would like you to rate each of the items which have been summarized in this form (*give form to observer*). This has nothing to do with usability testing, but in this form are all of the problems that occurred in the videotapes. These problems were identified by two usability test experts who also watched the videotapes and recorded the problems they

saw. I would like you to rate these items as you did before in terms of how important they are to fix. I'm having you do this so I can obtain your rating of importance on all problems including ones you may have missed when you summarized the data. You'll use the same rating scale and same criteria as before. *(give list of criteria)*

l) Completing the Training questionnaire

(read through the instructions at the top of the questionnaire and expand on the various types of training observers received)

m) Completing the Training Videotape questionnaire (extended training group only)

(read through the instructions at the top of the questionnaire)

n) Completing the Background questionnaire

(give questionnaire and briefly explain)

o) Debriefing

(give debriefing to subject to read and answer any questions)

Appendix I**Informed Consent Form****OBSERVATIONAL STUDY**

The purpose of the study you are about to participate in is to assess performance at observing and recording data in a usability test.

You will first receive training for learning how to observe and record data. You will then record data for three videotapes which you will watch, and summarize the data you have collected. At the end of the session, you will be asked to complete some questionnaires.

Your participation in this study is completely voluntary. You have the option to stop participating in the experiment at any time.

The data collected in this experiment will be used for statistical purposes only. Your responses on any of the materials you complete today will be kept confidential. Your name will not be identified on the materials.

Once you have read this form and agree with the terms of the experiment, please sign below.

NAME (print): _____

SIGNATURE: _____

DATE: _____

Appendix J

Debriefing

In this study I am assessing observer accuracy and reliability in a usability test. I asked you to record all the problems you saw occur in the user tests and then rate the importance of fixing each of those problems. To assess accuracy, I will compare your recordings of problems and ratings of importance to the recordings and ratings of two experts. To measure reliability, I will compare your recordings and ratings to those of other subjects. Determining performance accuracy and reliability will indicate how well observers can record data in a usability test.

In addition to measuring accuracy and reliability, I am also examining whether or not observer performance is affected by training, obviousness of the problems, and prior knowledge of problems that are predicted to occur in the user test. To assess the impact of training on performance, half of the subjects view a training videotape. The videotape teaches subjects about the importance of improving product design and the use of techniques such as usability testing to do this. I will compare performance of subjects who received the videotape to those who did not.

To examine if obviousness of problems affects performance, the problems that occurred in the videotapes were rated as high, medium, or low obvious by two experts in the area of product design and usability testing. Obviousness was rated on the basis of whether or not the problem was easy to see and whether or not observers would recognize the problem as a usability problem and record it. Based on the expert ratings, I will examine if subjects performed better on the high obvious problems than on the low obvious problems.

Finally, to examine if prior knowledge of problems influences performance, all subjects have been informed of some problems that are "expected" to occur in the videotapes. The problems were actually taken directly from the videotapes and were known to occur. Some subjects received the same predicted problems as you, and others received different problems. I will assess whether or not performance is better on those problems that were predicted than those that were not predicted.

In summary, I am measuring observer accuracy and reliability at recording data and identifying importance of problems in a usability test. I will assess the impact of training, obviousness of the problems, and prior knowledge of the problems on performance. Research in this area has not been done before and will be useful for determining an appropriate procedure to train observers to record data in a usability test.

Appendix K

Analyses on the Videotape Summaries

Analyses for the 54 problems in the general summary are presented in the Results section. The following analyses examined the effects of the independent variables on observer performance in the videotape summaries.

Problem Identification

Initial analyses examined the effects of obviousness, prior knowledge, and training on observer performance at identifying problems.

Preliminary Analysis

Before conducting the main analyses, an analysis was conducted to ensure that the two prior knowledge sets were "balanced" so that performance at identifying problems was equivalent for the two sets. A profile analysis appropriate for split-plot designs was conducted. Obviousness and prior knowledge were the repeated measures factors and prior knowledge set was the between-subjects factor. Paralleling the results in the general summary, the analysis revealed no significant effect on accuracy at identifying problems due to the set of problems about which subjects were informed ($F(1,22) = 2.5, p = .13$). Also, set did not interact with obviousness or with prior knowledge (both F values < 1).

Primary Analysis

The primary analysis addressed whether or not the independent variables (obviousness, prior knowledge, and training) affected accuracy at identifying problems in the videotape summaries. A profile analysis was

conducted with obviousness and prior knowledge as repeated measures factors and training as a between-subjects factor.

Interaction effects on problem identification

Only the two balanced sets of problems of which subjects had prior knowledge were included in the initial analysis. Statistical tests revealed a non-significant 3-way interaction involving obviousness, prior knowledge, and training ($F < 1$). Further, the 2-way interactions were non-significant between obviousness and prior knowledge ($F(2,44) = 1.38, p = .26$), between obviousness and training ($F < 1$), and between prior knowledge and training ($F < 1$).

As found in the general summary, the percent correct scores indicated that for the low and medium obvious problems there was a tendency for the observers to identify more of the problems of which they had prior knowledge than those of which they had no prior knowledge. A difference in the scores by prior knowledge did not occur for the high obvious problems. Figure K.1 shows this pattern. Simple-effect tests revealed that the differences were non-significant for both the low obvious problems ($F(1,22) = 3.12, p = .09$) and the medium obvious problems ($F(1,22) = 2.47, p = .13$).

Figure K.1. Accuracy scores for problem identification by obviousness and prior knowledge.

Effect of Obviousness on problem identification

A profile analysis involving all problems in the summaries was conducted to examine the effects of obviousness on accuracy at identifying problems. There was a significant main effect ($F(1,44) = 194.88, p < .001, \eta^2 = .81$)¹. As shown in Figure K.2, performance was poorest for the low obvious problems, and improved as the obviousness of the problems increased. The mean accuracy was 52% for low obvious problems, 69% for medium obvious problems, and 90% for high obvious problems.

¹ The effect was also significant when the analysis included only the two sets of which subjects were informed ($F(1,44) = 96.14, p < .001, \eta^2 = .67$).

Figure K.2. Accuracy scores for problem identification by obviousness.

Effect of prior knowledge on problem identification

There was a significant main effect for prior knowledge on accuracy at identifying problems ($F(1,22) = 4.13, p = .05, \eta^2 = .15$). The means in Figure K.3 indicate that subjects performed better on problems of which they had prior knowledge (mean accuracy = 79%) than on problems of which they did not have prior knowledge (mean accuracy = 72%).

Figure K.3. Accuracy scores for problem identification by prior knowledge.

Effect of training on problem identification

An analysis on the effects of training revealed a non-significant main effect ($F < 1$). As shown in Figure K.4, subjects in the extended training condition did not identify problems better than subjects in the basic training condition. The mean accuracy was 69% for the extended training group, and 72% for the basic training group.

Figure K.4. Accuracy scores for problem identification by training group.

Qualitative Analyses on Other Problem Reports

Additional Problems

Analyses were conducted to examine additional problem reports in the videotape summaries. Additional problems are problems identified by subjects that were not reported by experts. A total of 111 additional problems were reported in the videotape summaries. The problems are described under the categories invalid, possibly valid, and valid.

Invalid problems identified by subjects

Paralleling the results in the general summary, subjects reported several invalid problems. These problems, which are false alarms, accounted for 89% of the additional problem reports. The frequency of invalid problems reported by each training group is shown in Table K.1. The table shows that subjects in the basic training group reported 64 invalid problems while subjects in the extended training group reported 35 problems. The mean number of invalid problems was 2.9 for subjects in the extended training group, and 5.3 for subjects in the basic training group. A t-test to examine this difference was non-significant ($t(22) = 1.48, p = .153$). Also, 83% of the subjects recorded one or more of the invalid problems.

Table K.1

Description of invalid problems and frequency of invalid reports for each training group.

Category	Description of Category	Frequency of Reports	
		Extended Training	Basic Training
Incorrectly stated	The subject incorrectly described the problem.	22	38
Never happened	The subject described a behaviour or problem that never occurred.	13	26
TOTAL		35	64

Possibly valid problems identified by subjects

Possibly valid problems were very infrequent. As shown in Table K.2, there were only four occurrences of possibly valid problems for the basic training group and zero for the extended training group.

Table K.2

Description of possibly valid problems and frequency of possibly valid reports for each training group.

Category	Description of Category	<u>Frequency of Reports</u>	
		Extended Training	Basic Training
Vague	The subjects' description was too vague to determine if they correctly saw the behaviour or problem.	0	4

Valid problems identified by subjects

As shown in Table K.3, there was a small number of occurrences of Valid problems. There were only four occurrences for each training group.

Table K.3

Description of valid problems and frequency of valid reports for each training group.

Category	Description and Example of Category	<u>Frequency of Reports</u>	
		Extended Training	Basic Training
Possible interpretation	The subjects' interpretation of the problem was different from the experts. The subject or experts may have interpreted more from the problem.	3	2
Sub-problem	The problem was part of another problem and did not have to be specified separately. The subject separated one aspect of the problem as a separate type of problem.	1	2
TOTAL		4	4

Appendix L

Rules for Scoring

Following are the rules used to score problem identification by the experimenter and another scorer.

1. A score of 1 is assigned to each problem the observer identified that was also identified by the experts.
2. Identification of a problem can either be a statement of the usability problem or describing the behaviour of the user. Generally, by stating the behaviour, the observer is also identifying the problem. An example of describing the behaviour is "pressed 130 instead of 3 for temperature". An example of stating the problem is "did not know how to enter temperature - entered all 3 digits for temperature instead of just the middle digit". The subject receives a score of 1 for either type of description.
3. If an observer joined two or more problems as a single problem, then the observer receives a score of 1 for each problem. For example, an observer may have stated as a single problem "the user doesn't know how to enter the temperature and doesn't like how it has to be entered". The observer would receive a score of 1 for identifying the problem of not knowing how to enter temperature and a score of 1 for identifying the problem of not liking the procedure for entering the temperature.

4. If an observer described the problem and the behaviour, and one was stated incorrectly but the other correctly then a point of 1 is still given. For example, one observer reported the problem of not knowing to push the TIME button first as "Order Confusion. User didn't hit the Time button first, just tried to input the length of cooking time." The first part of this problem description "Order Confusion" is incorrect but the remaining part of the description is correct.²

² Although the observer was not completely accurate when they incorrectly reported either the problem or the behaviour of the user, there was awareness of the problem. Cases like this accounted for less than 3% of the problems for which observers received a score of 1.

Appendix M

Analysis on All Prior Knowledge Sets

An analysis was conducted to examine observer accuracy at identifying problems for all three levels of the prior knowledge variable. The purpose of this analysis was to compare performance on the problems of which no one was informed to performance on the prior knowledge sets, in particular, the problems of which subjects were not informed but other subjects were informed.

The analysis revealed that subjects' accuracy at identifying problems for the problems of which no one was informed was significantly lower than the accuracy for those problems of which subjects were not informed but other subjects were informed ($F(1,44) = 22.86, p < .001, \eta^2 = .24$).³ Figure M.1 shows the mean accuracy scores for the three prior knowledge sets. The mean accuracy was 75% for problems of which subjects were informed, 69% for problems of which subjects were not informed but other subjects were informed, and 55% for problems of which no one was informed. The results suggest that the problems of which no one was informed consisted of a different set than those counterbalanced in the informed and not informed sets.

³ A significant effect was also found on the videotape summaries ($F(1,44) = 15.64, p = .001, \eta^2 = .18$).

Figure M.1. Accuracy scores for the three levels of the prior knowledge variable.

To examine why the accuracy was lower in the set of problems about which no one had prior knowledge and, in particular, why the accuracy was significantly lower than the accuracy for the problems in the not informed condition, a qualitative analysis was conducted. It was suspected that the problems in the prior knowledge sets were more predictable than problems in the set about which no one was informed. The rationale for this suspicion is that problems in the two balanced prior knowledge sets were selected so that subjects would believe that the problems could have actually been predicted by designers.

To perform the qualitative analysis, problems that subjects themselves predicted while learning the microwave were examined. The predictions were based on spontaneous reports by subjects of potential problems and problems they experienced while learning the microwave.⁴

⁴ It is believed that if subjects experienced a problem, they would anticipate it to be a problem for the users.

Table M.1 shows the frequency and percentage of problems experienced or informally mentioned by at least one subject during the training. The table shows the predictions for the two prior knowledge sets and for the set of problems of which no one was informed. Seventy-two percent of the problems in the sets of which subjects had prior knowledge were predicted by at least one subject and only 14% of problems in the set of which no one had prior knowledge were predicted by at least one subject. Therefore, the problems in the two prior knowledge sets seem to have been more predictable than those in the set about which no one was informed. Unpredicted problems were likely more difficult for subjects to detect, or it was easier to identify problems that they themselves predicted or experienced. It is also possible, however, that the problems of which no one was informed were missed because subjects were concentrating on those problems they were instructed to look for, or those they predicted themselves or experienced. These results are, therefore, not conclusive.

Table M.1

Frequency and percent of predicted and unpredicted problems in the Prior Knowledge Sets and in the No Prior Knowledge Set.

	<u>Prior Knowledge Sets</u>		<u>No Prior Knowledge Set</u>	
	<u>Frequency</u>	<u>Percent</u>	<u>Frequency</u>	<u>Percent</u>
Predicted Problems	13	72%	5	14%
Not Predicted Problems	5	18%	31	86%
Total	18	100%	36	100%

Appendix N

Pearson Correlation on Importance Ratings

The correlation between subjects' ratings of importance on the first rating set and on the second rating set was examined. Mean ratings in each rating set were calculated for those problems identified and for those not identified. These mean ratings across all subjects are shown in Table N.1. The table shows that on the first instance that subjects rated importance, the mean rating for problems of which they identified was 3.1. Similarly, on the second rating set, the mean rating on problems of which they identified was 3.2. A Pearson correlation of .82 ($p < .001$) was obtained between mean ratings of importance in the first rating set and on mean ratings of importance in the second rating set. Therefore, subjects gave similar ratings of importance on the two rating sets.

Similarly, on the second rating set, a Pearson correlation of .74 ($p < .001$) was obtained between mean ratings on problems they identified and problems they did not identify. The mean rating on problems they identified was 3.2 and the mean rating on problems they did not identify, but rated in the complete list of problems was 2.9. The correlation and mean ratings indicate that subjects were rating the problems similarly and were not giving, for example, lower ratings to those they did not identify.

Table N.1

Mean ratings of importance for identified and unidentified problems in the two rating sets.

	<u>First rating set</u>		<u>Second rating set</u>	
	Not identified	Identified	Not identified	Identified
Mean Rating	N/A	3.1	2.9	3.2

Appendix O

Qualitative Analysis on Problems Identified

An analysis was conducted to examine subjects accuracy performance on individual problems. Average scores for each problem in the general summary were calculated. For example, an average score of 50 for a particular problem indicates that 50% of the subjects identified the problem. Problems for which subjects performed significantly lower than expected by chance, not significantly different from chance, and significantly higher than expected by chance were examined. This analysis involved computation of the 5% significance levels for the Binomial Distribution based on the hypothesis that subjects were guessing ($p = .5$). The lower and upper 5% critical regions were 30 and 70. These scores determined whether the average agreement score for each problem identified had a probability that is extreme (Siegel, 1956)⁵. The average score for each problem in the general summary is shown in Table O.1.

Using the scores of 30 and 70 at the 5% critical regions, the problems were categorized by three results; those in which performance was significantly lower than chance (mean score < 30), those in which performance was not significantly different from chance ($30 \leq \text{mean score} \leq 70$), and those in which performance was significantly greater than chance (mean score > 70). Subjects performed worse than expected by chance (mean score < 30) on only seven of the 54 problems (13%). Five of the seven problems were low obvious problems, one was medium obvious, and one was highly obvious. Also, on a 5-point rating scale, experts rated three of these problems as not at all important (1) and three of them as very

⁵ Siegel, S. (1956). Nonparametric Statistics for the Behavioral Sciences. New York. McGraw-Hill Book Company, Inc.

not significantly different from chance ($30 \geq \text{mean score} \leq 70$). For these problems, the level of obviousness varied from low to high obvious. For 29 of the 54 problems (54%) performance was better than expected by chance (mean score > 70). The majority of these problems were highly obvious (69%).

Similar results were found for the videotape summaries. As shown in Table O.2, subjects performed worse than expected by chance (mean score < 30) on only eight of the 79 problems (10%). Five of the eight problems were low obvious problems, two were medium obvious, and one was highly obvious. For 20 of the 79 problems (25%), agreement scores were not significantly different from chance ($30 \geq \text{mean score} \leq 70$), and for 51 of the 79 problems (65%) their performance was significantly better than expected by chance (mean score > 70). For the problems where subjects operated at a level not significantly different from chance, the problems varied in their level of obviousness. For problems detected at levels significantly better than expected by chance, the majority were highly obvious (66%).

Table O.1

Obviousness, importance, and percent of observers who identified each problem in the general summary.

Problem	Percent correct	Obviousness (1=Low, 3=High)	Importance (1=Not at all, 5=Very)
Forgot to CLEAR when had to restart	00	1	2
Poor feedback - think not working when enter 24 hour time	04	1	5
Sequence error - touch numbers before touching TOD	04	1	1
Thought oven start after putting in probe	04	1	1
Input temperature incorrectly - put 11	13	2	5
Not sure if like it that enter fraction of a pound	21	1	1
Not understand how enter temp. when prompted	29	3	5
Not sure if pressing TOD is correct because no feedback	33	3	5
Hit TEMP instead of TIME	33	2	1
Not know what 1stStg means	38	1	2
Reluctant to do temp. because doesn't understand	38	1	5
Entering TEMP is inconsistent with entering TIME	46	1	5
Didn't know to use MEMORY	46	2	5
Entered power incorrectly - 3 and 0	46	3	3
Not sure if 110 = 1 min. & 10 secs. or 110 secs.	50	3	2
Put 100 for 1 minute from current time	50	2	4
Entered POWER incorrectly - put 6 and 0	54	2	3
Not know how to start sequence - put # and START	58	3	2
Forgot to put in temperature values (slip)	58	3	2
Not know what info. to put in because no colon	63	3	5
Try to set POWER before TEMP	63	1	2
Forgot to put probe in	63	1	3
Pressed TIME before touching POWER (like TOD)	67	3	1
Didn't want to specify temp. again if same as cooking temp.	67	2	2
Expected to put 0 in after entering 1	67	2	2
Tried to set DELAY through TOD clock	71	1	3
Used MEMORY too often	71	3	5
Pressed TIME before touching START (like TOD)	75	2	1
Thinks system is dumb/doesn't look like temp	75	2	5
No indicator if 12 or 24 hour clock - enter 24 hour time	79	3	4
Confused by COOL feedback	79	2	2
Want HOLD WARM for time as well as temp.	79	1	2
Not know how to begin/button confusion-used TIME & TOD	79	3	3
Wants to see power level when cooking	83	2	3
Surprised to press 5 and get 50	83	1	3
Doesn't like way to specify temperature	83	2	5
Not know how to enter weight - not expect to put 0 in	83	3	3
Used MEMORY instead of HOLD WARM	83	3	1
Didn't know have to enter temp. for HOLD WARM	83	3	2
Want oven to stop between stages	83	2	1
Not know what to do after enter numbers - touched START	88	3	5
Not sure if 10 is 10 seconds	88	2	5
Entered 70 secs. - Not know how to enter time	88	3	5
Put 60 secs. instead of 100 for 1 minute	88	3	5
Expect/try to enter time as well as weight	92	3	2
Confused by meaning of HI%	96	3	4
Use QUICK-ON for speed instead of POWER	96	3	2
Not know how to terminate sequence with DELAY button	96	3	5
Confused TIME and TOD	100	3	3
Not know how to set power to half speed	100	3	5
Not know how to enter temperature	100	3	5
Problems entering TEMP	100	3	5
Want grams, not pounds	100	3	3
Not know where to use MEMORY	100	3	5

Note: Problems are divided into three sections. The first section is for problems where agreement scores were significantly worse than expected by chance, the second section for agreement scores that were not significantly different from chance, and the third section for agreement scores significantly better than expected by chance.

Table O.2

Obviousness, and percent of observers who identified each problem in the videotape summaries.

Videotape in which problem was reported/Problem	Percent Correct	Obviousness (1=Low, 3=High)
2/ Sequence error - put time before touching TOD	00	1
2/ Forget to touch CLEAR to restart when stuck	00	1
1/ Thought oven should start after put in probe	04	1
2/ Poor feedback - think not working when enter 24 hour time	08	1
3/ Touched 1 twice to enter temp. of 110°	13	2
2/ Not know TOD is correct because no change in display	21	3
1/ Not sure if like it that enter fraction of pound	25	1
3/ Not know TOD is correct because no change in display	29	2
1/ Lack of feedback when press correct command, TOD	38	3
1/ Not know meaning of 1stSig.	38	1
1/ Expected to put in 0 after entering 1	38	1
2/ Think single value does not look like temperature	42	1
3/ Sees way of entering temp. as weird	42	2
2/ Not sure what HI% means	50	1
2/ Entered POWER incorrectly, put 0 after 6	50	2
2/ Entered POWER incorrectly, put 0 after 3	50	3
3/ Not put temp. in after touching HOLD WARM	50	3
3/ Accidentally pressed TEMP when wanted TIME	50	2
1/ Not know what COOL feedback means - not useful	54	1
1/ Not know how to enter DELAY time - use 100	54	2
1/ Bit of confusion that HI% = fastest speed	58	2
3/ Did not know what COOL means	58	2
3/ Not know how to set DELAY time - use TOD with DELAY	58	1
3/ Extra MEMORY - used it before and after setting 1st stage	58	3
2/ Extra MEMORY - used it after setting 2nd stage	63	2
1/ Not know how to start time - touch START then TOD	67	2
2/ Didn't know to use MEMORY	67	2
3/ Tried setting power before temperature	67	1

1/ Not know how to enter time, no colon and enter secs	71	3
1/ Not know how to start sequence - put # and START	71	3
1/ Don't like how to enter temp.	71	1
1/ Used MEMORY too often	71	3
2/ Not sure if 110 = 1 min. & 10 secs. or 110 secs.	71	3
2/ Expected to put in 0 after entering 1	71	2
3/ Way of entering temp. is inconsistent with entering time	71	1
1/ Forgot probe and no feedback to insert	75	1
2/ Dislikes temp. feature before trying	75	1
3/ Tried to press START after entering time instead of TOD	75	2
3/ Forgot to put in temp. value/ pressed POWER prematurely	75	3
2/ Not understand temp. even when prompted	79	3
3/ Touched TIME after entering time	79	2
3/ Didn't like having to enter same temp. for HOLD WARM	79	2
1/ Surprised to see .1 for 1 pound - entering 0 unexpected	83	3
1/ Want HOLD WARM for time as well as temp.	83	1
1/ Wants to stop oven between stages to add food	83	2
2/ Confused TIME and TOD	83	3
1/ Thought could use HOLD WARM without entering temp.	88	3
1/ Not know have to enter DELAY again after specify time	88	3
2/ Does not know how to terminate - try START, then TOD	88	3
2/ Not know where to use MEMORY	88	2
3/ Wants to see power level while cooking	88	2
3/ Used MEMORY instead of HOLD WARM - name error	88	3
3/ Not know how to terminate - tried MEMORY after entering time	88	2
1/ Not know how to set POWER to half speed	92	3
2/ Try to put in 24 hour clock - no indication if 12 or 24	92	3
2/ Not sure if 10 = secs. - poor feedback from display	92	3
2/ Can't set temp.	92	3
2/ Think design is dumb/still doesn't look like temp	92	2
3/ Again, touched TIME after entering time	92	3
3/ Didn't know what HI% means	92	3
1/ Confused by names TIME and TOD	96	3
1/ Confusion with QUICK-ON and POWER	96	3
1/ Expect to enter 0 after entering 5 for 50% POWER	96	1
1/ Try to put 14 in for 140°	96	2
1/ Not know where to use MEMORY	96	3
2/ Think have to enter TIME as well as defrost weight	96	3
2/ Didn't know had to press DELAY after entering time	96	3
3/ Confused TIME and TIME OF DAY	96	3
3/ Pressed TIME after entering defrost value- think have to enter time	96	3
3/ Entered 60 secs. for time	96	3
1/ Can't enter temp. - expect to put in 130	100	3
2/ Entered 70 secs. - not know how to enter time	100	3
2/ Not know how to enter half speed for POWER	100	3
2/ Can't set temp.	100	3
3/ Didn't know how to change power to 50%	100	3
3/ Did not know how to specify temperature	100	3
3/ Not know where to use MEMORY, before or after first stage	100	3

Note: Problems are divided into three sections. The first section is for problems where agreement scores were significantly worse than expected by chance, the second section for agreement scores that were not significantly different from chance, and the third section for agreement scores significantly better than expected by chance

An analysis was also conducted to examine the importance ratings of subjects on individual problems in comparison to the experts' ratings. Agreement with the experts was defined as perfect agreement. This measure is a conservative estimate of agreement because small differences in ratings are considered as disagreement. For example, a rating of 4 by a subject, and a rating of 5 by an expert were considered as a disagreement even though both indicate fairly high importance. Average agreement scores on importance ratings for each problem in the general summary were calculated. For example, an average score of 50 for a particular problem indicates that 50% of the subjects agreed with the experts on the rating of importance. An analysis was conducted to examine for which problems average agreement scores with experts were significantly worse than expected by chance, not significantly different from chance, and significantly better than expected by chance. The 5% critical regions were computed for the Binomial Distribution based on the hypothesis that subjects were guessing ($p = .2$). Since 5 ratings were used, there was a 1 in 5 chance (.20) that the ratings would be the same based on chance alone. The lower and upper bounds for the 5% level were 4 and 38. These scores determined whether the average agreement score for each problem had a probability that is extreme. The agreement scores for each problem are shown in Table O.3.⁶

Using the scores of 4 and 38, the problems fall into two categories; those in which performance was not significantly different from chance ($4 \leq \text{mean score} \leq 39$), and those in which performance was significantly greater than chance (mean score > 38). There were no problems for which

⁶ Three of the problems were excluded from the analysis because the descriptions of the problems in the list presented to observers were not clearly articulated.

subjects performed worse than expected by chance. For 33 of the 51 problems (55%), the agreement scores were not significantly different from chance. These problems varied in their level of importance from not important to highly important. For 18 of 51 problems (35%), performance was better than expected by chance (mean score > 38). Thirteen of these 18 problems (72%) were highly important (importance rating = 5). The other five problems (28%) had low importance ratings (importance rating = 1 or 2). Therefore, subjects performed better on those problems that had extreme importance ratings which were either clearly important or unimportant. Also, 50% of the problems were high obvious problems, and 50% were either low or medium obvious problems.

Table O.3

Obviousness, importance, and percent of observers who gave the same importance rating as the experts for each problem in the general summary.

Problem	Percent correct	Obviousness (1=Low, 3=High)	Importance (1=Not at all, 5=Very)
No indicator if 12 or 24 hour clock - enter 24 hour time	4	3	4
Not sure if 110 = 1 min. & 10 secs. or 110 secs.	8	3	2
Not know how to begin/button confusion - used TIME & TOD	8	3	3
Tried to set DELAY through TOD clock	13	1	3
Pressed TIME before touching START (like TOD)	13	2	1
Want grams, not pounds	13	3	3
Didn't know to use MEMORY	17	2	5
Not sure if 10 is 10 seconds	17	2	5
Entered POWER incorrectly - put 6 and 0	17	2	3
Forgot to put probe in	21	1	3
Forgot to CLEAR when had to restart	21	1	2
Forgot to put in temperature values (slip)	21	3	2
Didn't know have to enter temp. for HOLD WARM	21	3	2
Want HOLD WARM for time as well as temp.	21	1	2
Put 100 for 1 minute from current time	21	2	4
Want oven to stop between stages	21	2	1
Expected to put 0 in after entering 1	21	2	2
Put 60 secs. instead of 100 for 1 minute	21	3	5
Confused TIME and TOD	25	3	3
Entered 70 secs. - Not know how to enter time	25	3	5
Surprised to press 5 and get 50	25	1	3
Confused by COOL feedback	25	2	2
Expect/try to enter time as well as weight	25	3	2
Didn't want to specify temp. again if same as cooking temp.	25	2	2
Entered power incorrectly - put 3 and 0	29	3	3
Pressed TIME before touching POWER (like TOD)	33	3	1
Confused by meaning of HI%	33	3	4
Wants to see power level when cooking	33	2	3
Not know how to enter weight - not expect to put 0 in	33	3	3
Not sure if like it that enter fraction of a pound	33	1	1
Used MEMORY instead of HOLD WARM	33	3	1
Not know what 1st Stg means	38	1	2
Use QUICK-ON for speed instead of POWER	38	3	2
Sequence error - touch numbers before touching TOD	46	1	1
Not know how to start sequence - put # and START	46	3	2
Try to set POWER before TEMP	46	1	2
Entering TEMP is inconsistent with entering TIME	46	1	5
Used MEMORY too often	46	3	5
Not sure if pressing TOD is correct because no feedback	50	3	5
Not know what to do after entered numbers - touched START	50	3	5
Input temperature incorrectly - put 11	54	2	5
Not know where to use MEMORY	54	3	5
Thinks system is dumb/doesn't look like temp	58	2	5
Reluctant to do temp. because doesn't understand	67	1	5
Not know how to terminate DELAY sequence with DELAY button	71	3	5
Hit TEMP instead of TIME	71	2	1
Thought oven start after putting in probe	71	1	1
Doesn't like way to specify temperature	75	2	5
Not know how to set power to half speed	79	3	5
Problems entering TEMP	79	3	5
Not know how to enter temperature	96	3	5

Note: Problems are divided into two sections. The first section is for problems where agreement scores were not significantly different from chance, and the second section for agreement scores significantly better than expected by chance.

Appendix P

Analyses on Gender and Microwave Usage

A series of analyses were performed to examine the effects of gender and prior experience with microwave ovens on accuracy at identifying problems and rating importance. Analyses were also planned to examine the effects of past experience with observational techniques and training. However, based on an examination of subjects' responses to the questions on observational experience and training, it was found that interpretation of these questions by subjects varied.⁷ Therefore, analyses with these two variables were not conducted.

Effect of Variables on Problem Identification

Multivariate analyses of variance tests were conducted on performance scores in the general summary. The repeated measures were obviousness and prior knowledge, and the between-subjects variable was either gender or microwave usage. The dependent variable was accuracy at identifying problems. The results revealed no main effects for gender or microwave usage on accuracy at identifying problems. Also, there were no significant interactions involving gender or microwave oven usage and obviousness or prior knowledge.⁸ All F values were less than 1.

⁷ Examples of observational experiences reported by subjects included examining in vitro tissue slices, observing mice under the influence of drugs, conducting hypnosis studies, and observing males and females in pubs. It is believed that some subjects' were liberal in their definition of observational experience.

⁸ The results for the videotape summaries parallel those in the general summary. All main effects and interactions were non-significant.

Effect of Variables on Importance Ratings

Analyses were conducted to examine the effects of gender and microwave oven usage on mean ratings of importance and on variability of importance ratings from the experts' ratings. Multivariate analyses of variance tests were conducted with prior knowledge and obviousness as repeated measures and microwave oven usage or gender as a between-subjects factor. The results are presented in Table P.1.

Table P.1

F values for effects of gender and microwave usage on importance ratings.

Statistical Test	F value
Mean importance ratings	
Gender	$F(1,22) = 1.96, p = .18$
Gender by obviousness	$F < 1$
Gender by prior knowledge	$F < 1$
Microwave	$F < 1$
Microwave by obviousness	$F(2,44) = 2.93, p = .06$
Microwave by prior knowledge	$F(1,22) = 3.12, p = .09$
Variability of ratings from experts' ratings	
Gender	$F(1,22) = 2.58, p = .12$
Gender by obviousness	$F < 1$
Gender by prior knowledge	$F < 1$
Microwave	$F(1,22) = 1.44, p = .23$
Microwave by obviousness	$F < 1$
Microwave by prior knowledge	$F < 1$

The analyses revealed no significant main effects or significant interactions on mean ratings of importance or on variability of importance ratings from the experts' ratings. There was, however, a marginal interaction between microwave usage and obviousness on mean ratings of importance. The mean ratings are shown in Figure P.1. Simple-effect

tests were conducted. It was found that for microwave users and non-microwave users, the mean importance rating significantly increased with an increase in obviousness. Also, there was no significant difference in the mean importance ratings of the microwave and the non-microwave users at each level of the obviousness variable. The difference in the mean ratings, however, was greatest for the medium obvious problems thus accounting for the marginal interaction.

Figure P.1. Mean ratings of importance by microwave and non-microwave users.

Appendix Q

Analysis at Identifying the Number of Users Who Experienced Usability Problems

When subjects completed the general summary they also identified how many of the users in the three videotapes experienced each problem. In this analysis, performance at identifying the number of users was examined. Performance was measured by examining variability from experts at identifying the number of users.

Variability from Experts at Identifying the Number of Users

Performance at identifying the number of users was calculated by taking the absolute value of the difference between the number of users reported by the experts and the number of users reported by the subjects.

$$\text{Absolute (Subject number of users - Expert number of users)}$$

The absolute value was included in the formula so that positive and negative differences between the experts' and a subject's identified number of users would not cancel out and sum to 0 to indicate no variability. A score of 0 indicates no variability from the experts while a score greater than 0 indicates variability.

There were few errors in recording the number of users for those problems that subjects identified. The average variability score from the experts was only 0.15. Because of this low variability, further analyses were not conducted.

END

2|3·1 2·92

FIN