Naval Submarine Medical Research Laboratory

NSMRL Report 1172 23 JULY 1991

A Comparison of the Usability of Three Versions of a Computerized Medical Diagnostic Assistance Program for Abdominal Pain

by

Elaine F. Chouinard
University of Hartford
Bernard L. Ryack, and Douglas M. Stetson
Naval Submarine Medical Research Laboratory

Released by:
R. G. Walter, CAPT, DC, USN
Commanding Officer
Naval Submarine Medical Research Laboratory

A Comparison of the Usability of Three Versions of a Computerized Medical Diagnostic Assistance Program for Abdominal Pain

by

Elaine F. Chouinard, University of Hartford,
Bernard L. Ryack, and Douglas M. Stetson,
Naval Submarine Medical Research Laboratory

Naval Submarine Medical Research Laboratory
Report Number 1172

Naval Medical Research and Development Command
Work Unit 63706N-M0095.005-5010

Approved and released by

R. G. WALTER, CAPT, DC, USN
Commanding Officer
NavSubMedRschLab

SUMMARY PAGE

THE PROBLEM

To compare the usability of three versions of the Abdominal Pain module of the MEDIC program. MEDIC is a computer based medical decision support program designed to be used by corpsmen onboard the submarine.

FINDINGS

A higher user satisfaction rating was associated with visual grouping of related items, ordering of items to coincide with the usual medical examination, the use of color to highlight information and direct the user, and minimal and consistent steps for data entry. Preference for graphic and list formats for the presentation of the diagnostic summary information was nearly equally divided. Confidence in the program-generated diagnosis was correlated with its perceived usability as measured by a user satisfaction questionnaire.

APPLICATIONS

These findings identify usability recommendations to guide the development of the final version of the MEDIC module for abdominal pain.

ADMINISTRATION INFORMATION

This investigation was conducted under the Naval Medical Research and Development Command, U.S. Department of the Navy, Research Work Unit 63706N-M0095.005-5010. The views expressed in this report are those of the authors and do not reflect the official policy or position of the Department of the Navy, Department of Defense, or the U. S. Government. It was submitted for review on 1 February 1991, approved for publication on 23 July 1991, and has been designated NSMRL Report #1172.

Abstract

Three versions of a computerized medical diagnostic assistance program for abdominal pain were tested for ease of use, ease of learning, user satisfaction, and time to complete the "Pain Site" screen. A higher satisfaction rating was associated with visual grouping of related items, ordering of items to coincide with the usual medical examination, the use of color to highlight information and direct the user, and minimal and consistent steps for data entry. Preference for graphic and list formats for the presentation of the diagnostic summary information was nearly equally divided. Longer learning time was associated with inconsistent rules for the handling of completed screens. Longer time to complete a screen was associated with a lack of grouping of related items, multiple steps for data entry, a lack of instructions identifying required and optional data entry items, and exclusive use of upper case text. Confidence in the program-generated diagnoses was found to increase with user satisfaction.

A Comparison of the Usability of Three Versions of a Computerized Medical Diagnostic Assistance Program for Abdominal Pain

A computerized diagnostic assistance program (MEDIC) to aid corpsmen in the diagnosis and treatment of illness and injury has been developed by the U.S. Navy and has been in limited use for approximately ten years. The individual modules have been repeatedly revised in response to advances in technology (Ryack, 1987). The availability of three versions of the abdominal pain module, differing only in presentation, not in content, provided an opportunity to determine the effect of presentation format on the usability of the program. A usability study was conducted to compare the three user interfaces against each other.

Background

The field of human-computer interaction is extensive. It covers such diverse topics as user behavior, systems evaluation, usability, and readability (Carroll, 1988); and it broaches such specific and general problems as fill- vs. right-justification of text, text vs. graphic presentation of materials, use of upper case only vs. upper and lower case letters, concrete vs. abstract instructional models, organization of menus, and comparisons of training methods, to name a few.

"A picture is worth a thousand words"

As diverse as the research areas have been, in reviewing the results, a common theme emerges. Within the medium afforded by the computer (the screen), an opportunity exists to present information is presented as both the actual information contained in the display, and the organization of that information, or "picture" that is seen when viewing the screen as a whole. That is, independent of the nature of the specific information displayed, the overall appearance of the screen communicates information. When the appearance of the screen is improved, it can lead to increased efficency. On the other hand, when the organization of the information is not usefully managed, the information it communicates can be confusing and distracting, leading to decreased efficiency.

Conventional Wisdom Prevails

When reading printed material, readers extract information from the appearance of the sentences and the page, in addition to the actual meaning of the words themselves. Specifically, one full space between letters indicates separate words and double spacing after a period indicates separate sentences. Trollip and Sales (1986) found support for the carry-over of this conventional usage of text appearance into computer generated printed text.

Subjects reading fill-justified text (text that is aligned such that both margins are straight by inserting unequal spaces between letters and words) took considerably longer to complete a passage than those reading the same text in the conventional left-margin justified form (text aligned to a straight left margin, leaving the right margin jagged) (Trollip & Sales, 1986). They concluded that the absence of the information usually provided by conventional spacing led to decreased efficiency in the reading of the fill-justified material.

Evidence also suggests that adherence to conventional page formats when presenting

narrative text on the screen increases efficiency, perhaps by its consistency with reader expectations. Specifically, screens displaying several lines at a time, across the full width of the screen, and in an eighty character per line format, were read faster (Duchnicky & Kolers, 1983), perhaps due to the screen's resemblance to the conventional printed page.

The use of upper and lower case letters also provides information to the reader that is independent of the actual meaning of the words printed (that a word denotes a proper name, for example). Upper and lower case text was found to be read 13% faster than all upper case text (Tullis, 1983). Presumably, as in the Trollip and Sales study, the absence of the information usually provided by the use of upper and lower case letters led to decreased efficiency when reading material presented in all upper case letters.

Payne, Sime, and Greene (1984) suggested building on reader expectations that a letter in upper case signals "something different" when designing computer applications. In a text editor program, in which letters could be used for both command statements and literals, a simple change was made such that placing a letter in upper case indicated that it was being used for a command statement. A considerable reduction in errors occurred with the version in which the use of upper case provided this additional information to the user (Payne et al., 1984). The authors reflected, "...a small syntactic change can produce a very large increase in usability, just by revealing the structure of the command more clearly" (Payne et al., 1984, p. 28).

"Let me illustrate..."

Many studies suggest that the more the appearance of a given screen draws a picture of its function, the greater will be the resulting increases in efficiency. For example, subjects using information displayed in a flow chart format solved problems faster and more accurately than those presented with the same information in simple prose (Wright & Reid, 1973). Speed and accuracy increased as the presentation of information became more graphic. The performance of subjects using a short sentence or table format was superior to those using simple prose, falling somewhere in between the prose and flow chart groups (Wright & Reid, 1973). Tullis (1980) also found performance with information displayed in graphics or structured text (grouping and highlighting related terms) to be superior to narrative text.

One of the best examples of a screen drawing a picture of its function can be found in the study by Hanson, Payne, Shively and Kantowitz (1981). Subjects were asked to monitor level indicators. The readings were presented as either the actual numerical values displayed in windows or as labeled horizontal bar graphs that moved right or left as the levels changed. As the number of indicators to be monitored increased, the reaction time of the subjects using the numerical display decreased as compared with those using the bar graph illustration, pointing to the superiority of the analog format for quick extraction of information (Hanson et al., 1981).

The display capabilities of computer screens provide opportunities to use screen appearance to communicate information in ways not possible on a printed page. For example, the display screen can be divided to show many different types of information at once. Logic suggests that using the same spaces to display the same type of information from screen to screen will soon lead the user to know the type of information being displayed,

merely by its placement on the screen. Indeed, this logic was supported by empirical evidence in a study for the NCR Corporation by Keister and Gallaway (1983). In the Keister and Gallaway study, several revisions were made to a data entry program. These revisions included: assigning specific screen areas for status messages, aligning dental entry fields, and a 2-field system for correcting errors such that both the error and the new input appeared on the screen simultaneously during error correction (Keister and Gallaway, 1983). Transaction times improved considerably and errors decreased significantly with the revised version of the program (Keister and Gallaway, 1983). Logic and empirical evidence also dictate that efficiency is highest when the overall picture is not cluttered with supplementary or embellishing information (Tullis, 1983). Thus, while designers may find the creative display of elaborative information aesthetically pleasing, providing information beyond that which is needed for the immediate task at hand is, in fact, distracting.

"But, do you see it my way?"

Although established conventions and empirical studies suggest general guidelines for screen appearance, specific designs are confounded by individual differences. For example, menus have been found to be most effective when based on the users' cognitive network (Roske-Hofstrand & Paap, 1986) and on users' goals (Mehlenbacher, Duffy & Palmer, 1989). Both Sein and Bostrom (1989) and Lamberti and Newsome (1989) found increases in usability when a program represented a task in a way more closely approximating the users' internal representation of the program. In other words, the more the screen's picture of the task resembled a user's internal picture of the task, the easier the program was to use. Appreciation of the implications of these findings has led to the most recent trend in human-computer interaction research: Ideas for development and design are now coming from the users themselves and then sent to the designers to be refined rather than the other way around (Carroll, 1988).

User-driven Research

If designers want their programs to be used, they must design usability into them (Schott & Olson, 1988). Carroll (1988), in his review of the history of the human-computer interaction field (HCI), describes how designers reached this conclusion.

Initial HCI research involved contrasting performance on different programs or program versions. This resulted in listings of studies organized on "...the basis of superficial features (e.g., as pertaining to variable names or menu systems)" (Carroll, 1988, p. 4). In addition, these contrasting studies often used the designers themselves as subjects and were performed in artificial laboratory conditions. This leads to a questioning of the ability to generalize the results to actual users with very different educational backgrounds and to user environments with unique real world needs. The second phase of HCI development was the trend towards linking research questions to existing models of human information processing. While this recognized the need to take the cognitive model of the user into consideration during the design phase, research results produced broad generalizations that were often impossible to translate into practical computer applications. The current developmental phase of HCI generates its hypotheses and performs its research with the actual users in the users' environments. The resulting "user-innervated invention" has produced innovations designed

for and within the context and situations of their intended use (summarized from Carroll, 1988).

Helen G. Bradley, of the Sabre Travel Information Network, is involved in ongoing testing of the Easy Sabre product (an airline ticketing service available to individual travelers using home computers). She states, "... observation of people actually using your service will provide you with invaluable information" that will assist in "design and refinement according to the mentality of the users rather than merely the logical flow of screen set-ups. The idea is to observe people using the product and then ask them why, in a given scenario, they did what they did" (H. G. Bradley, personal communication, January 1990). Fred Schott of the People/Technology Services at Aetna Life and Casualty in Hartford, Connecticut states, "The most instructive sessions occurred when the designers simply observed users making trial runs with the prototype" (Schott & Olson, 1988).

Such observation sessions generally take place in a usability lab. A typical usability lab contains a PC and three cameras in a room with a one-way glass in one wall. One camera is focused on the person and station, another on the screen, and a third on the keyboard. They record the entire session. Observers watch the user interacting with the program from behind the one-way glass. From these observations, the observers formulate the questions they will be asking the user immediately following the interactive session (Schott & Olson, 1988).

Method

Given the Navy's long range goal of increasing usage of the MEDIC programs, a usability study provided the best format for identifying usability problems and suggesting areas for revision. The three versions of the abdominal pain module illustrated three very distinctly different presentations of the same material. We hypothesized that user reactions to, and our measured usability of each version would also be distinctly different. Usability objectives can be as divverse as the software applications themselves, and have included "easy to use, fast to use, fun to use, easy to learn" (Potosnak, 1988, p. 89), user comments, user recall of good and bad features, user preference, user frustration, and user satisfaction (Whiteside, Bennett, and Holtzblatt, 1987). For the purposes of this study, usability was defined as ease of use, ease of learning, and user satisfaction.

Materials

The Versions

Three versions of the abdominal pain module of the MEDIC system were used. All the information in the three versions was identical. They differed only in their methods of presentation and elicition of information from the user. All versions were run on IBM or IBM compatible computer systems.

A detailed description of the three versions is given in Appendix A.

The Workstations

Three workstations were set up, each in a pale-colored cubical. Glare from the windows was blocked, and uniform lighting was provided. Each cubical contained one video camcorder aimed at the screen and keyboard.

Test Materials

Three sample cases of abdominal pain were randomly chosen from a library of cases used for training users with the abdominal pain module. Users were presented with a

printed narrative of history and physical examination for each case.

Users filled out a questionnaire to measure user satisfaction. It was composed of eleven items appropriate to this study taken from a questionnaire developed by Pearson (Bailey & Pearson, 1983). Pearson reported the instrument to have a predictive validity of .79 and a reliability of .93 (Bailey & Pearson, 1983). Baroudi and Orlikowski (1988) show that the use of a shortened and adapted version of the Pearson questionnaire only minimally compromises the validity and reliability reported by Pearson. Each questionnaire item was rated on scales composed of adjective pairs that represent opposite evaluations, such as "good" and "bad". Several of the adjective pairs in the form adapted for the current study were reverse scored to discourage users from marking straight down one side of the column. The score for each item was calculated by taking the average of its adjective pairs. The overall usability score is a sum of the item scores. An additional item of interest to the program developers was added, which was not used in the calculation of the overall usability scores, but was tallied only for the purpose of gathering information. This item addressed data entry and asked users to rate the method for entering their responses on speed, comprehension, ease of use, and efficiency.

Users

Ten male and one female active duty Independent Duty Corpsmen and one male U.S. Navy medical student participated as users of the program. Three had used one of the versions of the abdominal pain module of the MEDIC program previously, one had used a MEDIC module for another diagnostic category, and eight had never used any portion of the MEDIC program. Two reported that word processing was their only computer experience, while all others reported some experience beyond word processing. The additional experience was with data base and spreadsheet programs, computer games, and "other". Four users reported experience with three or more different computer applications.

Procedure

Three users participated in each of four sessions. Each user was instructed to complete the three test cases using all three versions of the diagnostic program. The order of presentation of the versions, as well as the order of presentation of the test cases within each version, was randomized. Users completed one copy of the questionnaire for each of the three versions, completing them immediately after using each version. At the end of the session, users were individually interviewed. The interview consisted both of questions asked of each user as well as specific questions asked of particular users after observing their interactions with the versions during the test sessions.

Results

The results of Pearson Product Moment correlations showed no effects for computer experience, experience with the MEDIC program, or usability session on any of the dependent variables. Usability was defined in terms of ease of use, ease of learning, and user satisfaction. In addition, participants were timed on the particular screen which asked for "Site of Pain at Onset" for all cases, in all versions. The means of these times are compared among versions as a method of comparing the efficiency of the different screen layouts for this question. All analyses reported below were performed using SPSS-X

version 4.0, using a p-value of .05 as the significance level for the statistical tests.

Ease of Use

Ease of use was defined as the mean time taken to complete the three cases. These data are incomplete for several reasons. The video of user eleven with version 1 was not available due to a camera malfunction, and user eight was interrupted while using version 2 due to a computer hardware failure. These sessions were excluded from the ease of use measure. Moreover, the usability sessions were limited to three hours to permit the users to return to their duties. If a user had not completed all three cases within a version after forty-five minutes, he/she was instructed to stop and begin filling out the questionnaire. For these reasons, conclusions from the response times are tentative. As shown in Table 1, the mean time to complete all three test cases was: 29 minutes (N=9) for version 1, 35 minutes (N=10) for version 2, and 36 minutes (N=9) for version 3. These differences were not significant according to a one-way analysis of variance, $F(2,25) = 1.32$, $p = .28$.

Forty-five minutes was not enough time for the completion of three cases for 17% of the users (two out of twelve) with version 1, for 8% of the users (one out of twelve) with version 2, and for 25% of the users (three out of twelve) with version 3. For those users not completing all three cases with version 1, one user completed two cases in forty-five minutes while one completed just the first case. With version 2, the user whose session was cut short completed one case in the forty-five minutes allotted. Of the three users who ran out of time with version 3, two completed only one case and one user completed two cases.

Ease of Learning

Ease of learning was defined as the difference in time taken to complete the first vs. the third cases within each version. The mean differences between the first and third cases were four minutes for version 1, seven minutes for version 2 and nine minutes for version 3. As shown in Table 2, the F-ratio of 3.06 obtained for the analysis of variance performed on version by the differences in times between the first and third cases has a

Table 1. Ease of Use.

Mean Time in Minutes to Complete Each Version		
Version 1	Version 2	Version 3
29 (n=9)	35 (n=10)	36 (n=9)

ANOVA Total Time* by Version

Source of Variance	Degree of Freedom	Sum of Squares	Mean Squared	F Ratio	F Probability
Between Groups	2	255.54	127.77	1.32	.28
Within Groups	25	2412.89	96.52		
Total	27	2668.43			

*Incomplete data, see text for explanation.

Table 2. Ease of Learning.

Mean Differences in Times to Complete the First and Third Cases					
Version 1		Version 2		Version 3	
4 Minutes (n=9)		7 Minutes (n=10)		9 Minutes (n=9)	
ANOVA Differences by Version					
Source of Variance	Degree of Freedom	Sum of Squares	Mean Squared	F Ratio	F Probability
Between Groups	2	151.49	75.75	3.06	.06
Within Groups	25	618.22	24.73		
Total	27	769.71			

probability of .06. This probability suggested that T-tests contrasting each version with the other two may reveal differences among the versions. The contrasts showed that the difference between version 1 and version 3 was significant (T = 3.4, p = .004).

The average time to complete the third case in both versions 1 and 3 is eight minutes. However, the average time to complete the first case with version 1 is twelve minutes, as compared to twenty minutes with version 3 (see Figure 1). Significantly more learning was required with version 3 than with version 1 to reach the same level of proficiency. A Pearson Product Moment Correlation correlating these time differences with the usability questionnaire scores was -.48 (p=.004), a

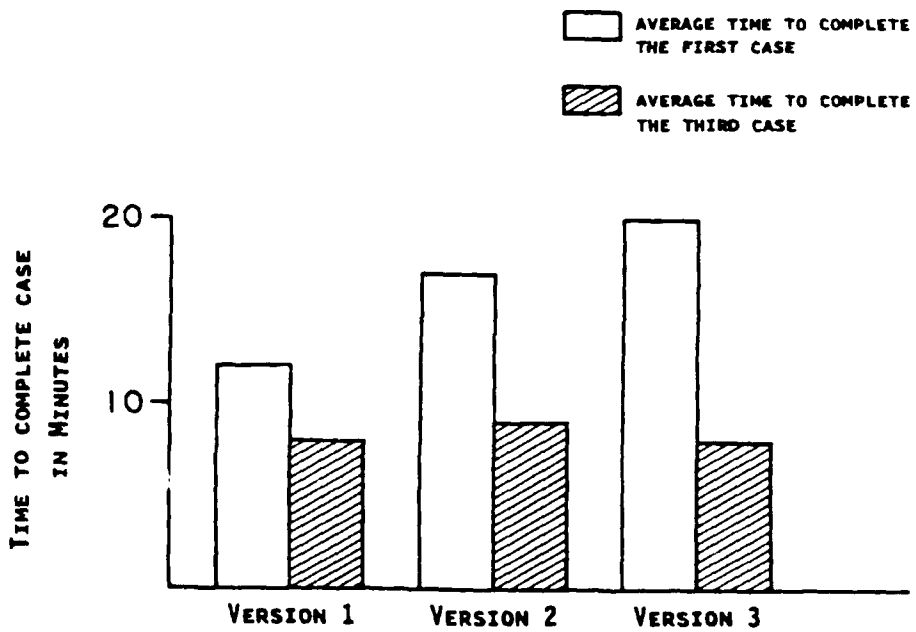


Figure 1. Version 3 required significantly more learning time than Version 1 to reach the same level of proficiency in data entry.

moderately high negative relationship. This shows that greater learning requirements are associated with a lower usability rating.

User Satisfaction

User satisfaction was defined by the scores on the modified Pearson usability questionnaire. The range of scores possible is -33 to 33. The mean score was 21.17 for version 1 (range = 14 to 31.5), 9.54 for version 2 (range = -4 to 28.25), and -2.58 for version 3 (range = -25.25 to 11.25). Thus, version 1 was the most preferred and version 3 the least. Nine users preferred version 1, three users preferred version 2 and none preferred version 3. These differences are highly significant, $F(2,33) = 18.94$, $p < .01$. A contrast analysis (Table 3) revealed significant differences within all possible pairings of the three versions (all T-values were significant at or below the .01 level). These results indicate the users felt uniquely different about each of the three versions. Appendix B gives the breakdown of the mean scores on each of the individual items.

The overall score on the questionnaire can be seen as an answer to the question, "Did you like this version?", and a review of the answers to the individual items indicate why the user liked or disliked it. The tallies of the users' responses to individual items on each of the three versions are given in Appendices C through E. The responses for version 1 are skewed in the positive direction. In particular, users indicated they found this version to be readable, organized, easy to understand, and easy to use. Responses to version 2 are widely scattered, with a slightly positive trend, indicating user opinion on version 2 was quite varied. Users liked the error recovery method in version 2, indicating they found it to be simple and fast. Responses for version 3 are skewed slightly in the negative direction overall. Users indicated they found the data entry method tedious and felt their understanding of the system to be low. The mean score for every item except one, "Language," was highest for version 1 and lowest for version 3. The order for "Language" is version 1, followed by version 3, with version 2 last.

Perhaps the most surprising result, and the one most remarkable in its implications, is the result obtained for the item "Confidence in the System." Recall that all three versions requested the same information from the users,

Table 3. User Satisfaction.

Mean Scores on the Usability Questionnaire					
Version 1		Version 2		Version 3	
21.17 (n=12)		9.54 (n=12)		2.58 (n=12)	
ANOVA Usability Score by Version					
Source of Variance	Degree of Freedom	Sum of Squares	Mean Squared	F Ratio	F Probability
Between Groups	2	3384.88	1692.44	18.94	< .01
Within Groups	33	2948.94	89.36		
Total	35	6333.81			

calculated the diagnosis in the same manner, and offered the same diagnosis in response to the same input. Yet, the users' confidence in that diagnosis differed greatly from version to version. The average score was 1.85 for version 1, 1.17 for version 2, and -0.78 for version 3 (-3 denotes low confidence and 3 denotes high confidence). When asked to explain the rating they assigned for "Confidence in the System," users associated their low confidence ratings for version 3 with a lack of guidance offered by the menus, inconsistent data entry rules, the fact that the values for the data entry response codes changed, the difficulty of changing a response, the fact that response choices are not easy to select, the lack of similarity between the commands for this version and other popular software, and the fact that "Press any key" was not always true. Two users summed their opinion by saying that, due to these difficulties, they were not sure that the program had "gotten" the information they were trying to give it. Therefore, since the program may have been computing its diagnosis with incomplete information, they could not place much confidence in that diagnosis. In explaining the high confidence assigned to version 1, users cited the fact that the method of presentation of the items highlighted for them the important points that lead to a diagnosis, reminded the corpsman to ask questions that might have been forgotten, and served as an information gathering aid. Also cited were the facts that the graphic display of probable diagnoses made it easy to see the most and least likely diagnoses, and that version 1 was "the easiest to use." One user, who assigned version 1 a confidence of 3, explained his rating this way, "[It] seems to work well!"

Two conclusions can be drawn from these results. First, the procedure for interacting with the system will influence users' con-

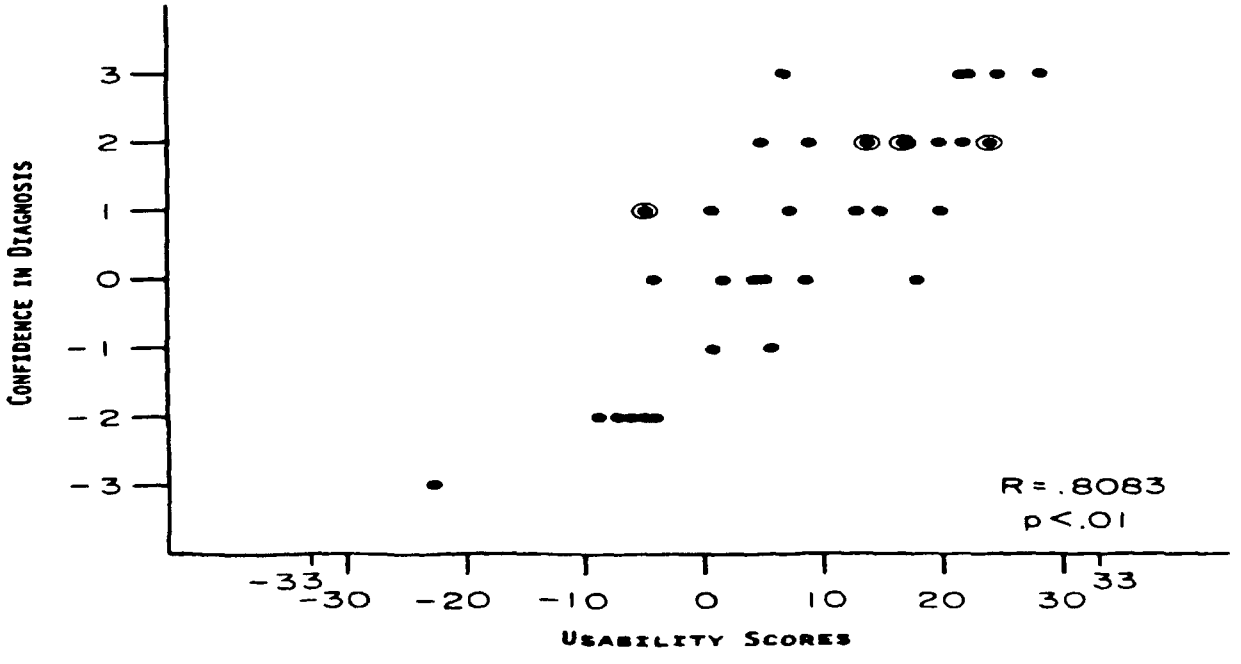


Figure 2. Correlation of confidence scores with overall usability score.

fidence in the system. Second, given the same output information by a computer system, the users' confidence in the information will be influenced by the way in which the information is presented. A correlation of the confidence ratings with the overall satisfaction ratings shows that confidence in the system increases as satisfaction increases ($r = .8083$, p .01) (see Figure 2). This suggests a third conclusion, that users' judgments of the functionality of a system may be influenced by their impressions of how much they "like" the system.

Pain Site Screen Comparison

In each of the three versions, the section covering "Site of Pain at Onset" was contained on one screen (although each version presented the questions in a uniquely different fashion) making this screen particularly suitable for comparing the characteristic style of presentation used by each version. There were significant differences in the mean time users spent responding to "Site of Pain at Onset" between version 3 (50 seconds) and both versions 1 (21 seconds) and 2 (28 seconds) according to T-tests ($T1\text{-}3 = -3.17$, $p < .01$; $T2\text{-}3 = -2.20$, $p < .05$). As shown in Figure 3, version 3 took the users significantly longer to complete.

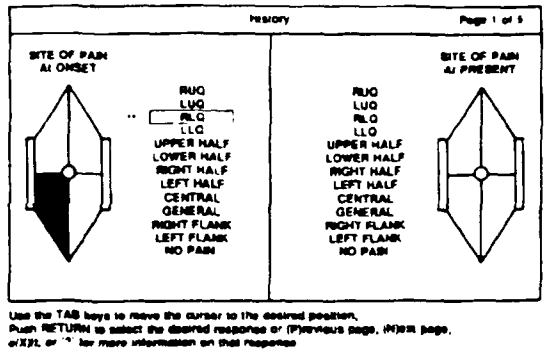
Usability Session Observations

Anecdotal reports of users' comments and behaviors, observed as they interacted with the each of the versions, appear in Appendix F.

Discussion

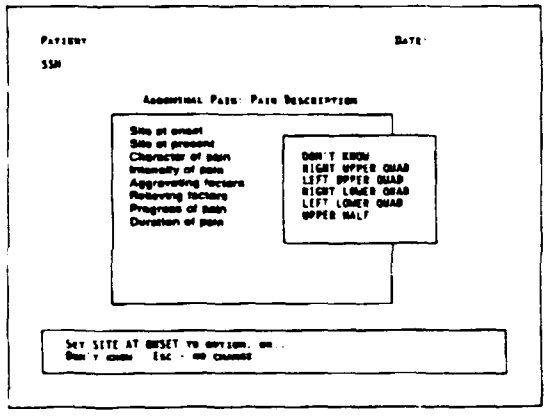
The results show that the differences in presentation and method of data entry among the three versions produced differences in

Version 1



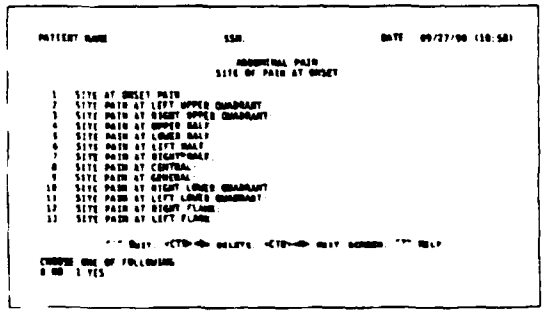
$\bar{X}$ = 21 seconds

Version 2



$\bar{X}$ = 28 seconds

Version 3



$\bar{X}$ = 50 seconds

Figure 3. Version 3's layout for pain site identification which took significantly longer to complete than both Versions 1 and 2.

their usability. Ease of use (average time to enter a case) results were inconclusive due to missing data. However, users took significantly longer to complete the Site of Pain at Onset screen in version 3 than in version 1. A significant difference for ease of learning (time difference between case 1 and case 3) was observed between versions 1 and 3. The most significant findings were in the area of user satisfaction (score on satisfaction questionnaire); specifically, the finding that users' confidence in the system correlated with their satisfaction with the system. Many of the user comments, made during the post-session interviews, suggest reasons for the obtained results.

Version 1 earned the highest satisfaction rating. Users commented that they prefered its separation and ordering of the history and exam sections, because this follows the order in which the same items are completed during the typical examination of real patients. One user explained that the questions in version 1 follow the S.O.A.P. Note (Subjective, Objective, Assessment, Plan) format typically used by corpsmen in completing their examinations. Users said that the bouncing back and forth between history and exam questions found in versions 2 and 3 was distracting. These reactions were consistent with findings by Wright and Reid (1973) and Hanson et al. (1981) who found that display formats which mimicked their function increased efficiency, and by Sein and Bostrom (1989) and Lamberti and Newsome (1989) who noted that usability increased as the program's match to the users' internal representation increased.

Users also felt that version 2 required too many steps for data entry, which may help to explain this version's lower satisfaction rating. Users were particularly frustrated by the dialogue boxes used in this version. Their reactions appear in Appendix F, Usability Session Observations. A puzzling result from the satisfaction questionnaire is the ranking of version 2 last in "Language". None of the users commented on the language used in this version, and none of the users exhibited difficulties during the usability test sessions that appeared to be language related. One user commented that he was confused by version 2's usage of "distension" and "swelling" and suggested the version include definitions distinguishing the two terms, but this alone cannot account for the low rating given version 2 in the area of language.

Users assigned version 3 a negative satisfaction rating. The majority of complaints with version 3 centered around the methods of navigating through the screens. One of the chief complaints was that this version did not allow a return to a previous "page". In addition, every user referred to the difficulties involved in having different rules for different screens. Users also commented on the procedure for moving to the next screen, saying it was too complicated and required too many steps. One user felt that the key commands differed too much from those used in other commonly used programs. Users also stated that the use of a code for data entry was too time consuming. These findings support Smith and Mosier's (1986) suggestions that such things as allowing users to change entries and staying with one method of data entry will increase usability.

User's confidence in the program generated diagnoses increased as the satisfaction ratings they assigned to the versions increased. Users commented that when entering data into the version(s) they judged as more usable, they felt more confident that the information they had entered was what had been actually "accepted" by the program.

Version 3 was found to require significantly more learning time than version 1. As all the users pointed out, version 3 had two sets of rules to learn for handling the screens. A sometimes amusing account of users' difficulties with this appears in Appendix F, Usability Sessions Observations. It makes intuitive sense that this characteristic of version 3 would lead to the observed increased learning time.

The "Site of Pain at Onset" screen in version 3 took significantly longer to complete than the corresponding screens in versions 1 and 2. Users asked why the first item in version 3, requesting if pain is present, is needed, given that the subsequent items ask the user to answer regarding the presence of pain in each individual site. Also, on-screen instructions stating that only the "yes" answers are needed would indeed have saved time in completing this screen. This is also consistent with Smith and Mosier's (1986) suggestion that a screen should distinguish between required and optional data entry.

Tullis (1980) found that grouping related items together to produce visual and conceptual "chunks" enhances performance. Related items are grouped in versions 1 and 2, but not in version 3. In addition, version 3 alone uses all upper case letters for items requiring user input. Tullis (1983), in a review of the literature, found that upper case text is read 13% slower than upper and lower case combined. Both of these factors could have contributed to the slower completion time for the pain site screen in version 3.

Lastly, some users commented that the use of the graphic representation of the abdomen in version 1's screen for identifying pain sites eliminated confusion between right and left, provided feedback, and summarized information quickly. Wright and Reid (1973) showed that similarity of a screen's layout to its intended function increases efficiency. This effect could account for the quicker completion time in version 1.

The lack of significant difference between the "Site of Pain at Onset" screen times for versions 1 and 2 can perhaps best be explained by the users' divided preferences concerning graphics. Version 1 uses a graphic format and version 2 uses structured text, both of which were shown to improve usability by Tullis (1980).

Version 1 uses graphics for the pain site and diagnostic summary screen, as well. Of the users who commented on this, six stated they preferred the graphic representations and four preferred to see the same information presented in list format, as it is in version 2. With specific reference to the diagnostic summary, those preferring a list suggested an ordering of the probable diagnoses in descending order of probability within the list as the most useful presentation of the information. They stated that the graphic representation was distracting. Those preferring the graph stated that it enabled them to extract a great deal of information in a short period of time. A more detailed account of users' reactions to these two presentation formats appears in Appendix F, Usability Sessions Observations.

Many users commented favorably on the use of colors. They stated that the colors aided in bringing information out quickly. Color also helped lead the user through the version and highlighted what should be attended to. In addition, color was stated to be more entertaining, less drab, and less likely to lead to boredom (and its accompanying fatigue). One user reported a sensation of eye

strain after using version 3, which appears in black and white. (A tally of frequently made comments appears in Appendix G).

Conclusions

A higher satisfaction rating was associated with visual grouping of related items, ordering of items to coincide with the actual physical exam, the use of color to highlight information and direct the user, and minimal and consistent steps for data entry. Both graphic and list formats for the presentation of information were also associated with higher satisfaction ratings. A negative satisfaction rating was associated with the inability to correct a previous entry, multiple steps for data entry and for proceeding to the next screen, and inconsistent rules for data entry and the handling of completed screens.

Longer times to complete a screen occured where there was a lack of grouping of related items, multiple steps for data entry, a lack of instructions identifying required and optional data entry items, and exclusive use of upper case text.

Longer learning times were associated with inconsistent rules for the handling of completed screens.

Confidence in a particular version was correlated with its assigned satisfaction rating. This supports current thinking that designing for user satisfaction is not just "icing on the cake," but, in fact, is a critical consideration in the software design process (Schott & Olson, 1988). User satisfaction can be a determining factor in whether a program is used for its intended purpose or not. The three versions of the current diagnostic program for abdominal pain are intended to be used with confidence by a corpsman performing his duties in the cramped environment of a submarine located a great distance from traditional medical support facilities. The information asked for and presented by the program needs to be understood quickly, accurately, and conveniently, and the user must have confidence in the results. Aspects of the interface that address data entry methods and presentation format have been shown to influence users' satisfaction with the program and confidence in its output. Since the corpsman's decision to consult the program will be influenced by the amount of confidence placed in its output, the user interface should be designed so as to increase this confidence. The following recommendations, based on the present findings, address these specific needs.

1. Reduce the steps required for data entry to the minimum number possible and keep them consistent throughout the program.

2. Organize the presentation of material so as to visually group related items.

3. Order items to follow the order in which they are addressed during the actual examination of the patient.

4. Allow the user to return to previous items to change responses at any point in the program.

5. Give the user the choice to view the diagnostic summary in list or graph form.

6. Use upper and lower case letters.

7. Use color to highlight information and to direct attention.

In addition, Appendix F, Usability Session Observations, describes issues related to specific items.

References

Bailey, J. E., & Pearson, S. W. (1983). Development of a Tool for Measuring and Analyzing Computer User Satisfaction. Management Science, 5(29), 530-545.

Baroudi, J. J., & Orlikowski, W. J. (1988). A Short-form Measure of User Information Satisfaction: A Psychometric Evaluation and Notes on Use. Journal of Management Information Systems, 4(4), 44-59.

Bradley, H. G. Personal communication. January 1990.

Carroll, J. M. (1988). Evaluation, Description and Intervention: Paradigms for Human-Computer Interaction. IBM T.J. Watson Research Center, Yorktown Heights, NY.

Duchnicky, R.L. & Kolers, P.A. (1983). Readability of Text Scrolled on Visual Display Terminals as a Function of Window Size. Human Factors, 25, 683-692.

Hanson, R.H., Payne, D.G., Shively, R.J., & Kantowitz, B.H. (1981). Process control simulation research in monitoring analog and digital displays. Proceedings of the Human Factors 25th Annual Meeting, 154-158.

Keister, R.S., & Gallaway, G.R. (1983). Making Software User Friendly: An Assessment of Data Entry Performance. Proceedings of the Human Factors 27th Annual Meeting, 1031-1034.

Lamberti, D.M., & Newsome, S.L. (1989). Presenting abstract vs. concrete information in expert systems: What is the impact on user performance? International Journal of Man-Machine Studies, 6(30), 27-45.

Mehlenbacher, B., Duffy, T., & Palmer, J. (1989). Finding information on a menu: Linking menu organization to the user's goals. Human-Computer Interaction, 4, 231-251.

Payne, S.J., Sime, M.E., & Green, T.R.G. (1984). Perceptual structure cueing in a simple command language. International Journal of Man-Machine Studies, 21, 19-29.

Potosnak, K. (1988). Setting objectives for measurably better software. IEEE Software, 89-90.

Roske-Hofstrand, R. J., & Paap, K. R. (1986). Cognitive networks as a guide to menu organization: An application in the automated cockpit. Ergonomics, 11(29), 1301-1311.

Ryack, B. L. (1987). A computer-based diagnostic/information patient management system for isolated environments: MEDIC Ten Years Later. Naval Submarine Medical Research Laboratory Report No. 1089.

Schott, F., & Olson, M. (1988). Designing usability in systems: Driving for normalcy. Datamation, 34, May 15, 68-70+.

Sein, M.K., & Bostrom, R.P. (1989). Individual differences and conceptual models in training novice users. Human-Computer Interaction, 4, 197-229.

Smith, S. L., & Mosier, J. N. (1986). Guidelines for Designing User Interface Software. MTR 10090 The MITRE Corporation, Bedford, Massachusetts.

Trollip, S.R., & Sales, G. (1986). Readability of computer-generated fill-justified text. Human Factors, 28, 159-167.

Tullis, T.S. (1980). Human performance evaluation of graphic and textual CRT displays of diagnostic data. Proceedings of the Human Factors Society, 310-311.

Tullis, T.S. (1983). The formatting of alphanumeric displays: A review and analysis. Human Factors, 25, 657-682.

Whiteside, J., Bennett, J., & Holtzblatt, K. (1987). Usability Engineering: Our Experience and Evolution. DEC-TR 547.

Wright, P., & Reid, F. (1973). Written information: Some alternatives to prose for expressing the outcomes of complex contingencies. Journal of Applied Psychology, 57, 160-166.

Author Notes

Acknowledgements

The authors wish to thank HM1(SS) Gregory Prunier, HMCS(SS) Mikel Middleton and HM1 (ret.) Patrick Flaherty for their valuable help during the study. The authors also wish to thank Russell Eberhart and Roy Dobbins of the Johns Hopkins University Applied Physics Laboratory, Baltimore, Maryland, William Pugh and Anthony Gino of the Naval Health Research Center, San Diego, California, and David Southerland and Karen Fisherkeller of the Naval Submarine Medical Research Laboratory, Groton, Connecticut, for providing the program versions used in this study.

Appendix A
Descriptions of the Versions

Appendix A
Descriptions of the Versions

Version 1 begins with an introductory screen explaining the purpose of the program and instructions for its use. Also contained is a message to the user to rely on his/her professional judgement when the diagnosis suggested by the program does not totally agree with the user's assessment. A main menu offers choices such as "Real Case", "Simulated Case", and "Display Treatment". Selection is made by moving the cursor to the desired selection and hitting Enter or by typing in the number associated with the selection. Patient information screens follow which require the information to be typed into the appropriate data entry fields. A second menu screen appears offering Data Entry Options such as "Go To History Pages", "Make Diagnosis", and "Return to Main Options". The pace of data entry is user controlled. Entry errors are corrected by typing over the error, and error messages explain the nature of the error. Help is available for every field via the "?" key.

The screens addressing the history of the illness and the physical exam findings present groups of related questions with available responses appearing in list form beneath each question. Help is again available for every item via the "?" key. The user selects a response by moving the cursor to it and highlighting it with the Enter key. Highlighting of the desired response is indicated on the screen by a double asterisk appearing to its left. The following keys are used to navigate through the version: up, down, right and left arrows = up, down, right and left cursor movement, the Tab key moves the cursor to the next question on the screen, shift-Tab to the previous question, P = previous page, N = next page, and X = exit. These instructions appear at the bottom of every screen. All choices for a given screen are saved when N or P are used. Users can return to previous items to change a response using the arrow keys and the P key. The quickest way to move through the version is to use the left hand on the Tab key to move between questions and the right hand on all other keys for movement within questions and between pages.

The screens requesting the site of pain at onset and at present display a diagram of the abdomen. The operator can choose more than one area, and each area chosen is filled in with color on the graph, resulting in a picture of the total affected area.

Following all data entry, the diagnostic summary is presented as a histogram in graph form. The probability associated with each possible diagnosis is shown by a vertical bar appearing on a background which indicates critical values with horizontal lines of dashes.

Version 1 uses color against a black background to communicate the functions of the various program parts. Grey lettering is used overall. Instructions appearing at the bottom of each screen are in green lettering and data entry field labels appear in yellow. Each screen is enclosed in a thin blue border and is functionally titled in red. All textual

material is left-margin justified, using conventional spacing with conventional use of upper and lower case letters. Labels appear in all upper case letters.

Version 2 begins with the main menu. Selection is made by moving the cursor to the desired item and hitting "Enter" or by typing in the number associated with the selection. Movement through the version is accomplished with the following keys: the arrow keys for the corresponding movements, Pg Up and Pg Dn to go on to the next or return to the previous pages, Esc = quit and F1 for help. These instructions appear at the bottom of each screen. Version 2 uses dialogue boxes in its interaction with the user. Each screen contains a group of related questions. The user moves the cursor to the question to be answered and hits Enter to select it. This opens up a box. Within the box, either fields for data entry or acceptable responses are presented. The user either types in the value or moves the cursor to the desired response and then hits Enter to register an answer. For fields requiring typed in values, prompts as to the acceptable format appear at the bottom of the screen below the general instructions. For fields requiring specific answers, the list of options is displayed within the box. It is sometimes the case that not all options can be displayed at once. The user must move the cursor to the bottom of the list and continue pressing the down arrow key to reveal the remaining items. No prompt appears to indicate the necessity of this action to the user. Once a choice is registered with the Enter key, the box disappears and the response appears on the screen in the field following the question. The user controls the pace of data entry. Users can go back and make changes or corrections by re-opening the box associated with the desired item and typing over the error. The correction appears in the box while the original response remains on the screen for comparison until the Enter key is hit again.

Questions requesting the site of pain at onset are handled the same as all other questions. Not all the available responses are displayed at once in the box, requiring the user to scroll down the list to view all options. Only one response option can be chosen to indicate the pain site in Version 2. The diagnostic summary is presented as a list of possible diagnoses with their associated probabilities appearing to the right of each diagnosis. The list is presented in the same order for every case.

Version 2 is presented against a pale blue background with the current screen of questions contained in a medium blue square. As in Version 1, color is used to communicate the different parts of the program. The dialogue boxes are grey. Help appears in a green box. The red lettering, used for questions and field labels turn white to indicate which item the cursor is resting on. Yellow lettering is used for functional screen titles and for the instructions that appear along the bottom of each screen. All textual material is left-margin justified, using conventional spacing with conventional use of upper and lower case letters. Labels appear in upper and lower case letters but all data entered appears in upper case.

Version 3 opens with a main menu. Selections are made by moving the cursor to the desired item. (Hitting the Enter key is the required next step although this instruction does not appear in the prompts at the bottom of the screen.) The following keys are used to move through the version: the up and down arrow keys for their corresponding movements, Ctrl N = next page, Ctrl D = delete, ? = help, and the ^ key = quit. These instructions appear at the bottom of each screen. The Enter key is used to register a choice or response, although this is not stated. For the menu screen requiring the user to choose the diagnostic module desired (the sixth screen presented), an additional step is added. The user must use the movement keys to position the cursor on the desired response, use the Insert key to select it, then hit the Enter key to register the choice. If the user hits Enter at the desired choice without first hitting Insert, he/she is returned to the first screen of the version.

Each screen presents the user with numbered items consisting of labels or phrases. When the user positions the cursor on the item to be answered, a key appears at the bottom of the screen. This key presents the available responses with a number associated with each. The user types in the number assigned to his/her choice. That number then appears after the item. When the user hits the Enter key, the number changes to the text of the chosen response.

Each screen's items are related although they are not separated into groups within each screen. For example, on the page titled "Other Patient Symptoms", the item "bowels" is followed by phrases that appear to be response choices related to bowel status rather than separate items in themselves (such as "constipated" and "blood in stool") which are then immediately followed by the item "urination" followed by other phrases that appear to be choices related to the item "urination". This is in contrast with the grouping that appears in versions 1 and 2 which presents response choices related to an item as an indented list under or next to the item itself. The user can move up the screen to correct or change responses on that screen but cannot return to a previous screen to make changes.

The screens relating to pain site require the user to choose the number corresponding to "yes" or "no" in answer to an item indicating the presence of pain, followed by choosing the number corresponding to "yes" or "no" for each of twelve subsequent items identifying a possible pain site. The diagnostic summary states the probability associated with the most likely diagnosis.

Version 3 uses white lettering on a black background for all program components. In this version, all textual material is fill justified, using conventional spacing. Conventional use of upper and lower case letters appears in menu screens only. All items requiring a response appear in upper case.

Appendix B
AVERAGES

Appendix B

AVERAGES

Possible ranges - overall score: -33 to 33, item scores: -3 to 3. (Actual ranges are provided in the bottom row of each cell.) N=12 for all cells except those indicated by asterisks (*).

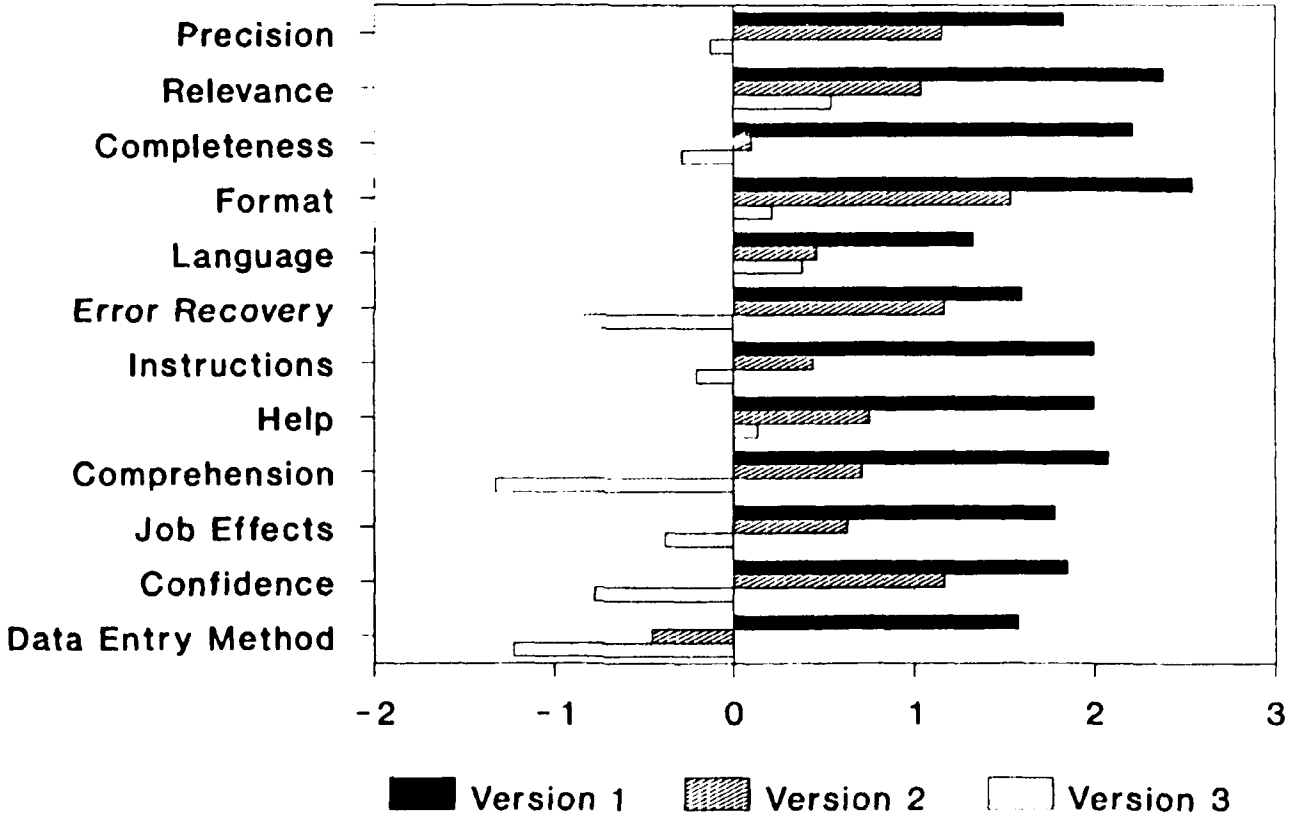
	Version 1	Version 2	Version 3
Overall Satisfaction	21.19 14 to 31.5	9.52 -4 to 28.25	-2.58 -25.25 to 11.25
Precision	**1.83** 0 to 3	**1.16** -1.5 to 3	**-0.13** -2.5 to 3
Relevance	**2.38** 1 to 3	**1.04** -2.5 to 3	**0.54** -2.5 to 2.5
Completeness	**2.21** 0 to 3	**0.10** -2.5 to 3	**-0.29** -2.5 to 2.5
Format of Output	**2.54** 1 to 3	**1.54** .5 to 3	**0.21** -2.5 to 2
Language	**1.33** 0 to 2	**0.38** -2 to 1.5	**0.46** -1 to 1.5
Error Recovery	**1.60** -2.25 to 3	**1.17** -2.25 to 3	**-0.83** -3 to 2.5
Documentation - Instructions	**2.00** 1 to 3	**0.94** -1.25 to 3	**-0.21** -2.75 to 1.25
Documentation - Help	**2.00** 1 to 3**	**0.75** -3 to 3***	**0.13** -2 to 2.75
Understanding of System	**2.08** .5 to 3	**0.71** -2 to 3	**-1.33** -3 to 0
Job Effects	**1.73** -.25 to 3	**0.63** -1 to 3	**-0.38** -2.75 to 3
Confidence in System	**1.85** 0 to 3	**1.17** 0 to 3	**-0.78** -3 to 3
Data Entry* Method	**1.58** .25 to 3	**-0.06** -2.25 to 2.75	**-1.23** -3 to 2

* The questions on the method of data entry were included in the questionnaire for the purpose of gathering information only and were not included in the calculation of the overall usability score.

** N = 11.

*** N = 10.

Average Item Scores



Appendex C
Screens Evaluation

Tally of Questionnaire Responses

Note: Items that were reverse-scored on the questionnaire have been "re-reversed" such that for all items the most positive response is on the left and the most negative on the right.

Version 1

	3	2	1	0	-1	-2	-3	
Precision								
high	***	******	**	*				low
definite	*	******	*	*	*			doubtful
Relevance								
useful	*****	******	*					useless
relevant	*****	*****	*					irrelevant
Completeness								
sufficient	*****	******		*				insufficient
adequate	***	*******		*				inadequate
Format								
simple	********	**	*		*			complex
readable	*********	***						unreadable
useful [1]	*****	******	*					distracting
organized [1]	********	**	**					cluttered
professional [1]	*******	**	***					unprofessional
easy to understand [1]	*********	***						difficult to understand
Language								
simple	***	*****	**	*	*			complex
powerful		***	******	***				weak
easy to use	*****	******	*					difficult to use
ErrorRecovery								
simple	***	*******		*			*	complex
fast	***	******	*		*	*		slow
superior	***	***	****	*		*		inferior
complete	***	******	**			*		incomplete
easy to access [1]	****	******	*				*	difficult to find
easy to understand [1]	******	*****	*					difficult to understand
easy to use [1]	******	***	**				*	difficult to use

Tally of Questionnaire Responses, Version 1, cont.

	3	2	1	0	-1	-2		
Documentation - Instructions								
clear	******	***						hazy
available	**	*******						unavailable
complete	*	**********						incomplete
current	****							obsolete
Documentation - Help								
clear	*****	*****	*					
available	****	*****	*		*			*
complete	*****	*****	*					
current	****	*******						
relevant [1]	******	**	***					+
easy to access [1]	********	***						
Data Entry [1]								
speedy [1]	****	***	***					**
simple [1]	*******	***	*					
easy to use [1]	*********	***						confusing to use
efficient	*	****	*	**	**	*	*	error prone
Understanding of System								
sufficient	***	******	**	*				
complete	*****	*****	**					
comfortable [1]	*******	****	*					
in control [1]	******	****	*	*				
Anticipated Job Effects								
liberating	****	****	*	**	*			
significant	**	******	**	**				
good	**	*******	*	**				
valuable	**	********		**				worthless
Confidence								
high	***	*****	***	*				low

[1] These items were not part of the original questionnaire developed by S. Pearson, Ph.D. and therefore were not used in calculating the overall usability score. They were included in this questionnaire for the purpose of gathering additional feedback from the users.

Appendix D
Screens Evaluation

Tally of Questionnaire Responses

Note: Items that were reverse-scored on the questionnaire have been "re-reversed" such that for all items the most positive response is on the left and the most negative on the right.

Version 2

	3	2	1	0	-1	-2	-3	
Precision								
high	****	***	*	**	*	*		low
definite	***	*	***	***	**			doubtful
Relevance								
useful	***	****	*		*	*	*	useless
relevant	****	***	*	*	*	**		irrelevant
Completeness								
sufficient	****		*	*	*	***	**	insufficient
adequate	***		*	**	***	***		inadequate
Format								
simple	***	****	***		*	*		complex
readable	****	******	**					unreadable
useful [1]	***	***	******	*				distracting
organized [1]	****	***	*	*	*	**		cluttered
professional [1]	****	****	**	**				unprofessional
easy to understand	***	*****		**		**		difficult to understand
Language								
simple	**	****		****	*		*	complex
powerful			***	******	**	*		weak
easy to use	*	****	****		***			difficult to use
Error Recovery								
simple	******	***	**				*	complex
fast	*****	**	**	*	*		*	slow
superior	*	*****	*	***		*	*	inferior
complete	**	**	****	*	**		*	incomplete
easy to access [1]	*	*****	**	*	**	*		difficult to find
easy to understand [1]	**	***	**	*	**	**		difficult to understand
easy to use	***	***	*	*	**	**		difficult to use

Tally of Questionnaire Responses, cont.

	3	2	1	0	-1	-2	-3	
Documentation - Instructions								
clear	****	***	**	*	**			hazy
available	***		***	**	*	**	*	unavailable
complete	***	**		**	**	***		incomplete
current	***	***	***	***				obsolete
Documentation - Help								
clear	****	*	*	**			**	hazy
available	***	**	*	*	*	*	*	unavailable
complete	***	**		***		*	*	incomplete
current	**	*	**	***		*	*	obsolete
relevant [1]	**	*	**	*	*	**	*	useless
easy to access [1]	***	****		**		*		difficult to find
Data Entry [1]								
speedy [1]	**	**		***		***	**	tedious
simple [1]	*	***		**		****	**	complex
easy to use [1]		****		*	**	*****		confusing to use
efficient [1]	*	*****	*	*	*	**	*	error prone
Understanding of System								
sufficient	*	***	***	*	***	*		insufficient
complete	*	*****	**	*	*	**		incomplete
comfortable [1]	**	****	*	*	****			intimidating
in control [1]	**	***	**	**	**	*		helpless
Anticipated Job Effects								
liberating	*	*	****	*	***	*	*	inhibiting
significant	*	***	**	****	**			insignificant
good	*	***	**	******				bad
valuable	*	***	**	****	**			worthless
Confidence								
high	*	****	***	****				

[1] These items were not part of the original questionnaire developed by S. Pearson, Ph.D. and therefore were not used in calculating the overall usability score. They were included in this questionnaire for the purpose of gathering additional feedback from the users.

Appendix E
Screens Evaluation

Tally of Questionnaire Responses

Note: Items that were reverse-scored on the questionnaire have been "re-reversed" above such that for all items the most positive response is on the left and the most negative on the right.

Version 3	3	2	1	0	-1	-2	-3	
Precision								
high	*	*	**	***	**	***		low
definite	*	*	**	***	***		**	doubtful
Relevance								
useful	*	*	****	***	**		*	useless
relevant	**	**	***	**	**	*		irrelevant
Completeness								
sufficient	*		****	*	****		**	insufficient
adequate		*	**	****	***	*	*	inadequate
Format								
simple		**	**	*	*	****	**	complex
readable	**	*****	**	*	*		*	unreadable
useful [1]		**	**	**	*	**	***	distracting
organized [1]	*	**	*		**	***	***	cluttered
professional [1]	*	**	*	****	*	***		unprofessional
easy to understand [1]	*	**	**	**	*	***	*	difficult to understand
Language								
simple	**	**	****	***		*		complex
powerful		*	*	*******	**	*		weak
easy to use		****	**	*	***	**		difficult to use
Error Recovery								
simple	*	*		**	**	***	***	complex
fast	*	**	*	**	*	**	***	slow
superior		*	*	***	*	***	***	inferior
complete		*	***	*	**	***	**	incomplete
easy to access [1]		*	*	**	*	***	****	difficult to find
easy to understand [1]		*	*	****		*****	*	difficult to understand
easy to use [1]			*	****	*	****	**	difficult to use

Version 3, continued

	3	2	1	0	-1	-2	-3	
Documentation - Instructions								
clear		**	*	**	***	***	*	hazy
available		**	***	*****	*	*		unavailable
complete		*	*	****	**	***	*	incomplete
current	*	*	**	****	***		*	obsolete
Documentation - Help								
clear	**	*	**	***	*	***		hazy
available	*		*****	**	*	*	*	unavailable
complete	*		**	*****	**	*		incomplete
current	*	***	**	**	*	**		obsolete
reelevant [1]		***	***		**	**	*	useless
easy to access [1]		**		**	***	***	*	difficult to find
Data Entry [1]								
speedy [1]		**			*	***	*****	tediious
simple [1]	*	*	*	*	***	**	***	complex
easy to use [1]		*	**	*		****	****	confusing to use
efficient [1]		*	**	**	*	***	***	error prone
Understanding of System								
sufficient			*	**	***	****	**	insufficient
complete				**	*****	****	*	incomplete
comfortable [1]				*	*****	**	****	intimidating
in control [1]				*	****	*****	**	helpless
Anticipated Job Effects								
liberating	*	*	*	**	***	***	*	inhibiting
significant	*	*	*	**	**	*****		insignificant
good	*	*	**	***	**	**	*	bad
valuable	*	*	**	***	*	**	**	worthless
Confidence								
high	*	*	*	*	**	*****	*	low

[1] These items were not part of the original questionnaire developed by S. Pearson, Ph.D. and therefore were not used in calculating the overall usability score. They were included in this questionnaire for the purpose of gathering additional feedback from the users.

Appendix F

Appendix F
Usability Session Observations

Version One

The quickest way to move through version 1 is to use the left hand to move between questions with the Tab key and the right hand for all other keystrokes. One out of 12 users actually used both hands. In fact only two of the users utilized the Tab key at all. All others used the arrow keys to move through all response options of a given question to get to the start of the next question. This may be due to the fact that although the Tab key is listed at the bottom of each screen as a movement key, it is not specified that the particular movement produced by the Tab key is one of jumping from one question to the next.

Eight users accessed the Help files from various screens in this version without difficulty. At the Diagnostic Summary page, five users chose and executed the option to review their previous responses without difficulty. One user was confused by the graphic representation of the probable diagnoses on the Diagnostic Summary page, asking if the values shown on the bar graph indicated the percentage of users choosing the wrong diagnosis.

One user always used the arrow keys to move to the last option on a screen before hitting the N key to move to the next screen, as if he felt he had to go to the "bottom" of the page before "turning" it! (This is in fact not the case, and time can be saved by using the N key from anywhere on a screen to move to the next screen).

Version Two

Enter must be used to open a dialogue box before version 2 will accept any type of input. For questions requiring the values to be typed in, 11 out of 12 users did not understand this and attempted to key in their responses without first opening the dialogue box. Nine of these users discovered the proper keystroke sequence through trial and error; two users became frustrated and called for assistance. Even after the need to open the box was understood, five users continued to try to type in responses directly before resorting to opening the box on subsequent questions. Three users apparently forgot the procedure when starting their second and third cases and initially again tried to type in their responses directly. One user, frustrated and amused by the need to open the dialogue box, called a staff member over and said, "This one's harder to use. I don't see why you have to go to the box. Why can't you just type it? That's... (demonstrates for staff member) I have to get a box! (chuckling as he continues) Ask for a box!"

Also, as described earlier, the box presenting options for Site of Pain does not present all options at once. The user must use the arrow key to scroll the rest of the list into view, although this is not stated on the screen. This confused five out of 12 of the users. One user, after opening the box three times and failing to find the desired response, left this question blank. One asked for assistance, and three discovered the other options by trial and error.

A user suggested that a simple solution would be to have the last item in the box be "scroll down for more". Also, the Site of Pain box permits only one response. Two users indicated that the inability to combine abdominal areas prevented them from inputting the total area affected.

Some of the Help files for version 2 were not available at the time of testing; however, all nine users who attempted to access Help experienced no difficulties in reaching the part of the program where the Help files would be located.

In addition, the one user who wished to review past responses successfully chose and executed "Review Past Encounter" from the main menu.

The field for indicating the patient's temperature does not accept decimal values. Two users stated that they found this to be inadequate for entering accurate data. In addition, one user indicated that the range of acceptable values for respiration rate accepted by this version was inadequate, stating, "People breath more than twenty respirations per minute. More than twenty is not uncommon if you're under stress."

Version Three

Version 3 contains unique instructions at the menu asking for the desired diagnostic module (the sixth screen). After moving the cursor to the desired response, this screen requires the added step of hitting the Insert key. If the user hits Enter instead (which is the proper sequence for all the other screens), he/she is returned to the first screen in the version. The prompt "Insert to select" does appear on the bottom of the screen with the other prompts. However, once Insert is used, then the Enter key must be hit to register the selected response. This instruction does not appear at the bottom of the screen. Eleven out of 12 users did not notice the change in the instructions at the bottom of the screen and used Enter to try to register their choice, thereby returning to the start of the program. They then re-keyed the information asked for in the four screens appearing between the first screen and sixth screen, used Enter again on screen six and were again returned to the start of the program! The usability lab staff allowed users to run through this loop three times before intervening and prompting them to review the instruction lines on the sixth screen. After reading the instructions and appropriately hitting the Insert key, four of the 11 then sat waiting for the program to respond, and one hit nearly every key but the Enter key in an attempt to register his response (he was reluctant to use the Enter key on this screen after the trouble it had caused previously)! Staff again intervened to inform them that the Enter key now needed to be used. The one user who did notice the changed instructions on this screen properly used the Insert key the first time he was presented with screen six and then waited for a program response (until staff informed him that Enter was his next move). Five of the users repeated the above error of failing to use the Insert key for screen six on their second and third cases. The users expressed their frustration with screen six in a variety of ways including holding their head in their hands, throwing their hands into the air, moaning, and glaring angrily at the screen. One user called on the staff for help saying,

"It just doesn't do anything. You *have* to hit the Return key otherwise it doesn't do anything, then...I can't even enter the information!"

Six users experienced difficulty responding to the question asking for their ship's name. They attempted to type in a name which disappeared when Enter was used. Pressing Enter a second time revealed a list of ship names and corresponding numbers that took an average of two minutes to scroll through and read. All six tried unsuccessfully to escape from the list. Five users entered the number corresponding to their choice, but the sixth user typed in a name from the list, which the version did not accept, and needed to view the entire list again to find the number. The remaining six users bypassed the question altogether.

Five users tried to return to previous screens to correct input only to discover that this version does not allow this.

Four users accessed Help successfully.

The first item on the Site of Pain screen asks the user to indicate the presence or absence of pain by typing 1 for "yes" or 2 for "no". It then lists the possible pain sites requiring the user to indicate if each site is affected by typing 1 for "yes" or 2 for "no" for each possible site in turn. Three users tried to type in the affected location directly. Two of the users discovered that skipping the sites with the answer "no" and typing only the "yes" responses was a shortcut for moving through this screen.

Several users expressed dissatisfaction with what they viewed as discrepancies in version 3. The second screen asks the user *to* enter either the patient's Social Security number *or* name. The third screen then places this input in the name field even if it is the Social Security number that was entered. Two users commented on this and several took the trouble to erase the Social Security number from the name field before moving on. One user expressed frustration that the field for entering the patient's Social Security number did not accept the hyphens that normally divide the number. Two users pointed out screens where the prompt "Press any key" is not accurate since only the Enter key evoked a response. Four users did not receive the version's diagnosis at the end of their interaction and instead were returned to the first screen.

As described earlier, version 3 presents items and possible response options to those items in a list format running down the full screen instead of grouping the items and their response options separately. This produces a screen that appears very full. Several users responded to this by sighing heavily whenever one of these screens appeared.

Appendix G

Appendix G
Interview Comment Tally

#	Comment
Use of Color	
(4)	Color helped direct the user.
(4)	Color in the graphics was best combination for quick extraction of information.
(1)	Color in the graphics was distracting.
(1)	Color was interesting.
Data Entry Method	
(4)	Preferred data entry by highlighting with cursor movement.
(2)	Didn't like data entry by opening dialogue boxes - mentally fought doing this.
(1)	Data entry in version 2 (dialogue boxes) quickest to use.
(1)	Preferred data entry by typing in response.
Order of Contents	
(2)	Liked how version 1 followed S.O.A.P. note.
(1)	Liked separation of items into History and Exam sections.
Diagnostic Summary	
(2)	Liked seeing listing of all diagnoses with their probabilities.
(1)	Suggestion: List diagnoses in descending order of probability.
(1)	Liked seeing the most likely and the least likely diagnoses.
(1)	Would like to see 2 or 3 alternate diagnoses given with primary diagnosis.
(1)	Liked graphic display of diagnoses - quicker than having to read.
Miscellaneous Comments	
(4)	Version 1 easier to work with.
(3)	Suggestion: Have cursor automatically move to next item after completion of each ite
(1)	Liked the idea of computer-aided diagnosis - it is a good review of diagnostic guidelin
(1)	The program should not state that the corpsman is wrong; instead it should suggest checking other sections.
(1)	Liked being able to enter data with only one hand.

UNCLASSIFIED
SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1a. REPORT SECURITY CLASSIFICATION
Unclassified

1b. RESTRICTIVE MARKINGS

2a. SECURITY CLASSIFICATION AUTHORITY

2b. DECLASSIFICATION/DOWNGRADING SCHEDULE

3. DISTRIBUTION/AVAILABILITY OF THE REPORT
Approved for public release; distribution unlimited

4. PERFORMING ORGANIZATION REPORT NUMBER(S)
NSMRL Report 1172

5. MONITORING ORGANIZATION REPORT NUMBER(S)
NA

6a. NAME OF PERFORMING ORGANIZATION
Naval Submarine Medical Research Laboratory

6b. OFFICE SYMBOL (If Applicable)

7a. NAME OF MONITORING ORGANIZATION
Naval Medical Research and Development Command

6c. ADDRESS (City, State, Zip Code)
Box 900, Naval Submarine Base NLON, Groton, CT 06349-5900

7b. ADDRESS (City, State, Zip Code)
National Naval Medical Center, Bldg 1, Tower 12, Bethesda, MD 20889-5044

8a. NAME OF FUNDING/SPONSORING ORGANIZATION
Same as 7a

8b. OFFICE SYMBOL (If Applicable)

9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER

8c. ADDRESS (City, State, Zip Code)
Same as 7b

10. SOURCE OF FUNDING NUMBERS

PROGRAM ELEMENT NO.	PROJECT NO.	TASK NO.	WORK UNIT ACCESSION NO.
63706N	M0095	.005	5010

11. TITLE (Include Security Classification)
A comparison of the usability of three versions of a computerized medical diagnostic assistance program for abdominal pain

12. PERSONAL AUTHOR(S)
E. F. Chouinare, B. L. Ryack, and D. M. Stetson

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Interim	FROM ____ TO ____	1991 July 23	54

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	Computer assisted diagnosis; MEDIC Program; Abdominal Pain

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

(U) Three versions of a computerized medical diagnostic assistance program for abdominal pain were tested for ease of use, ease of learning, user satisfaction, and time to complete the "Pain Site" screen. A higher satisfaction rating was associated with visual grouping of related items, ordering of items to coincide with the usual medical examination, the use of color to highlight information and direct the user, and minimal and consistent steps for data entry. Preference for graphic and list formats for the presentation of the diagnostic summary information was nearly equally divided. Longer learning time was associated with inconsistent rules for the handling of completed screens. Longer time to complete a screen was associated with a lack of grouping of related items, multiple steps for data entry, a lack of instructions identifying required and optional data entry items, and exclusive use of upper case test. Confidence in the program-generated diagnoses was found to increase with user satisfaction.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT
[X] UNCLASSIFIED/UNLIMITED [ ] SAME AS REPORT [ ] DTIC USERS

21. ABSTRACT SECURITY CLASSIFICATION
Unclassified

22a. NAME OF RESPONSIBLE INDIVIDUAL
Susan D. Monty, Publications Office

22b. TELEPHONE (Include Area Code)
(203) 449-3967

22c. OFFICE SYMBOL

DD Form 1473, JUN 86 — Previous editions are obsolete. S/N 0102-LF-014-6603

SECURITY CLASSIFICATION OF THIS PAGE
UNCLASSIFIED