A Comparison of Evaluative Techniques to Improve the Reliability of Maintenance Documentation

Bonnie Lida Rogers
Christopher J. Hamblin
Alex Chaparro

Follow this and additional works at: https://corescholar.libraries.wright.edu/isap_2005

Part of the Other Psychiatry and Psychology Commons

Repository Citation
https://corescholar.libraries.wright.edu/isap_2005/33

This Article is brought to you for free and open access by the International Symposium on Aviation Psychology at CORE Scholar. It has been accepted for inclusion in International Symposium on Aviation Psychology - 2005 by an authorized administrator of CORE Scholar. For more information, please contact corescholar@www.libraries.wright.edu, library-corescholar@wright.edu.
A COMPARISON OF EVALUATIVE TECHNIQUES TO IMPROVE THE RELIABILITY OF MAINTENANCE DOCUMENTATION

Bonnie Lida Rogers\textsuperscript{1}, Christopher J. Hamblin\textsuperscript{1}, and Alex Chaparro\textsuperscript{1,2}

\textsuperscript{1}National Institute for Aviation Research
\textsuperscript{2}Department of Psychology
Wichita State University
Wichita, Kansas

The purpose of this research was to investigate the applicability of usability methods in evaluating aviation maintenance documentation and to document the types of errors found. A diverse set of participants were recruited to participate in the evaluations in order to document how experience and training affect error detection. The results are similar to the findings of usability testing of software and web design – system experts and users identify unique errors and roadblocks.

Introduction

Maintenance procedures and information have been cited as primary factors contributing to maintenance errors (Dekker, 2002; Hobbs & Williamson, 2003; McDonald, Corrigan, Daly, & Cromie, 2000; Reason & Hobbs, 2003). A review of Naval Aviation Maintenance mishaps that occurred between 1990 and 2003 (Ricci, 2003) showed that 28\% of the accidents involved problems in maintenance procedures including missing procedural steps, incorrect sequence of steps, inadequate procedures for inspection and troubleshooting, and incorrect technical information and diagrams. However, because mishaps are rare events, they underestimate the frequency of incidents in which poor documentation resulted in maintenance errors. Also, mishaps do not account for the other effects of poor documentation including the costs of incorrectly executed or slowed maintenance.

Maintenance documentation has recently begun to receive attention from academic researchers, the Federal Aviation Administration, and manufacturers. Many of these studies have focused on employing human factors principles to document and workcard design (Drury, Sarac, & Driscoll, 1997; Patankar & Kanki, 2001; Patel, Prabhu, & Drury, 1993). More recently, the methods and techniques employed by the aviation industry to develop maintenance documentation have also been investigated. Chaparro and Groff (2001) identified a number of problems with the development of maintenance documentation, including: reactive rather than proactive evaluation of the manuals, the limited use of aircraft maintenance technicians\textsuperscript{'} (AMTs\textsuperscript{}) input and procedure validation, the absence of systematic attempts to track error, and the lack of standards for measuring document quality.

In addition to improving maintenance documentation through design guidelines and manual usability, the accurate and clear communication of information is also critical. In other words, the AMT\textsuperscript{'}s interpretation of the procedure must match the intent of the writer for successful maintenance task completion. A mismatch has two likely outcomes. First, the AMT may become frustrated and call customer support for assistance in performing a procedure; or secondly, the AMT may \textquotesingle work-around\textquotesingle the procedure. The \textquotesingle work-around\textquotesingle approach entails trying to deduce the writers\textquotesingle intent when a procedure is confusing, or the information is incomplete or inaccurate.

This is not an uncommon occurrence. A study by Hobbs and Williamson (2000) conducted for the Australian Transportation Safety Bureau found that 67\% of AMTs report having been misled by maintenance documentation, 47\% report having opted to perform a maintenance procedure in a way they felt was superior to that described by the manual, and 73\% of mechanics surveyed reported failing to refer to maintenance documents either occasionally or often. Chaparro, et al. (2002) also found that 64\% of AMTs reported finding their own way of performing a procedure. Nearly 60\% of AMTs reported continuation of an unfamiliar task despite not being sure if they were performing it correctly (Hobbs & Williamson, 2000). Similarly, McDonald et al. (2000) reported that 34\% of routine maintenance tasks are performed in ways different than outlined in the maintenance documentation (MD).

Surveys reveal that aviation manufacturers rely on aircraft maintenance technicians (AMTs) to identify problems in MD (Chaparro et al., 2002). Most corrections to the MD are post-release through reports of problems by AMTs, called Publication Change Requests (PCRs). However, assuming that AMTs will report errors in maintenance procedures may be incorrect. Chaparro et al. (2002) found that 53\% of AMTs reported only occasionally, rarely, or never reporting errors they found.
AMTs are often very good at deriving a plausible interpretation of incomplete information by drawing on their knowledge and that of other mechanics. This ability may result in an AMT misinterpreting procedures in such a manner that it is difficult to discover the error in their interpretation and subsequent actions. Although the AMTs’ training and experience may allow them to correctly identify the writers intent, this will not always be the case. This uncertainty can be reduced by the proactive approach of assessing documentation quality before publication using tools originally developed to test the usability of computing software programs and documentation.

The purpose of these experiments is to investigate the applicability of two usability methods in evaluating aviation documentation and to document the types of errors found in MD. A diverse set of participants were recruited to participate in the evaluations in order to document how familiarity and training effect error detection.

Based on interviews with aviation technical writers, two usability techniques (described below) were chosen for the evaluation: Cognitive Walkthrough (CW) and User Performance (UP). Two experiments were performed to evaluate each of these evaluative methods.

Cognitive Walkthrough (CW) is a review technique in which evaluators review or “walk through” each step of a procedure to identify incorrect technical and factual information, poor wording choices, and inadequate information. Participants are instructed to visualize performance of each step as if they were doing the task. Normally, CW reviews are conducted in the early stages of document development to make corrections and changes before actual user testing.

User Performance Evaluation (UP) involves a participant physically performing a task. Participants are chosen who are not familiar with the task procedure or its development, to ensure that they are representative of users (AMTs) and the procedure can be evaluated without the potential biases arising from knowledge of the developer’s, i.e. technical writer’s, intent or familiarity with the system’s design. Two forms of the UP were also compared: 1) a single user (SU) (i.e., AMT) performs the evaluation and 2) a two-person team work together, referred to as a Co-discovery (CD) user performance technique. In this study, an AMT performs the task as written in the MD and a Customer support engineer observes and makes comments.

Methods

CW Participants

Typically, CW evaluators are “expert” reviewers, familiar with the product’s design and development; however, in this evaluation, we selected both “expert/familiar” and “naïve/ unfamiliar” participants to review the MD in order to investigate the role experience (expert vs. naïve) and training (AMT vs. engineer) play in error detection at earlier stages of document development.

Nineteen participants, 17 male and 2 female, completed the CW evaluation. The participants were assigned to one of four groups (expert vs. naïve) and technical background (engineers vs. AMTs). A total of three expert engineers, 5 expert AMTs, 6 naïve engineers, and 5 naïve AMTs participated in the evaluation. Naïve mechanics and engineers watched a short animated video of the procedure that illustrated the key parts and provided an overview of the task’s process. One naïve engineer participant’s responses were not included in the analysis as she reported more than the combined total of the other members in her group.

UP Participants

A total of ten naïve AMTs and five naïve engineers (all unfamiliar with the new procedural task) from the manufacturer’s service facility participated in the UP Evaluations. Five of the AMTs were assigned to the single-user (SU) evaluation and five were assigned to the Co-discovery (CD) evaluation. The five naïve Customer Service engineers were teamed with the five naïve AMTs in the CD evaluations. All of the participants in this evaluation were male.

Materials

A general aviation aircraft manufacturer provided an unpublished maintenance procedure for the usability testing. This procedure was chosen because 1) it was unfamiliar to the pool of AMTs and their prior experience did not transfer readily to the new design, and 2) a computer simulation and physical prototype were available for use in testing. Prior to the experiments, the maintenance procedure was evaluated by production line mechanics and design engineers familiar with the task to estimate the number and types of errors within the document. The procedure was not modified as it was judged to have a sufficient number and types of errors.
**CW Procedure**

All participants read a paper copy of the MD and were asked to note any errors they found including typos, missing or incorrect information and any instructions that were out of sequence or unclear. Any materials typically referenced while proofing the MD (e.g. engineering drawings) were available to the participants while they reviewed the written procedure. The time required to complete the cognitive walkthrough was recorded upon completion (M = 40 minutes, range 26-70 minutes).

**UP Procedure**

AMTs were instructed to perform the procedure as written in the MD and to verbally describe what they were doing at each step and why they were doing it. In the CD evaluation, CS engineers were to observe. In the S and CD user performance evaluations, both types of participants (CS engineers and AMTs) were asked to verbalize their actions and inform the researchers of any instruction (or part of an instruction) that was incorrect, missing, out of sequence, or unclear. The time required to complete the cognitive walkthrough was recorded upon completion (M = 142 minutes, range 105-210 minutes).

Prior to the experiment all participants were informed of the purpose of the experiment and were asked to read and sign a consent form and privacy statement. Two researchers conducted the evaluations and recorded and coded the comments made by all participants into the error taxonomy, see Results section. A Cohen’s Kappa (κ) of .85 was calculated on a sample of 50 comments reflecting an excellent level of consistency between the coders (Fleiss, 1981). Following the experiments, each participant completed a short background and satisfaction questionnaire.

**Results**

**Error Taxonomy.** To facilitate analysis and interpretation, a coding scheme was developed to categorize the errors identified by the participants. Within the context of this study, errors are defined as those items identified by participants as potential problem areas in the documentation. Four error-type categories and twelve specific reason categories were identified in the evaluations: 1) Technical (tools, values, parts); 2) Language (clarity of wording/terminology, grammar, typos, incorrect information); 3) Graphics (dimensions, part diagram, caption/text); and 4) Procedural (step(s), ordering). The associated corrective actions (add, delete, or change information) suggested by the participants’ comments were also coded for analysis.

**Cognitive Walkthrough (CW)**

The results in Table 1 show that experts (AMTs and engineers) identified more than twice the errors (154 vs. 72) than their naïve counterparts. This is true despite the fact that there were fewer expert participants (n = 8 vs. 10). Both naïve and expert evaluators reported language error types most frequently (naïve, 41; expert, 63), followed by procedural-type errors (naïve, 19; expert, 47).

<table>
<thead>
<tr>
<th>Type of error</th>
<th>CW Naive Engineer</th>
<th>CW Expert Engineer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Technical</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>Language</td>
<td>30</td>
<td>40</td>
</tr>
<tr>
<td>Graphic</td>
<td>15</td>
<td>25</td>
</tr>
<tr>
<td>Procedure</td>
<td>5</td>
<td>10</td>
</tr>
</tbody>
</table>

**Table 1.** Number of errors reported in the CW method by evaluator group.

A review of the comments made by each user group revealed several differences. Comments by naïve participants typically regarded the meaning or interpretation of the text and “what ifs?” (i.e., the absence of instructions regarding what actions to perform if a stated value or condition was not met.)

The experts reported more errors that were factual in nature including incorrect technical values, language, and procedural sequences. This result is not surprising since only individuals familiar (i.e., experts) with the design and operation can readily identify whether descriptive or factual information is incorrect.

These results illustrate the unique contributions made by the different experience (i.e., familiarity) levels of evaluators at an early stage of document development. Because of their familiarity with the procedure, system experts were better able to identify errors in technical information and system descriptions. However, due to their familiarity with the system they were less likely to identify vague, unclear, and imprecise procedural descriptions reported by the naïve participants.
User Performance Testing (UP)

AMT participants who were unfamiliar with the task and used the documentation to actually perform the procedure reported errors more frequently than any group in CW or the Customer Service Engineers in the Co-discovery (CD) method of user performance testing. The CD evaluation method was relatively more effective in identifying errors than the SU method—roughly twice as many total errors were reported by participants using the CD vs. the SU method (CD, 331; SU, 162).

Figure 2. Number of errors reported in SU and CD user performance methods by evaluator group.

A comparison of the contributions made by AMTs and engineers in the CD method show that AMTs identified many more errors (roughly three times) associated with procedural and language than did the engineers. Like the results from the CW, procedural and language errors were again the most frequently cited problems. The most common types of procedural errors were missing steps including the absence of instructions regarding what actions to perform if a stated value or condition was not met, steps for disassembling or reassembling, and simple instructions which aid the AMT frame of reference (e.g., open/close door).

Comparison of CW and UP. Figure 3 illustrates the average number of the four major error types (language, graphic, procedural and technical) reported by participants using the two evaluation methods (CW and UP). These results demonstrate the benefits of performing the maintenance procedure on an aircraft. As illustrated in the differences between the frequency of language and procedural errors in Figure 3, the CW was relatively more effective at detecting language errors while the UP evaluations resulted in greater detection of procedural errors.

Figure 3. Error frequencies as a function of evaluation method and error type.

Procedural errors identified in UP evaluations were most frequently missing steps (n = 95), followed by the need to change the sequencing of the steps (n = 44). Both of these specific reasons were reported more than three times as often in UP as in CW.

Figure 4. Comparison of the corrective actions by evaluation method.

Corrective actions of adding, deleting and changing information were implied when the errors were reported. As illustrated in Figure 4, the majority of these comments for both User Performance (SU and
CD) and Cognitive Walkthrough (CW) techniques requested either changing or adding more information to the procedure. Note that more than twice as many comments requesting that information be added to the procedures were obtained through UP (n = 280) than CW (n = 118).

**Unique Errors.** In many instances, the same error was reported by more than one participant in the experiment; these redundant reports were eliminated and the sums of these single instance or “unique” errors for each method were calculated. Sixty-seven percent of the 226 reported errors in CW and forty-four percent of 493 in UP were unique errors. This analysis also shows that the two techniques (i.e., CW and UP) were not redundant as the CW method had only 21 errors in common with the SU and 45 errors in common with the CD method.

**Satisfaction Measures.** A scale was developed to assess the participants’ satisfaction with the written procedure and was administered following the usability testing. The scale had ten individual statements of satisfaction measured on a 5-point agreement scale; Strongly Disagree to Strongly Agree. A Cronbach’s Alpha of .92 was calculated revealing excellent scale reliability in measuring participants’ satisfaction with the technical documentation (Nunnally, 1978). Three additional statements asked for 1) a judgment of the procedure’s complexity relative to other procedures; 2) whether additional instructions would be needed to complete the procedure; and 3) an open-ended query of what would improve the procedure. Results of the satisfaction measures were analyzed by method, i.e., CW and UP evaluations (Single-User (SU) & Co-Discovery (CD), and by user group, (expert engineer, expert AMT, naïve engineer, and naïve AMT).

Generally, participants in the CW method were more satisfied with the written procedure, giving it a mean rating of 3 or higher (i.e., greater satisfaction) on the ten satisfaction statements and the overall satisfaction query; whereas, those who participated in UP evaluations rated the procedure <3, (less satisfaction) for those statements. The total satisfaction score for the CW group (M = 68.33, SD = 14.86) was significantly higher than for the UP group (M = 54.00, SD = 14.38), t(31) = 2.79, p = .009, d = .93. Responses for the following satisfaction queries were significantly higher for the CW participants in comparison to those tested by UP: “I am satisfied with the number of steps included.” (CW: M = 3.78, SD = 1.11; UP: M = 2.33, SD = .90), t(31) = 4.04, p = .001; “The procedure was clearly written.” (CW: M = 3.47, SD = .96; UP: M = 2.40, SD = .83), t(31) = 3.43, p = .002; “The illustration was helpful.” (CW: M = 3.68, SD = 1.64; UP: M = 2.80, SD = 1.27), t(31) = 2.08, p = .046; and “The amount of information included was useful.” (CW: M = 3.78, SD = 1.00; UP: M = 3.20, SD = .90), t(31) = 1.98, p = .056. Both groups indicated that the procedure needed more instructions and were neutral that this procedure was “more complex than most.”

A comparison of the number of errors reported and satisfaction score reveals that satisfaction scores are negatively related to the number of errors found – as the number of errors reported increases, the level of satisfaction significantly decreased (r = -.66, p < .01).

**Discussion**

The results of this investigation show that 1) User Performance and Cognitive Walkthrough evaluations are complementary techniques for evaluating maintenance documentation, 2) the errors identified by individual participants varied in significant ways according to familiarity (expert vs. naïve) and training (engineers vs. AMTs), 3) procedural and language errors are the most commonly cited errors reported in the maintenance documentation usability testing., and 4) satisfaction levels are higher in a CW evaluation compared to UP evaluations.

**Cognitive Walkthrough (CW).** Most commonly in usability evaluations, the user does not review the task at the early stage of development; however, results from this study show that in this domain (i.e., aviation maintenance), the information from a naïve user (AMT) and naïve engineer may provide the technical writer with valuable feedback as to what areas may need additional clarity and where procedural steps, such as checks and functional tests may need to be added.

Several issues identified by the naïve participants in the CW were later reported as problems in the UP evaluation. For example, three naïve engineers in CW testing reported that the wording “Adjust …until the force needed to close … are the best between them.” needed clarification. In the UP evaluations, this step was also cited as unclear by three of the naïve AMTs in SU evaluations, two naïve engineers and two naïve AMTs in the CD evaluations. When it is not possible to test MD using a UP, CW may be a viable alternative using naïve users (AMTs) and naïve engineers for evaluations.

**User Performance (UP).** User performance testing identified specific areas in the documentation that were incomplete, unclear, or incorrect. Ambiguities
are more salient to the user when they have to convert written statements into action. In addition, physical obstructions that make the procedure difficult or impossible to perform become obvious. The results also demonstrate the benefits derived from having evaluators work as a team.

Problems with language clarity included the use of unfamiliar part names, lack of consistency in the procedure, and subjective language, such as “…seal can be removed.” As one AMT commented, “Does it need to be removed or not?” Another statement in the procedure was “make sure … operates correctly” to which an AMT commented, “What is correctly? Correct gap or correct position?). When unfamiliar part names were referenced, the AMTs would often rely on their experience to identify the relevant part. This was not always sufficient as several of the AMTs volunteered that they would have taken apart or adjusted the wrong component.

Given that the same types of information obtained in usability evaluations of MD are similar to those cited as contributory to accidents and incidents (Ricci, 2003), it would seem that adapting usability techniques to improve MD is a feasible and proactive alternative to the current MD development methods. The two methods tested in this research yielded a significant number of instances in which both inaccurate and unclear information could be corrected before publication.

Additional benefits to employing these methods include increasing technical writers’ awareness of the information necessary for the AMT to perform maintenance and a consideration of the constraints under which the AMT is working. As part of this research, an aviation technical writer’s “toolbox” was developed that outlines evaluative methods which have been adapted for aviation technical documentation. The toolbox consists of descriptions of each evaluation technique, guidelines for using the methods, and various supporting documents (questionnaires, data collection forms, etc.) that can be used during the evaluations. The toolbox is available at http://www.niar.wichita.edu/humanfactors/toolbox/default.htm .

References


