A Comparison of Usability Assessment Methods Applied to the U.S. Navy's Performance Management and Appraisal System

Elizabeth Dean
Michael Schwerin
Kimberly Robbins

RTI International
3040 Cornwallis Road
P.O. Box 12194
Research Triangle Park, NC 27709
edean@rti.org

Introduction

Usability testing is an important method for examining how system users understand and use a system to complete a specific task. This research method is used by system developers to test and evaluate automated tools ranging from information systems, web-based questionnaires, and websites. RTI International recently implemented a large-scale, multi-site usability testing effort to assist in the development of the U.S. Navy's new web-based performance management system. As part of this testing effort, several methodological variations of usability testing were applied. This paper describes the relative utility of these usability methods and examines the differences that resulted from each. The data presented in this paper include behavior coding of usability tests, timing estimates from usability tests, and subjective data from pre- and post-test user surveys and focus group results.

The U.S. Navy's Usability Testing Requirement

As a result of the 2001 Executive Review of Navy Training, ordered by the Chief of Naval Operations (CNO), the U.S. Navy created the Task Force for Excellence through Commitment to Education and Learning (EXCEL). Task Force EXCEL's goal is to identify new ways for the U.S. Navy to train, grow, place, and utilize personnel who maximize the Navy's ability to accomplish its military mission while developing a more productive yet satisfying workplace. One working group, or "vector," of Task Force EXCEL is the Performance Vector, which was charged with an examination of the Navy performance management and appraisal system. After reviewing the existing system of performance review, the Performance Vector recommended the development of an electronically based performance management and appraisal system.

Within the Department of Defense (DoD) and the Department of the Navy (DoN), there is a growing emphasis on the importance of human system integration (HSI) in the development of new systems for military personnel. While HSI evaluations are routinely integrated into training systems, it is unclear whether HSI is a critical part of the system development process for manpower and personnel systems. The Undersecretary of Defense for Acquisition, Technology, and Logistics (USD AT&L) recently issued DoD Instruction 5000.2 (DoD, 2003) that specifically calls for DoD acquisition program managers to ". . . ensure human factors engineering/cognitive engineering is employed during systems engineering over the life of the program to provide for effective human-machine interfaces and to meet HIS requirements. Where practicable and cost effective, system designs shall minimize or eliminate system characteristics that require excessive cognitive, physical, or sensory skills; entail extensive training or workload-intensive tasks; result in mission-critical errors; or produce safety or health hazards." (Enclosure 7, paragraph E7.1.1, p43) It is clear that it is DoD's intent to ensure that all systems with a human-machine interface – including manpower and personnel systems – are tested for ease of use and that acquisition program managers need to consider system usability through the life cycle of system development. Hence, as the Commander, Navy Personnel Command (CNPC) began implementing the new performance management and appraisal system, the final performance feedback and appraisal forms required usability testing with Navy personnel to meet the requirements of DoD Instruction 5000.2; that is, the system must be tested in order to reduce cognitive, physical, and sensory burden; minimize necessary training and additional workload; and avoid errors and safety and health hazards. Specific questions the designers of the web-based tool faced included:

- Is the tool compatible with Navy culture?
- Does the tool meet functionality requirements?
- Are users able to navigate the system?

- Is the system equally usable for a wide range of Navy personnel (from junior enlisted to senior officers)?
- Is the system operable in ship and shore-based environments?
- Will the system result in increased job satisfaction and a greater perception of fairness?

To answer these questions, the goals of the Navy usability testing study were to capture data from participants to identify potential sources of error and burden in the new performance management and appraisal system. Specifically, RTI was required to

- Conduct usability tests with non-supervisory and supervisory Navy personnel with a range of skill sets in a variety of work environments
- Collect data on the type and frequency of user errors and the time required to complete key tasks using the system
- Collect self-reported data on users' comfort using the system, perceptions of effectiveness of the system, satisfaction with the system, and ease of using the system
- Conduct focus group interviews with users after they completed usability tests, to identify features they liked and recommendations for improvement.

This variety of assessment methods (usability testing, self-administered surveys, and focus groups) was employed to ensure that CNPC obtained a broad view of the types of problems that might occur with a Navy-wide implementation and especially to assess the likely impact of the shift in performance management and appraisal on Navy culture and Sailor morale.

Best Practices in Usability Testing

Nielsen (1993, p. 165) describes usability testing as "the most fundamental usability method" and "irreplaceable" because it is the only mechanism that allows the researcher to obtain direct, detailed information on the user's experience with the product or tool being tested. Every usability test has the primary goal of improving the usability of the end product, coupled with specific goals and concerns, such as the needs of a particular group of users or the implementation of the product in a particular environment (Dumas & Redish, 1993). For the Navy usability testing study, specific concerns were the differential needs of non-supervisors and supervisors as well as shipboard and shore-based personnel. According to Dumas and Redish, in any usability test, the following four key factors must be present:

- The participants represent real users.
- The participants do real tasks.
- The usability researcher observes and records what participants do and say.
- The usability researcher analyzes the data, diagnoses the problems, and recommends changes to fix the problems.

Usability researchers agree that multiple methodologies can be used to effectively assess the user experience. In fact, most usability test plans include several types of data collection. Other methods not included in this study include surveys of user needs, participatory design experiences, heuristic evaluations, task analysis, and paper prototyping. In the usability testing literature, the two most consistently emphasized assessment practices are an iterative design and consideration of user context.

In a survey of usability researchers, Nielsen (1993) identifies the six most effective methods for usability improvement. Iterative design (tied with task analysis) is the most important consideration. There are several reasons why iterative design of usability tests is so important. Changes to a system as a result of usability testing sometimes do not solve a problem. In fact, new solutions may create new problems. Furthermore, new solutions may reveal additional problems that were previously hidden or outbalanced by the original problem identified. Nielsen's research analyzing the effectiveness of iterative testing finds a median improvement in system usability, defined by the usability metrics employed for the particular test plan, of 38% per iteration. While five out of 12 iterations in Nielsen's analysis showed that one dimension of usability had gotten worse, significant improvements in usability continued to be made in later iterations.

In the early days of usability testing (the 1970s and 1980s), the norm was one large-scale test of 30 users, conducted very late in the design process when most of the design features were stabilized and thus averse to change. The problem with this approach was that it found pervasive system problems, but at a stage in the development cycle

where it was too late to fix them. In addition, 30 users were not needed to identify such large and pervasive problems. The solution adopted was to test earlier prototypes of systems, even using paper prototypes when necessary, with multiple iterations of five to 10 users. This approach allows early identification of large-scale systemic problems. Since 1990, iterative testing with small samples has been the preferred approach (Dolan & Dumas, 1999).

Valid usability measurement cannot take place outside the user's context, and usable systems require this context be incorporated into the development cycle. When considering tools such as guidelines and checklists for user-centered design, Bevan and Macleod (1994) warned against dependence on checklists, because guidelines for usable system features need extensive detail to be useful, but if checklists are detailed enough, they are likely to be too specific to apply in multiple real-world contexts. For example, a highly interactive web-based performance management evaluation form that requires frequent communication with a server to complete may be desirable in an office setting because it will allow the user's data to be saved through many interruptions. Conversely, this approach may not be desirable on board a deployed Navy ship, since the satellite Internet connection may be unavailable or regularly interrupted. The solution is to conduct scenario-based assessments that reflect the environments of real users. A true-to-life environment can be replicated in a lab setting, but the most realistic approach is to conduct on-site usability testing in the field. Bevan and Macleod add a fifth factor to Dumas and Redish's list above: The participant's real-life context is represented in the usability test.

This evaluation of the Navy's web-based performance management and appraisal system incorporated the two key design features of iterative testing and context awareness. Usability tests were conducted at three very different Navy installations with time between iterations to make changes to the system. Our data collection approach is described in the next section.

Data Collection Approach

Data collection took place in three iterations at three different sites. At each location, approximately 20 participants took part in the pretest and posttest surveys, the usability tests, and the focus group interviews. Each participant completed all of the assessment protocols (pretest and posttest surveys, usability tests, and focus group interviews), except in rare situations where unforeseen circumstances or work schedules kept them from completing a full cycle. Each site's group was composed of about half non-supervisors and half supervisors.

Iteration 1 took place at Naval Air Station (NAS) Brunswick in Brunswick, Maine. NAS Brunswick is an aviation community in the Atlantic Fleet. A total of 21 personnel took part. Of these 21, 14 were supervisors, and seven were non-supervisors. Ten participants were NAS Brunswick personnel, 10 were squadron personnel, and one participant was from a ship pre-commissioning unit. Iteration 2 took place aboard the USS KITTY HAWK (CV 63), an aircraft carrier in Yokosuka, Japan. Twenty personnel participated. Of the 20, 14 were supervisors and 6 were non-supervisors. Participants were members of the Air Department, Air Immediate Maintenance Department, Operations Department, Combat Systems Department, Executive Officer Administration, Supply, and Weapons. Iteration 3 took place at the Trident Training Facility at Naval Base Kitsap in Bangor, Washington. Naval Base Kitsap is a submarine community in the Pacific Fleet. A total of 19 personnel participated. Of the 19, 10 were supervisors and 9 were non-supervisors. Participants were mostly from submarine commands, including the USS ALABAMA (SSBN 731), USS ALASKA (SSBN 732), USS NEVADA (SSBN 733), and the USS KENTUCKY (SSBN 737).

Data collection instruments included usability test scenarios, surveys, and focus group guides. Usability test scenarios were developed to evaluate the effectiveness of screen layouts, performance appraisal item structures, and on-screen features of the performance management and appraisal system. Usability test scenarios required users to complete actual tasks that personnel are likely to encounter. Some example tasks included:

- Log into the performance management system
- Open your performance document
- Complete your performance appraisal
- Check spelling
- Submit the performance document.

The usability tests were conducted on desktop computers and recorded using a portable usability lab – a coordinated system of digital audio and video capture equipment. Users' facial expressions and on-screen activity were

recorded. After the usability tests, coders reviewed the digital videotapes and logged the errors and problems encountered as well as the durations for each task.

The paper-and-pencil self-administered surveys were developed to obtain Sailors' subjective impressions of the performance management and appraisal system. The pretest survey included demographic measures, frequency of computer use at home and work, satisfaction with the current performance management and appraisal process, and expectations regarding difficulty of using the system. The posttest survey asked participants to report their impressions after completing the test scenarios. Items included overall satisfaction with the system, comfort completing the tasks, perceived success completing the tasks, ease of use, appearance and efficiency of the system, and retest measures of satisfaction with the current performance management and appraisal process.

The focus group interviews were conducted as a means of identifying features of the performance management and appraisal systems that users liked or features that needed to be improved. Following the conventions and best practices of conducting focus group interviews (Edmunds, 1999; Morgan, 1997; Krueger, 2003), 6-10 participants made up each group. Within each group personnel had similar performance management and appraisal responsibilities and work environments. The interviews were designed to capture qualitative information, centered on aspects of the usability testing process, such as the tasks listed above. A final question asked for a summary rating of the system participants used, reasons for that rating, and what might be done to improve users' overall ratings.

In addition to using a variety of data collection methods, several different testing methods and analytic styles were used to obtain a comprehensive picture of the usability of the performance management and appraisal system. When analyzing the usability test data we examined both timing records and counts of the number of problems users experienced in the system. Each type of analysis contributed different information. Timing data assessed user burden while problem incidence assessed poor design and generated suggestions for improvements. Additionally, we combined subjective and objective measures of usability. That is, we intended to capture self-reported perceptions of usability of the system (in the focus groups and surveys) as well as observed user errors and successes (in the usability tests). Third, when conducting the usability tests, we employed two different data collection techniques – conversational (think-aloud) testing and task-oriented testing. The use of both techniques was designed to balance the needs for user feedback and neutral observation of the user experience. Below we present the results of implementing each of the above three methodologies.

Results

When analyzing the usability test data, we used two independent variables: supervisor status and test site location. These were key variables of interest to the system designers. Initially, we also examined current paygrade, years served in the Navy, education, gender, race and ethnicity. Ultimately, little variation was observed in those variables so they were excluded from the final analysis.

The tables presented in this section and later sections show the results of significance testing. Two analytic techniques were used – analysis of variance (ANOVA) and chi-square tests of significance. ANOVA was used for continuous variables, such as task time and error frequency, and for ordinal variables, such as survey questions using a five-point agreement scale. For categorical variables, we used a contingency table. The differences in continuous and ordinal variables within the demographic analyses were investigated with a Bonferroni *t*-test that accounts for multiple comparisons. Since the Bonferroni *t*-test is a more stringent test of significance for between-group mean score comparisons, Tukey's *t*-test was used to determine if a less stringent test would affect the results. Tukey's *t*-test for group comparisons produced the same results. For categorical variables we used chi-square tests of significance for group differences. Although these tests require random, normally distributed samples, cautiously applying this statistical test to convenience samples is a common practice in the usability testing literature (e.g., Westerman, 1997; Wiedenbeck, 1999; Norman *et al.*, 2000).

In usability testing, researchers typically manipulate experimental usability stimuli to compare the effect of system usability between groups or between conditions. While this may be a subject of study in a follow-up full-scale pilot study, the objective of this study was to examine system usability in a group of potential system users. As a result, no experimental effects were examined, but rather usability was examined between user groups (i.e., supervisors and non-supervisors, and users at different geographic locations).

Given these two constraints, the interpretation of the results should take into account the following points. First, the findings may not be generalized to either the general population or to the Navy population. Generalization may be

possible only through large-scale studies employing probability samples of target populations. Second, since this study did not have experimental and control conditions, the associations between the independent and dependent variables should be viewed as correlational rather than causal.

Analyses of Task Durations vs. Usability Errors

Usability test results were examined along two dimensions: the time users took to complete specific tasks and the number and types of errors that were observed while users were completing specific scenarios. The duration analysis provides estimates of the user burden (the time users spent on tasks) of the performance management and appraisal system. The usability errors analysis provides information on the errors and poor design features of the system that may have led to the increased durations in completing tasks.

The average durations for each task in completing a self-evaluation, or Human Performance Feedback and Development (HPFD) document, in the performance management and appraisal system by supervisor status and by location appear in Table 1. Results indicated the tasks that took the longest were completing the tutorial (an average of 1633.6 seconds) and completing the HPFD document (an average of 716.3 seconds). These results are not surprise, given that the tutorial is designed to train Sailors on all aspects of the performance management and appraisal system, and was expected to take up to 2 hours to complete. Additionally, completing the HPFD document is the central task of using the system for any Sailor. It is a fairly involved self-appraisal with a series of different dimensions on which the Sailor must rate him- or herself and write a brief explanation of the ratings.

Results indicate few statistically significant differences across groups: logging into the system (Task 2), finding the "target behaviors" description (Task 6), and changing ratings and cutting and pasting comments (Task 7).

- Sailors on the USS KITTY HAWK (CV 63) had the longest durations for logging into the system, an average of 539.5 seconds compared to 247.5 seconds at Naval Base Kitsap – Bangor and 135.6 seconds at NAS Brunswick. The extended login durations for the USS KITTY HAWK (CV 63) appears to be the result of multiple server problems at the site. The server was frequently down and users could not log in. Additionally, the ship shifted from wire to satellite communications on Day 4 of data collection, as the ship prepared to go to sea the following week.
- There is no apparent reason for the differences in time for finding the "Target Behaviors" button shown by Sailors at NAS Brunswick (31.3 seconds) compared to the USS KITTY HAWK (CV 63) (70.4 seconds) and Naval Base Kitsap – Bangor (71.4 seconds) participants.
- The difference in time spent collapsing sections of the document is not apparent among locations but is visible between supervisors and non-supervisors. Supervisors took an average of 75 seconds to change ratings and cut and paste, whereas non-supervisors took an average of 120.6 seconds.

Overall, differences between supervisors and non-supervisors and among the three sites are minimal.

There are some limitations of the burden estimates. Burden estimates only tell part of the story for the analyst; they do not point to the cause of burden. Hence, behavior coding of all usability tests was conducted. Table 2 illustrates total frequency and average occurrence of key errors within each task. The most common types of errors or problems included not following screen instructions, inability to set passwords, navigational errors, and the need to refer to the Quick Reference Guide (QRG). The QRG is a "cheat sheet" developed after iteration 1 to help users navigate through particularly challenging or unfamiliar tasks (e.g., finding and opening the HPFD document). The results in Table 2 supplement the durations displayed in Table 1 with a more qualitative picture of what it was like for a participant to use the system. Some examples of this kind of information include:

- Many of the tasks were so difficult at times that the Sailor had to ask for help to complete them (all but Task 1). This difficulty suggests that the system design was counterintuitive in parts and did not conform to users' expectations about how it would operate.
- Most tasks required more instructive information, as the relatively high rate of use of the QRG indicates.
- Problems setting a new password and problems following screen instructions in the tutorial were only encountered in their respective tasks but tended to occur far more frequently than some of the other problems that occurred across tasks.

Table 1. Estimate of Average Time to Complete Usability Testing Task by Task[1, 2]

Task Description	(n)	Overall	Supervisor Status		Location		
			Supervisor (S)	Nonsup. (NS)	Naval Base Kitsap (K)	USS KITTY HAWK (Y)	NAS Brunswick (B)
Task1: Complete the CBT tutorial	26	1633.6	1822.5	1444.6	1480.1	1797.1	1641.7
Task2: Log in to system	38	323.8	317.0	332.4	247.5	539.5 [B]	135.6 [Y]
Task3: Open the HPFD document	45	212.0	209.8	216.1	209.8	213.6	213.1
Task4: Complete the HPFD document	51	716.3	732.3	695.1	742.9	633.4	747.9
Task5: Check spelling	38	88.1	88.3	87.8	78.0	103.8	85.6
Task6: Find the "Target Behaviors" description	36	50.1	45.4	56.6	70.4 [B]	71.4 [B]	31.3 [Y]
Task7: Change ratings and cut and paste comments	35	89.3	75.0 [NS]	120.6 [S]	102.6	88.8	80.4
Task8: Collapse all sections of the document	41	32.1	30.5	34.5	45.4	30.5	23.1
Task9: Submit the HPFD document	38	63.7	73.7	44.5	58.9	85.0	55.4
Task10: Enter a performance note	39	143.5	135.1	154.4	168.7	175.0	108.2

[1] Time was measured by second.

[2] Tasks 11 through 19 were completed only by study subjects with supervisor status. Therefore, these tasks were excluded from the analysis.

Note: Superscripts [S], [NS], [K], [Y], [B] indicate significantly different estimates at the 0.05 level from *t*-test. Bonferroni *t*-test was used to account for multiple comparisons for location variable.

Table 2. Frequency and Average Number of Errors by Usability Scenario Task

Problem Category	Non-Supervisors		Supervisors	
	Total Number of Incidents Across Sessions	**Average Number of Incidents Per Session**	**Total Number of Incidents Across Sessions**	**Average Number of Incidents Per Session**
Task 1: Complete the CBT Tutorial				
User does not follow screen instructions	125	17.8	170	24.2
Tutorial button error	16	5.33	42	6
Task 2: Log in to NSIPS				
User is not able to set new password	21	3	16	2.28
User asks for help	7	1.16	11	1.57
System or server error	7	2.33	7	1.4
User refers to QRG	3	3	1	1
Task 3: Open the HPFD document				
Navigational error	25	3.57	19	2.71
User refers to QRG	21	3	14	2
User asks for help	8	1.14	7	1.4
System or server error	6	1	7	1.16
User retries action because the system did not react the first time	5	1.66	1	1
Task 4: Complete the HPFD document				
User refers to QRG	22	3.14	6	2
User asks for help	11	1.83	9	1.8
General button error	6	1.2	2	1
User is timed out	4	1	5	2.5
Navigational error	7	2.33	1	1
Task 10: Enter a performance note				
Navigational error	16	2.28	9	2.25
User refers to QRG	12	1.71	1	1
User is timed out	4	1	2	1
User asks for help	3	1	1	1
User retries action because the system did not react the first time	3	1.5	9	4.5

In sum, the task duration analysis provides information on the relative burden of each task in the usability scenarios, whereas the usability error analysis provides more information about the sources of burden. Additional analyses of usability errors provided more information. For example, estimates of the total error frequency per task, and rate of error occurrence per task supplement the overall burden estimates of the duration analyses. (For more details on these results, see Dean *et al.*, 2004.)

Subject and Objective Results

Data collected from subjective measures were reported from the pretest and posttest surveys and in the posttest focus group interviews. Data from objective measures were obtained from behavior coding and timing analyses of the usability test scenarios. As described above, the behavior coding and timing analyses provided key insights into the burden associated with the most difficult components of the performance management system. Survey and focus group interview data supplemented this information with details on a different aspect of implementing the new performance management system—the effect of the change on Navy culture. Subjective results included recommendations to incorporate a paper QRG to assist with navigating the performance management system, ensure that one-on-one personal counseling sessions are maintained with the transition to an automated system, and develop a communication plan for supervisors on guiding their direct reports on the implementation of the system.

Supervisory and non-supervisory personnel at each of the three sites participated in focus group interviews and were asked about the phases involved in using the system to complete the HPFD document. Table 3 displays results of the focus group interview data. In this paper, we will only present the supervisors' results. (For more information on results for both the supervisors and non-supervisors, see Schwerin *et al.*, in press.) Supervisors were also asked about completing an ePerformance document. The ePerformance document is a performance management document in which a supervisor evaluates his or her direct reports. Hence, non-supervisors were not required to use the ePerformance document. Those results are not presented here, since the ePerformance document was functionally similar to the HPFD document. We discuss the ePerformace document in the next section, on conversational vs. task-oriented interviewing.

The "Completing the HPFD form" task generated the greatest number of user comments. Participants provided general positive comments: "The form was easy to complete and fill in the block with comments; it was simple and efficient." The terminology used with the performance dimensions were "well organized and worded logically" and reflected "phrases I'd use in the written text of a FITREP [senior enlisted and officer performance appraisal] or EVAL [junior to mid-grade performance appraisal]." When asked for features of the HPFD system that could be improved, supervisors cited the lack of an electronic or paper back-up option, overall system performance, the performance dimensions, and concerns with the HPFD process. Supervisors felt that a paper back-up copy would be helpful to ensure against document loss in the event of system failure and provide a form that Sailors could use when Internet access is unavailable.

While the terminology used for the performance dimensions was mentioned as a positive feature of the HPFD process, several supervisors felt the terminology was "too civilian" and that several of the performance dimensions seemed redundant. Additionally, concern about how the HPFD process would be implemented in a web-based environment led to comments stating "I want to do a face-to-face counseling with Sailors first, complete the counseling document and send it to the Sailor for review, and they can call if they have any problems." Participants at each testing site echoed this sentiment.

The "other concerns" segment, which concluded the focus groups, generated the second highest number of comments. A number of implementation and procedural concerns throughout the entire performance management process emerged: at what level in the organization would the raters be given the responsibility for the performance appraisal, the notion of raters' comments remaining unchanged as the appraisal is routed up the chain of command, concerns of procedural fairness and grade inflation, and the impact of removing the reporting senior's promotion recommendation.

Other tasks generated fewer but still substantial numbers of comments. Most participants expressed dissatisfaction with the tutorial. The most frequently cited criticism was a general lack of simplicity in relaying the information about the system. As a result of comments in the first iteration of the study (i.e., "Sailors work from checklists – they're used to that; the training should be set up like a checklist to work people through the training"), the research team developed the QRG. Supervisors generally found this useful, citing that it was "critical" and "easy to

understand." Participants stated that the QRG could be improved if functions and critical features of the program were highlighted and presented in a sequential fashion that led participants through the major tasks.

When asked about logging into the system, several supervisors reported that it was an easy process when the web system, Navy Systemwide Integrated Personnel System (NSIPS), was functioning. NSIPS had recurring interruptions in service, which led to a great deal of frustration among participants. Beyond the system challenges, participants cited two aspects of the password change process that should be improved: making the password non-case sensitive and outlining the password requirements when users are first asked to reset passwords (i.e., one uppercase character, one lowercase character, minimum password length, special characters, and numeric characters). As one participant commented, "The rules for password generation should be published." Frequently users learned that their passwords did not satisfy the password requirements only after several trial-and-error attempts.

When closing the HPFD document, several supervisors repeated concerns mentioned in previous phases of system use (i.e., electronic or paper back-up copies, text box limitations). Other concerns included confusion between the "Save" and "Complete" buttons and questions pertaining to document routing and processing. Users reported uncertainty as to what differentiated the "Save" and "Complete" functions. Several participants recommended a confirmation screen that prompts system users to confirm their intent (i.e., "Pressing 'Yes' will save the document and forward it to your supervisor for review and processing.")

In addition to the focus groups, other subjective measures were the pretest and posttest surveys. The surveys were designed to collect information expected to relate to the users' perceptions of the web-based performance management system. Questions addressed topics such as the participants' experiences using computers and their overall impressions of the test version of the web-based tool. Over 85% of participants said they either used a computer at work every day or that most of their work was done on a computer. Furthermore, at home, nearly 37% of participants reported using a computer every day, and another 58% reported using a computer "sometimes." Discrepancies exist between supervisors and non-supervisors. Approximately 32% of non-supervisors reported using a computer at work only "sometimes" or "never," compared to only 3% of supervisors while 9% of non-supervisors reported never using a computer at home, compared to 3% of supervisors.

When asked on the posttest survey about their perceptions of "overall ease of using the web-based system," nearly 23% reported that the system was "somewhat easy to use," 46% said it was "neither difficult nor easy to use," and 23% reported it was "somewhat difficult to use." Only 2% of participants thought the system was "very difficult to use." No one reported that the system was "very easy to use," an important take-away message for system designers. There appears to be significant room for improvement in the web-based tool.

Overall, the objective, tester-observed usability test results (detailed in the previous section) gave information on individual burden associated with using the system and particular errors users were likely to encounter. Details of the test results highlighted the task instructions that were not intuitive, the tasks that required more on-screen instructions, and the navigational problems with the system design. On the other hand the subjective methods elicited extremely valuable suggestions for improving the system (such as adding the QRG). The focus groups in particular gave information on cultural issues such as the language being "too civilian" and the concerns about insufficient face-to-face contact between supervisors and direct reports during the performance appraisal process. The great advantage of using the subjective data collection method was that it expanded the evaluation to more than just an assessment of the user interaction with the web-based form; it became an evaluation of the total feasibility of implementing the new performance management and appraisal procedures.

Conversational versus Task-Oriented Usability Testing

The third methodological variation we implemented was the use of both conversational and task-oriented testing. Specifically, we were interested in the differential impact on the test results and on the user experience, of using a conversational style of interview—someone who sat beside the user as he or she completed the tasks—versus using a task-oriented interview, in which the tester observed from a different cubicle or office and gave the user instructions via a desktop microphone. Sailors on the USS KITTY HAWK (CV 63) completed the tests in a conversational mode. Sailors at Kitsap Naval Base – Bangor and NAS Brunswick completed the tests in a task-oriented mode. Since the QRG, which proved to be very helpful for participants, was not used at NAS Brunswick, for our comparison of the two modes we examined only the results from the USS KITTY HAWK (CV 63) and Kitsap Naval Base – Bangor. Furthermore, we limited the analyses to supervisors only so as to limit the impact of supervisory status on the results of the testing. Finally, we examined the total durations and errors of completing the

Table 3. Supervisory Personnel—Summary of Major Themes and Frequency of Comments for Each Phase of HPFD Process

Focus Group Interview Topic	What Worked Well—Major Themes	Freq.	Opportunities for Growth—Major Themes	Freq.
Training and reference materials—CBT	General satisfaction	5	Simplicity of learning process/didn't relay document navigation information	11
			Clarity of information	3
			Relevance	1
Training and reference materials—Quick Reference Guide (QRG)	Usefulness of QRG	9	Availability and accessibility	2
			Presentation of information—relevance	2
Logging in to the system	Ease of login (when system was functioning)	6	Trouble navigating to the document	13
			Unclear password requirements	6
Selecting and opening your document	Ease of selecting the document facilitated by QRG	8	Login failures due to NSIPS problems	6
			Trouble identifying the document	5
	Document was easy to find	5	Page loading	1
			Terminology "too civilian"	1
Completing your HPFD form	General positive comments	6	Paper document back-up	10
	Performance dimensions – well phrased	5	NSIPS problems	7
	HPFD process	2	Performance dimensions – repetitive and "too civilian"	6
	Web-based form	5	HPFD process—need for "face to face" interaction	6
	Process ownership	1	Navigation and process clarification needed	6
			Text box limitations necessary	3

Table 3. Supervisory Personnel—Summary of Major Themes and Frequency of Comments for Each Phase of HPFD Process (cont'd)

Focus Group Interview Topic	What Worked Well—Major Themes	Freq.	Opportunities for Growth—Major Themes	
Closing your HPFD or ePerformance session	General positive comments	2	Text box limitations necessary	6
			Save and complete/forward documentation	5
			Form routing	4
			Paper document back-up	3
			Forced distribution/promotion summary recommendation concerns	3
			Spell check/language check	3
Other concerns	Performance notes	2	Workflow concerns: first-level rater and form routing	10
			Concerns about procedural fairness	8
			Implementation concerns	6
			Forced distribution questions	6
			Connectivity concerns	6
			Resistance to change	1

ePerformance tasks only, since these tasks were only completed by supervisors. The ePerformance tasks included finding, opening, and completing a performance appraisal document (or review document) for a direct report.

With the conversational approach, the mean completion time for ePerformance was 893.2 seconds. The task-oriented approach allowed users to complete the tasks more quickly, at an average of 773.1 seconds. However, this difference is not statistically significant at a .05 level. The difference in total number of errors yielded statistically significant results. For conversational interviews, the average number of errors across all ePerformance tasks was 10.8 errors, as compared to 2.4 errors for the task-oriented interviews. Not assuming equal variance, these differences are statistically significant at the .057 level. It seems that while the usability tests did not take longer for users with the conversational approach, that approach did uncover more usability errors. This is probably a result of the coding scheme used to identify usability errors. In addition to incorrect clicks within the web page, behaviors such as expressing confusion or frustration, or asking for help were also coded as "errors."

The total number of observations of key behaviors in conversational and task-oriented interviews appears in Table 4. For this analysis we did not look at overall errors or at specific user errors within the web-based tool. Instead we examined the occurrence of behaviors that might have been affected by using either the conversational or task-oriented style: participants requesting help or volunteering that they did not understand how to complete a task, and the participants' expressions of confusion, boredom, and frustration. Obvious facial expressions as well as verbal statements received a positive code.

Table 4. Occurrence of Specific Types of Errors by Type of Interview

	Total Error Incidence by Interview Type		
	Conversational	**Task-Oriented**	**Total Incidence of Error**
Asked for help	10 (45.5%)	12 (54.5%)	22
Said "I don't know how to do this."	2 (50%)	2 (50%)	4
Expressed confusion	9 (50%)	9 (50%)	18
Expressed boredom	7 (70%)	3 (30%)	10
Expressed frustration	5 (100%)	0	5

The results in Table 4 suggest that varying the style of interview did not have an impact on participants' willingness to ask for help or to express that they did not know how to complete a task. The two types of behaviors that perhaps varied according to interview type were expression of boredom and expression of frustration; (expression of frustration was the only analysis that yielded a statistically significant chi-square test at the .05 level). Both were more likely to occur in conversational interviews, possibly because participants felt more comfortable criticizing aspects of the system with the tester sitting nearby. Conversely, having the tester nearby made them feel less comfortable—more bored and frustrated.

The final analysis comparing the results of conversational versus task-oriented interviewing, examined the posttest score results for both types of interviews. Table 5 shows the results, which are overwhelmingly similar across interview style. Only one question looks even slightly different: "Overall, how professional or unprofessional did the system appear?" Task-oriented participants were more likely to report that it looked more professional. However, none of the results generated a statistically significant difference between means.

Overall, it appears that there is little impact of conducting the interviews conversationally as opposed to using a task-oriented manner, except that conversational interviews may yield more usability errors. Users were equally likely to express their reactions to the system in either type of interview, and were likely to give the system the same type of rating in the posttest interviews. Furthermore, conversational interviews, contrary to what one might expect, did not take longer to complete.

Table 5. Posttest Survey Outcomes by Type of Interview

Variable Description	Type of Interview	
	Conversational Mean Score (n)	**Task-oriented Mean Score (n)**
How comfortable or uncomfortable did you feel performing the tasks in the test? 5: Very comfortable ~ 1: Very uncomfortable	3.4 (18)	3.2 (17)
How certain or uncertain are you that you completed the tasks successfully? 5: Very certain ~ 1: Very uncertain	3.2 (18)	3.5 (17)
Compared to other similar software you have used, how would you rate this performance management system in terms of ease of use? 5: Much less complicated ~ 1: Much more complicated	3.2 (18)	3.1 (17)
Overall, how easy or difficult was the system to use? 5: Very easy ~ 1: Very difficult	3.5 (18)	3.2 (17)
Overall, how easy or difficult was the system to understand? 5: Very easy ~ 1: Very difficult	3.4 (18)	3.5 (17)
Overall, how professional or unprofessional did the system appear? 5: Very professional ~ 1: Very unprofessional	3.8 (18)	4.5 (17)
Overall, how efficient or inefficient was the system? 5: Very efficient ~ 1: Very inefficient	3.1 (18)	3.3 (17)
Overall, as you worked through the tasks, did the product become… 5: Much easier to use ~ 1: Much harder to use	3.9 (18)	4.0 (17)
Overall, how effective or ineffective do you think the performance management system will be as a career development and career planning tool? 5: Very effective ~ 1: Very ineffective	3.5 (18)	3.5 (17)
I have a clear understanding of the performance management system. 5: Strongly agree ~ 1: Strongly disagree	3.1 (18)	3.1 (17)
The performance management system seems fair/accurate. 5: Strongly agree ~ 1: Strongly disagree	3.8 (18)	3.6 (17)
The performance management system allows performance reviews to be conducted in a timely manner. 5: Strongly agree ~ 1: Strongly disagree	3.4 (18)	3.9 (17)
I am satisfied with the test version of the performance management system. 5: Strongly agree ~ 1: Strongly disagree	3.3 (18)	3.4 (17)

Discussion

Since usability tests were not conducted as a probability-based sample of any population and methodological variations were not implemented as an experimental design, it is not appropriate to draw broad conclusions about the types of usability testing approaches which should be used in the future. However, results presented in this paper provide those conducting usability tests with several key insights. First, analyzing task durations does not provide the detailed results which may lead to a greater understanding of test system problems. Specifically, analysis of usability errors identified the kinds of problems users encounter with a system, uncovered functional problems such as slow servers, and demonstrated the need for navigational guides like the QRG. Although performing such error analysis may have perceptibly negative cost and schedule implications, the amount of detail this method uncovers far surpasses simple timing data and should be considered a worthwhile investment of resources. Second, subjective measures greatly enhance usability test results by providing an awareness of the specific user context or culture and by producing recommendations for improvement. Understanding of the greater context and culture into which a new system is to be implemented is of the utmost importance while planning for the full-scale implementation of a system. Such awareness can help inform a multitude of decisions such as the process by which the system will be implemented, system training content and education, and gaining cooperation and support from the users. Although objective measures of usability testing may be able to create a functionally perfect system the system may still fail upon implementation if it is not integrated into the current culture in a conscientious manner. It is difficult to build an optimal system without considering the greater context to which it belongs. Thus, the subjective measures elevate the inquiry from the micro level of system planning to the macro level of system acceptance and implementation. This component is especially important if one of the goals of usability testing is to create a system that facilitates organizational change. For example, one of the goals of this study is to create a system that increases job satisfaction and perception of fairness amongst personnel. It would be impossible to accomplish this goal without considering the perceptions and opinions of actual personnel. Third, conversational interviewing does not appear to have an effect on the results of posttest surveys or on the durations of the tests, but it may yield more errors. Since the goal of a usability test is often to identify as many errors as possible, a conversational approach may be the most effective. Although a desire for objectivity may lead us to believe that the more naturalistic approach of task-oriented testing, in which the usability analyst is more of an unobtrusive observer, is more valid, the conversational approach's benefit of yielding more useful data may outweigh its perceived limitations. To increase the effectiveness of user testing, we recommend more rigorous analyses to evaluate these usability test methods.

Referentes

Bevan, N., & Macleod, M. (1994). Usability measurement in context. *Behavior and Information Technology, 13*, 132-145.

Dean, E., Schwerin, M.J., Robbins, K.M., & Bourne, M.J. (2004). *Usability testing of the U.S. Navy's performance management system:* Technical Report #2. Research Triangle Park, NC: RTI.

Department of Defense Instruction (DODINST) 5000.2 (2003). *Operation of the Defense Acquisition System*. Washington, DC: Department of Defense (USD AT&L).

Dolan, W.R., & Dumas, J.S. (1999). A flexible approach to third-party usability. *Communications of the ACM, 42*(5), 83-85.

Dumas, J.S., & Redish, J.C. (1993). *A practical guide to usability testing.* Norwood, NJ: Ablex Publishing Corporation.

Edmunds, H. (1999). *The focus group research handbook*. Chicago, IL: NTC Business Books.

Krueger R. (2003). *Focus groups: A practical guide for applied research* (3rd ed.). Thousand Oaks, California: Sage Publications.

Morgan D. (1997*). Focus groups as qualitative research* (2nd ed.). Thousand Oaks, California: Sage Publications.

Nielsen, J. (1993). *Usability engineering*. Boston: AP Professional.

Norman, K.L, Friedman, Z., Norman, K., and Stevenson, R. (2000). Navigational Issues in the Design of On-Line Self-Administered Questionnaires. Human-Computer Interaction Laboratory, University of Maryland. *Unpublished Manuscript.* Accessed from ftp://ftp.cs.umd.edu/pub/hcil/Reports-Abstracts-Bibliography/2000-22html/2000-22.pdf.

Schwerin, M.J., Dean, E., Robbins, K. M., Bourne, M.J., and Reed, L. (2005, In Press). "Subjective and Objective Results of Usability Testing for the U.S. Navy's Performance Management System." *Military Psychology.*

Westerman, S.J. (1997). Individual Differences in the Use of Command Line and Menu Computer Interfaces. *International Journal of Human-Computer Interaction*, 9(2), 183-198.

Wiedenbeck, S. (1999). The use of icons and labels in an end user application program: an empirical study of learning and retention. *Behavior & Information Technology*, 18(2), 68-82.