COMPARISON OF EMPIRICAL TESTING AND WALKTHROUGH METHODS IN USER INTERFACE EVALUATION

Clare-Marie Karat, Robert Campbell, and Tarra Fiegel***

IBM T. J. Watson Research Center
PO Box 704, Yorktown Heights, NY 10598
ckarat@watson.ibm.com

* Now at Department of Psychology, Clemson University, Clemson, SC 29634
**Now at Department of Psychology, New Mexico State University, Las Cruces, NM 88003

ABSTRACT

We investigated the relative effectiveness of empirical usability testing and individual and team walkthrough methods in identifying usability problems in two graphical user interface office systems. The findings were replicated across the two systems and show that the empirical testing condition identified the largest number of problems, and identified a significant number of relatively severe problems that were missed by the walkthrough conditions. Team walkthroughs achieved better results than individual walkthroughs in some areas. About a third of the significant usability problems identified were common across all methods. Cost-effectiveness data show that empirical testing required the same or less time to identify each problem when compared to walkthroughs.

KEYWORDS: Empirical testing, walkthroughs, problem severity, cost-effectiveness, scenarios

INTRODUCTION

Software development teams work within cost, schedule, personnel and technological constraints. In recent years, usability engineering methods appropriate to these constraints have evolved and become increasingly incorporated into software development cycles. Human factors practitioners currently rely on two types of techniques to evaluate representations of user interfaces: (1) empirical usability testing in laboratory or field settings; and (2) a variety of usability walkthrough methods. These latter methods have substantive differences and are referred to as pluralistic walkthroughs, heuristic evaluations, cognitive walkthroughs, think-aloud evaluations, and scenario-based and guideline-based reviews [1, 4, 10, 12, 16, 18, 19, 20, 21, 22, 23]. Empirical usability testing and walkthrough methods differ in the experimental controls employed in the former.

Human factors practitioners must make tradeoffs regarding time, cost, and human factors issues in selecting a usability engineering method to use in a particular development situation [15]. Use of walkthroughs has been encouraged by development cycle pressures and by the adoption of development goals of efficiency and user-centered design [1, 2, 5, 12, 20, 23]. Many questions remain about how walkthroughs compare to empirical methods of usability assessment, and when and how walkthroughs are most effective.

1992 ACM 0-89791-513-5/92/0005-0397 1.50

Questions About Empirical Testing Versus Walkthrough Methods

Usability problems. How do the two methods compare in the number of usability problems identified in a user interface? Is one method better than the other in identifying serious problems? How many of the problems are found by both methods and how many are found solely by one method?

Reliability of differences. If the methods differ in their effectiveness in identifying usability problems in user interfaces, do these differences persist across different systems? Or is the effectiveness of an evaluation method system dependent, based on the type of interface style and metaphor used in the interface?

Cost-effectiveness. What is the relative cost-effectiveness of the two techniques in identifying the usability problems in an interface?

Human factors involvement. What amount of human factors involvement is necessary in the use of the two techniques? What issues arise in analyzing and interpreting data?

Questions About Walkthrough Techniques

Individuals versus teams. Are walkthroughs more effective when conducted individually or in teams? Social psychology has documented that groups seldom perform up to the level of their best member [17]. One exception in this area is that groups do offer the possibility of more accurate judgments than individuals, especially when working on complex tasks [17]. The use of interaction-enhancing procedures may heighten group productivity as well [7].

Evaluator expertise. Are members of development teams and representative end users effective evaluators

in walkthroughs, or should the evaluators be exclusively human factors or user interface (UI) specialists?

Prescribed tasks versus self-guided exploration. In a usability walkthrough with prescribed tasks, evaluators step through a representation of a user interface (e.g., paper specification, prototype) or the actual system while performing representative end user tasks. Other walkthrough procedures rely on self-guided exploration of the interface by evaluators who may or may not generate scenarios for that purpose. Do evaluators in walkthroughs think that one approach is more useful than the other in identifying usability problems?

Utility of guidelines. What is the role of usability heuristics or guidelines in usability walkthroughs? Are heuristics useful and necessary for experienced members of development teams?

Recent Data

Recent studies provide some data on these issues. Jeffries, Miller, Wharton, and Uyeda [10] compared the effectiveness of usability testing, guideline, heuristic, and cognitive walkthrough [16] methods in identifying user interface problems. The heuristic method diverged from the Nielsen and Molich [20] method as it was completed by UI specialists and did not include the use of written heuristics or guidelines. Prescribed task scenarios were employed in usability testing and cognitive walkthroughs. Results showed that the heuristic method identified the most usability problems and more of the serious problems, and did so at the lowest cost of the four techniques. Usability testing was generally the second-best method of the four in identifying problems. A number of questions arise: Was evaluator expertise the key component in the effectiveness of both the heuristic and usability testing methods? What was the amount of overlap in the problems identified by the four methods? What was the inter-rater reliability of problem identification?

Desurvire, Lawrence, and Atwood [4] compared the effectiveness of empirical usability testing and heuristic evaluations in identifying violations of usability guidelines. The heuristic method differed from others [10, 20] in that evaluators rated guidelines on bipolar scales for each of a set of tasks. Laboratory testing identified violations of six of ten relevant guidelines, while the combined results from the heuristic evaluations identified only one violation of a guideline. The heuristic ratings from UI experts and empirical usability test participants were predictive of laboratory user performance data. Non-UI experts' ratings were not predictive of performance. Heuristic ratings were effective in identifying tasks where problems would occur, but not the specific user interface problems themselves. UI experts' "best guess" predictions of performance were comparable to their heuristic ratings.

Bias [1] describes a systematic group evaluation procedure called the pluralistic usability walkthrough that includes end user, architect, design, developer, publication writer, and human factors representatives who complete scenario-driven walkthroughs of software prototypes. Human factors staff lead the group sessions. The design-test-redesign cycle is reduced to minutes through the use of low-technology prototypes to illustrate alternative designs, and through the presence and cooperation of individuals with the varied skills required to complete the work. This technique highlights the value of multidisciplinary activity in design [6] and group problem solving [7], and the iterative design possible within tight time constraints. A question arises about the group facilitation skills and procedures required for human factors engineers to achieve high group productivity and accurate judgments in the walkthroughs [7, 17].

Nielsen and Molich [20] tested heuristic evaluations completed individually by evaluators who were not human factors experts. In three of the four studies reported, evaluators received a lecture on nine usability heuristics and related reference material. No prescribed task scenarios were employed. Aggregates of data from five computer science students or industry computer professionals generally found about two-thirds of the problems that the authors had previously identified in the interfaces. While the study measured how hard it was to find problems, there was no measure of the severity of the identified usability problems.

There is some evidence that evaluators other than UI specialists can carry out useful and successful think-aloud and heuristic walkthroughs [19, 20]. Also, Jorgensen [12] and Wright and Monk [23] both provide evidence of the success of a think-aloud technique used by developers who had minimal training on the procedure and limited use of human factors resource. Developers observed users who thought aloud while working through tasks on the developers' systems. Walkthroughs were done by individuals in the former study and by teams in the latter. Questions arise about the numbers and types of problems that were not identified in these studies, and how data on problems were interpreted and analyzed.

Goals of the Study

Our study had three goals: The first goal of the study was to better understand the relationship between empirical testing and walkthrough results. This information would improve the understanding of the tradeoffs in selecting one rather than another in a particular situation. This goal included assessment of the number and severity of usability problems identified by the two methods and the resource required to identify them. The second goal was to determine whether the results regarding the relative effectiveness of empirical and walkthrough methods were reliable and would replicate across systems, or whether these results were system dependent. The third goal of the study was to understand how well walkthroughs work in user interface evaluation and how to improve their effectiveness. An effective walkthrough method would be one that identifies most usability problems in an interface, and especially the most severe usability problems. The role of individuals and teams, evaluator characteristics, scenarios, self-guided exploration, and usability heuristics in walkthroughs were explored as part of this third goal.

METHOD

Design

Three user interface evaluation methods were assessed: empirical usability test, individual usability walkthrough, and team usability walkthrough. The usability walkthrough procedure developed and used in this study included components to maximize effectiveness of usability problem identification, based on previous research. The walkthrough included separate segments for 1) self-guided exploration of a graphical user interface (GUI) office system, and 2) use of prescribed scenarios. The procedure utilized a set of 12 usability guidelines. Walkthroughs were conducted individually by six evaluators in one condition and by six pairs of evaluators in another condition. The evaluators in the team condition conversed with each other about issues and problems during the sessions. The evaluators were responsible for documenting the usability problems they identified in the walkthroughs.

The empirical usability test method also had separate segments for 1) self-guided exploration of a GUI system, and 2) use of prescribed scenarios. The six users in the usability tests were asked to describe usability problems they encountered, and problems were recorded by the human factors staff who were observing the sessions.

The usability problems identified through use of the three methods were categorized using common metrics. Thus data could be compared across methods on dimensions including number and severity of usability problems identified in the interface. The three methods were each applied to two competitive software systems in order to assess the reliability of the findings. The usability tests and walkthroughs were completed as if part of a realistic development schedule with resource constraints so that the data could provide practical information on usability engineering in product development.

Participants

Six separate groups of experienced (GUI) users participated in the study of the two systems. For each system, the empirical test and individual walkthrough each utilized six participants, and the team walkthrough utilized six pairs of participants. A total of 48 participants took part in the study. Participants were randomly assigned to methods; team members did not know each other. Participants had not previously used the GUI system they worked with in the study. The six groups were comparable based on background data gathered prior to usability sessions. Participants were predominantly end users and developers of GUI systems, along with a few UI specialists and software support staff. Most of the participants had advanced educational degrees; used computers in home, work, and school settings; and had used a variety of computers, operating systems, and applications. They used computers approximately 20 hours a week, including over 10 hours a week on GUI systems. Except for more formal education, the participants were typical of those who would participate in usability walkthroughs and empirical testing of GUI systems in product development.

Materials

The two systems selected for the study were commercially available GUI office environments with integrated text, spreadsheet, and graphics applications. They will be referred to as Systems 1 and 2. The two systems differed substantially in the type of interface style and office metaphor presented.

Human factors staff consulted with end users and developed a set of nine generic task scenarios to be used with both systems. These scenarios were representative of typical office tasks involving text, spreadsheet, and graphics applications, and use of the system environment. The tasks included a range of one to thirteen subtasks and covered document creation, moving and copying within and between documents, linking and updating documents, drawing, printing, interface customization, finding and backing up documents, and use of system-provided and user-generated macros.

A two-page document of guidelines was developed for the evaluators in the walkthrough conditions. The document told evaluators their assignment was to identify usability problems with the interface initially by exploring the interface on their own and then by walking through typical tasks provided to them. A usability problem was defined as anything that interfered with a user's ability to efficiently and effectively complete tasks. Evaluators were asked to keep in mind the guidelines about what makes a system usable and to refer back to them as necessary. Following the precepts of minimalism [3], the document provided brief definitions and task-oriented examples of twelve guidelines. These guidelines were compiled from heuristics used by Nielsen and Molich [20], the ISO working paper on general dialogue principles [9], and the IBM CUA user interface design principles [8]. The twelve usability guidelines included:

- Use a simple and natural dialog,
- Provide an intuitive visual layout,
- Speak the user's language,
- Minimize the user's memory load,
- Be consistent,
- Provide feedback,
- Provide clearly marked exits,
- Provide shortcuts,
- Provide good help,
- Allow user customization,
- Minimize the use and effects of modes, and
- Support input device continuity.

A usability problem description form was developed for use by the walkthrough evaluators. The form instructed the evaluators to briefly describe each problem and then rate its impact on end user task completion.

Procedure

The authors completed all usability engineering work in the study and became familiar with each GUI system prior to commencement of the usability sessions. All sessions were completed in a usability studio in Hawthorne, NY. The GUI systems were set up on an

IBM PS/2 Model 80 with an 8514 display and a printer. Usability sessions for all methods each took about three hours. The first half of the usability sessions included an introduction by the usability engineer who administered the sessions, and a self-guided exploration of the system by the participants. During the self-guided exploration, participants could go through online tutorials, read any of the hard copy documentation shipped with the system, use and modify example documents created using the different applications and system functions, or create new application documents and experiment with system functions. In the second half of the sessions, participants worked through a set of nine typical tasks presented in random order and completed a debriefing questionnaire given by the administrator.

The empirical testing and walkthrough sessions differed in human factors involvement in the session and in how usability problems were documented. In empirical testing, two usability engineers administered each session in its entirety with an individual user. One person in the control studio interacted with users (who were in the usability studio and described usability problems they encountered during sessions), controlled the videotape equipment, and observed usability problems. The second person in the control studio logged user comments, usability problems, time on task, and task success or failure.

Usability staff involvement in the walkthrough sessions was limited to test the resource requirements of the method. One administrator was available on-call during the session in case of unexpected events. A few sample sessions were videotaped and observed by human factors staff; no session logging occurred. One administrator introduced the session and instructed walkthrough evaluators in the use of the guidelines for usability walkthrough document and the usability problem description forms. The administrator emphasized in both individual and two-person team walkthrough conditions that the problem identification sheets were the deliverable for the session. In the team conditions, evaluators were given additional instructions to help each other by providing relevant information [7]. They were told that if either one of the team members thought something was a usability problem, they should record it. Also, team walkthrough evaluators were instructed to take turns with the mouse and with recording usability problems so that each person had direct experience with the interface and with the usability problem description forms. Evaluators read the guidelines document and the administrator then left the studio. After the self-guided exploration phase, the administrator returned briefly to present the task scenarios and emphasize that it was more important to identify usability problems than to complete all the tasks.

RESULTS

Data Analysis

The usability problems recorded during empirical testing and the usability problem descriptions documented during the walkthroughs were classified by the usability engineers using a generic model of usability problems that evolved during the course of the study. The classification completed a content analysis of the problems and prepared the data for subsequent problem severity ratings. The hierarchical model consisted of a total of 47 categories of potential user interface problem areas. Because of the functional differences between the systems, all 47 categories applied to System 1 while only 43 applied to System 2. Subcategories were created when a main category had several problems that were related, yet addressed different aspects of the higher-level category. For example, the fifth main category was Move & Copy and it had two subcategories: 1) Clipboard and 2) Direct Manipulation. We distinguished between problems that were pervasive through the environment and those that were application specific. For example, we classified icon complaints (e.g., cannot understand icon meaning, icons hard to read, no icon status information or it is hard to distinguish) as pervasive office-level problems, while confusion about specific spreadsheet functions (e.g., how does the sum function work?) were regarded as application specific.

Item classification was discussed until consensus was reached about its placement in the model. To assess inter-rater reliability, 50 problem statements were randomly selected from the data for the three conditions and classified by two usability engineers who had not observed the participants or been involved in data analysis. They each classified the data using the generic model of usability problems. The inter-rater reliability scores between the third-party usability staff and the staff involved in the study were 87% for the empirical testing data, 70% for the individual walkthrough data, and 71% for the team walkthrough data. For each empirical testing and walkthrough group, data were analyzed regarding the number of usability problem tokens (all instances), usability problem types (instances minus all duplicates), and problem areas (higher level categories of problem tokens and types, e.g., Move and Copy) in the generic model. These problem areas were assigned Problem Severity Classification (PSC) ratings.

A version of the PSC measure, which is used in IBM, was employed in the study. It provides a ranking of usability problems by severity that can be used to determine allocation of resources for addressing user interface problems (see Table 1). PSC ratings are computed on a two-dimensional scale, where one axis represents the impact of a usability problem on end user ability to complete a task, and the other represents frequency (the percentage of end users who experience the problem). Categories of the impact dimension (high, moderate, low) and frequency dimension (high, moderate) were combined to form an index of PSC ratings that ranged from 1-3 where 1 is most severe. High impact was defined as a problem that prevented the user from completing the task, moderate impact represented significant problems in task completion, and low impact represented minor problems and inefficiencies. Given the small sample sizes in the condi-

tions, moderate frequency was defined as 2 (33%) users, evaluators, or evaluator teams; and high frequency was defined as three (50%) or more of them. For example, if three or more of the six evaluators (high frequency) reported a problem that caused significant difficulty (moderate impact) in completing a task, a PSC rating of "1" would be assigned to the problem area.

	Frequency (Percentage of Users)	
	High	Moderate
Impact on Task		
High	1	1
Moderate	1	2
Low	2	3

Table 1. Problem Severity Classification rating matrix.

We generated PSC ratings for each of the categories in the generic model of user interface problems. There were 47 PSC ratings for System 1, and 43 for System 2. To generate PSC ratings in the empirical conditions, the human factors staff calculated the frequency of users experiencing a problem and assigned an impact score. Disagreements about impact scores were discussed until consensus was reached. In the walkthrough conditions, human factors staff calculated the frequency data and averaged the impact scores provided by the evaluators. Problem areas that did not have problem tokens from at least two participants or teams were assigned a PSC rating of 99 (i.e., no action required). Problem areas with PSC ratings of 1-3 were called significant problem areas (SPA), and those with ratings of 99 were called "no action" areas.

For the walkthrough conditions, evaluator questionnaire data on aspects of the walkthrough procedure were collected during the debriefing sessions and analyzed. For the empirical conditions, data on time on task, completion rates, and the debriefing questionnaire were collected but are not reported here.

Empirical Testing and Walkthrough Results

For Systems 1 and 2, empirical usability testing identified the largest number of usability problem tokens (all instances), followed by team walkthrough and then individual walkthrough (see Table 2). For both systems, the total number of usability problem tokens found by empirical testing was about four times the total number of problems identified by team walkthroughs, and about five times the total number found by individual walkthroughs. The difference in the distribution across the groups of the total number of tokens found was statistically significant for each system at the $p < .01$ level according to χ^2 tests.

	Empirical Test	Team Walk	Individual Walk
System 1			
Problem Tokens	421	115	78
Problem Types	159	68	49
System 2			
Problem Tokens	401	107	64
Problem Types	130	54	39

Table 2. Total identified usability problems.

Empirical testing also identified the largest number of usability problem types (instances minus duplicates), followed by team and individual walkthroughs. For both systems, the total number of usability problem types found by empirical testing was about twice the total number found by the team walkthroughs, and three times the total number found by individual walkthroughs. Again, the difference in the distribution of the total number of problem types found in the three groups was statistically significant for each system at the $p < .01$ level.

The data on PSC ratings assigned to problem areas for Systems 1 and 2 are presented in Table 3. For both systems, empirical testing identified a larger total number of significant problem areas (Total SPAs) assigned PSC ratings of 1-3 than did either team or individual walkthroughs. However, the variation in total number of SPAs across methods was statistically significant for System 1 ($p < .01$) but not for System 2. For both systems, there was no bias or tendency towards more severe ratings (i.e., more PSC 1s versus 2s) in one group as compared to another.

	Empirical Test	Team Walk	Individual Walk
System 1			
PSC 1	19	9	8
PSC 2	18	13	9
PSC 3	3	1	1
Total SPAs	40	23	18
No Action Areas	7	24	29
Total Problem Areas	47	47	47
System 2			
PSC 1	10	3	6
PSC 2	15	10	10
PSC 3	2	1	1
Total SPAs	27	14	17
No Action Areas	16	29	26
Total Problem Areas	43	43	43

Table 3. PSC ratings of usability problem areas.

Table 4 provides information on the number of unique usability problem areas identified by each of the methods. A problem area that is unique to a method is a SPA that is identified by only one method. For Systems 1 and 2, empirical testing identified the largest number of unique problem areas. For both systems, two-thirds or more of these unique problem areas were assigned a PSC rating of 2, representing relatively important problems in the user interfaces.

A common usability problem area across methods occurred when a SPA was identified by all three methods. For example, a common problem area was the basic model for linking, where empirical testing and team walkthroughs generated a SPA with a PSC rating of 1 and individual walkthrough generated a SPA with a

PSC rating of 2. To analyze the proportion of problem areas that were common, the total number of SPAs for each system was computed. The total number of SPAs identified for System 1 was 41 (40 identified by empirical testing plus 1 unique problem area identified by team walkthrough). The total number of SPAs identified for System 2 was 29 (27 identified by empirical testing plus 2 unique problem areas found by individual walkthrough). Regarding common problem areas for System 1, 13 of the total number of 41 SPAs (32% of total) were common across the three techniques. For System 2, 10 of the 29 SPAs (35% of the total) were common across techniques.

	Empirical Test	Team Walk	Individual Walk
System 1	13	1	0
System 2	8	0	2

Table 4. Unique usability problem areas.

Walkthrough Results

Additional analysis of the effectiveness of team as compared to individual walkthroughs was conducted by studying the total number of problem tokens found by each individual walkthrough evaluator or walkthrough evaluator team. This analysis showed that teams found more problem tokens than did individual walkthrough evaluators for each system ($p < .01$ according to t-tests). For System 1, the average number of problems identified by the walkthrough teams was 19 while the average for individual walkthrough evaluators was 13. For System 2, these values were 18 for team and 11 for individual walkthroughs respectively. However as shown in Tables 2 and 3 above, while more problem tokens and types were identified by team walkthrough conditions, the total number of SPAs identified was similar for both team and individual walkthroughs, and the pattern held across systems.

During the debriefing, evaluators rated the relative usefulness of scenarios as compared to self-guided exploration in identifying usability problems in the systems. The evaluators used a 5-point scale where a score of 1 was the most positive response for use of scenarios. All walkthrough groups favored the use of scenarios over self-guided exploration; the average score across systems was 1.8. Evaluators were also asked about the added value of using the guidelines during the walkthrough. A 5-point scale was again used and a score of 1 was the most positive response for guidelines. For both systems, the walkthrough evaluators thought the guidelines were of limited added value to them in identifying usability problems; the average score across systems was 3.9. The evaluators said they thought the brief document was very effective in explaining and giving examples of the guidelines, and that they would not change the format. They stated that because of their experience with GUI and other systems, they were already familiar with the guideline concepts, but that less experienced users would find it very useful. It was noted that almost all evaluators tried to take the guideline document with them at the end of the session. When asked about this, they said they were very pleased with it and wanted to keep it for reference.

Cost-Effectiveness Data

Table 5 shows the cost-effectiveness data for the three methods on the two systems. This analysis includes the time required by human factors staff and participants; no laboratory facility costs are included. Human factors time includes preparation of all materials (35-45 hours across methods), administration of sessions (10-55 hours), and data analysis (16-50 hours). Time to analyze the data using the generic model of problem areas and the PSC matrix are included for all groups. As expected, the total hours required (human factors plus participant) for a method was highest for empirical testing for System 1 and System 2.

	Empirical Test	Team Walk	Individual Walk
System 1			
HF staff hours	136	72	70
Participant hours	24	48	24
Total hours	160	118	94
Problem Types	159	68	49
Hours/Type	1.0	1.7	1.9
SPAs	40	23	18
Hours/SPA	4.0	5.1	5.2
System 2			
HF staff hours	116	77	76
Participant hours	24	48	24
Total hours	140	125	100
Problem Types	130	54	39
Hours/Type	1.1	2.3	2.6
SPAs	27	14	17
Hours/SPA	5.2	8.9	5.9

Table 5. Cost-effectiveness data for the three methods.

However, empirical testing needed only about half as much time as the walkthroughs to find each usability problem type. For System 1, the hours required to identify each significant problem area were fairly similar across techniques. For System 2, the resource required for team walkthrough was higher than for both empirical testing and individual walkthrough.

DISCUSSION

The findings regarding the relative effectiveness of empirical testing and walkthrough methods were generally replicated across the two GUI systems. It is not clear whether these patterns would be replicated on non-GUI systems, however, the significant differences in the style and presentation of the two GUI systems in the study support the reliability of the results across these types of systems.

The empirical testing condition identified the largest number of problems, and identified a significant number of relatively severe problems that were missed by the walkthrough conditions. These data are consistent with Desurvire et al. [4] data and at odds with Jeffries et al. [10] data. The difference between our data and

those of Jeffries et al. [10] in the number of problems found might be explained by the difference in evaluator expertise, but Desurvire et al. [4] also utilized UI experts, and their data are consistent with ours. All three studies do provide strong support for the value of UI expertise though. We recognize that the basis of our empirical usability testing results was the experimental controls employed, the skills required to conduct the test, experience with the two GUI systems prior to observing test sessions, and the UI expertise required to recognize and interpret the usability problems encountered by the users. Our data suggest that this type of empirical usability testing should be employed for baseline and other key checkpoint tests in the development cycle where coverage of the interface and identification of all significant problems is essential. Walkthroughs of the type in this study are a good alternative when resources are very limited [19] and may be the preferred method early in the development cycle for deciding between alternative designs for particular features. In Jeffries et al. [10] the heuristic method found a larger number of severe problems than usability testing. This might be explained partially by the differences in procedures. Data for the methods in our study were collected across a three-hour time period. In the Jeffries et al. [10] study, the UI experts in the heuristic condition documented the problems they found over a two-week period.

About a third of the significant problem areas identified were common across all methods. The degree of overlap is encouraging, but it should caution human factors practitioners about the tradeoffs they are making in employing one method rather than another. These methods are complementary and yield different results; they act as different types of sieves in identifying usability problems. Jeffries [11] stated that there was less overlap in problems found by any two of the methods in their study [10] than in ours. The higher degree of overlap in our study might be partially due to the fact that all methods used the same scenarios. These scenarios were rich and complex examples of typical work that end users need to perform and may have greatly aided in the evaluation of the systems by all methods. The overlap between the two methods that used their scenarios in Jeffries et al. [10] was no higher than that between the others though, and may reflect differences in the two sets of scenarios.

Team walkthroughs achieved better results than individual walkthroughs in some areas. The fact that any differences emerged between the team and individual walkthroughs is encouraging. The brief period of the usability session, the lack of an established working relationship between team members, and the small size of the teams may have contributed to the small differences found. Many usability walkthroughs in product development are done by moderate-sized teams (e.g., 6-8 people) because of the wide range of skills and backgrounds necessary to identify and then resolve usability problems. Therefore, due to practical and organizational considerations, team walkthroughs may be an area warranting future research. Work by Bias [1] and Hackman and Morris [7] may help identify ways to facilitate and enhance the performance of team walkthroughs.

All walkthrough groups favored the use of scenarios over self-guided exploration in identifying usability problems. This evidence supports the use of a set of rich scenarios developed in consultation with end users. And as evaluation work attempts to predict what will occur in real world settings, the use of well-founded scenarios can provide some assurance that real world problems will be identified.

The evaluators, who were all experienced GUI users, generally thought that the guidelines for usability were of limited added value to them in conducting the walkthroughs, but would be helpful for less experienced users. These data are consistent with the Desurvire et al. [4] results showing no difference between UI experts' heuristic and "best guess" ratings. Guidelines may serve to promote consensus about usability goals for development teams, and may be more useful for less experienced evaluators during walkthroughs.

The results also demonstrate that evaluators who have relevant computer experience and represent a sample of end users and development team members can complete usability walkthroughs with relative success. Specific UI or human factors expertise may be very helpful but is not required, and there are a multitude of practical, individual, end user, organizational, and product benefits to be achieved by involving more members of development teams in walkthroughs [5, 12, 19, 20].

Cost-effectiveness data show that empirical testing required the same or less time to identify each problem as compared to walkthroughs. The differences between these data and the cost-benefit data for usability test and heuristic methods in Jeffries et al. [10] may be due to the differences in the walkthrough procedures utilized and in the type of data analysis performed in the two studies. If our walkthrough data had been analyzed in other ways or by different individuals (e.g., developers), the resource required might have varied significantly. However, the resource and skills applied to data analysis may be reflected in the quality of the analysis and the resulting changes to systems. Ultimately, the true cost-benefit of these methods will be realized through their ability to facilitate the achievement of usability objectives for systems in iterative development, and to provide measurable benefits that exceed the costs of their use [13, 14, 15]. Analysis of data from one iteration in isolation is of limited utility.

The identification of usability problems is not an end in itself. Rather, it is a means towards eliminating problems and improving the interface. The part of the development process concerned with making recommendations for change based on the usability problems identified is not covered in this study. We did find that the larger the number of problem tokens and types identified regarding a significant problem area of the interface, the richer the source of data was for forming recommendations for changes to improve that portion

of the interface. The data from this study show that the empirical and team walkthrough conditions have the advantage over the individual walkthrough conditions in this area. The empirical test data contained four times as many problem tokens describing a significant problem area and providing context about it as compared to team walkthrough data, and teams produced 33-50% more information as compared to individual walkthroughs. The quality of the data analysis completed and the recommendations that arise from them are issues for future research.

How could walkthrough methods be improved? Users in the empirical testing sessions were given opportunities to provide recommendations for changes to the usability problems they encountered, and the walkthrough sessions could be improved to capture evaluator recommendations as well. Another area of walkthrough procedures that needs attention is the difficulty of interpreting problems. Walkthrough evaluators used different language than the staff who analyzed the problem reports, and the data analysts' job of understanding these problem statements was made more difficult by a lack of context and lack of session observation. The difficulty experienced in interpreting walkthrough data was supported by the lower inter-rater reliability data reported for walkthroughs compared to usability tests. Moreover, from walkthrough sessions that human factors staff observed, it became evident that evaluators misattributed the sources of problems. We also observed that evaluators sometimes became so involved in the task scenarios that they forgot to document problems they encountered and identified. We attempted to overcome this demand characteristic of the walkthroughs by emphasizing the importance of problem identification over task completion, but it was not effective in some cases, and further refinement of intervention strategies should be explored [7]. A better debriefing of evaluators that included reviewing identified usability problems, capturing undocumented ones that evaluators mentioned in passing, and collecting evaluator recommendations for changes might improve walkthrough effectiveness.

ACKNOWLEDGEMENTS

The authors thank Amy Aaronson, Catalina Danis, Tom Dayton, John Gould, Robin Jeffries, John Karat, Wendy Kellogg, Jakob Nielsen, John Richards, Kevin Singley, Cathleen Wharton, and the anonymous reviewers for comments on earlier versions of this paper.

REFERENCES

1. Bias, R.G. Walkthroughs: Efficient collaborative testing. *IEEE Software* , 8, 5, 1991, pp. 94-95.
2. Bellotti, V. Implications of current design practice for the use of HCI techniques. In Jones, D.M. and Winder, R., (Eds.), *People and Computers IV.* Cambridge University Press, Cambridge, 1988, pp. 13-34.
3. Carroll, J., Smith-Kerker, P.L., Ford, J.R., and Mazur-Rimetz, S.A. The minimal manual. *Human-Computer Interaction*, 3, 1987-88, pp. 123-153.
4. Desurvire, H., Lawrence, D., and Atwood, M. Empiricism versus judgment: Comparing user interface evaluation methods on a new telephone-based interface. *SIGCHI Bulletin*, 23, 4, pp. 58-59.
5. Gould, J.D., and Lewis, C. Designing for usability: Key principles and what designers think. *Communications of the ACM*, 28, 1985, pp. 300-311.
6. Grudin, J. and Poltrock, S.E. User interface design in large corporations: Coordination and communication across disciplines. In *Proceedings of CHI'89* (Austin, TX, April 30-May 4, 1989), ACM, New York, pp. 197-203.
7. Hackman, J.R., and Morris, C.G. Group tasks, group interaction process, and group performance effectiveness: A review and proposed integration. In Berkowitz, L., (Ed.), *Advances in Experimental Social Psychology*, Vol. 8, Academic Press, New York, 1975.
8. International Business Machines Corporation. *Systems Application Architecture, Common User Access, Guide to User Interface Design*, (SC34-4289), 1991.
9. International Standards Organization. *Working Paper of ISO 9241 Part 10, Dialogue Principles, Version 2* , ISO/TC 159/SC4/WG5 N155, 1990.
10. Jeffries, R.J., Miller, J.R., Wharton, C., and Uyeda, K.M. User interface evaluation in the real world: A comparison of four techniques. In *Proceedings of CHI'91*, (New Orleans, LA, April 28-May 3, 1991), ACM, New York, pp. 119-124.
11. Jeffries, R. J. Personal communication, Sept. 1991.
12. Jorgensen, A.K. Thinking-aloud in user interface design: A method promoting cognitive ergonomics. *Ergonomics*, 33, 4, 1990, pp. 501-507.
13. Karat, C. Cost-justifying human factors support on development projects. *Human Factors Society Bulletin*, 35, 3, 1992, pp. 1-4.
14. Karat, C. Cost-benefit and business case analysis of usability engineering. *ACM SIGCHI Conference on Human Factors in Computing Systems*, New Orleans, LA, April 28-May2, Tutorial Notes.
15. Karat, C. Cost-benefit analysis of usability engineering techniques. In *Proceedings of the HFS Society*, (Orlando, FL, Oct., 1990), pp. 839-843.
16. Lewis, C., Polson, P., Wharton, C., and Rieman, J. Testing a walkthrough methodology for theory-based design of walk-up-and-use interfaces. In *Proceedings of CHI'90* (Seattle, WA, April 1-5, 1990), ACM, New York, pp. 235-242.
17. McGrath, J.E. *Groups: Interaction and Performance* Prentice-Hall, Englewood Cliffs, N.J., 1984.
18. Nielsen, J. Finding usability problems through heuristic evaluation. In *Proceedings of CHI'92* (Monterey, CA, May 3-7, 1992), ACM, New York.
19. Nielsen, J. Usability engineering at a discount. In Salvendy, G., and Smith, M.J., (Eds.), *Designing and Using Human-Computer Interfaces and Knowledge-Based Systems.* Elsevier Science Pulishers, Amsterdam, 1989, pp. 394-401.
20. Nielsen, J. and Molich, R. Heuristic evaluation of user interfaces. In *Proceedings of CHI'90*, (Seattle, WA, April 1-5, 1990), ACM, New York, pp. 249-256.
21. Whitten, N. *Managing Software Development Projects: Formula for Success.* Wiley and Sons, New York, 1990, pp. 203-223.
22. Wharton, C., Bradford, J., Jeffries, R., and Franzke, M. Applying cognitive walkthroughs to more complex user interfaces: Experiences, issues, and recommendations. In *Proceedings of CHI'92* (Monterey, CA, May 3-7, 1992), ACM, New York.
23. Wright, P.C., and Monk, A.F. The use of think-aloud evaluation methods in design.*SIGCHI Bulletin*, 23, 1, 1991, pp. 55-57.