

AFOTEC Questionnaire Handbook

Revised August 2000

| Index | Pg. |
|---|------------|
| I. Introduction | 2 |
| II. General Guidelines | 3 |
| III. Construction Guidelines | 4 |
| IV. Data Collection and Administration Guidelines | 15 |
| V. Questionnaire Analysis | 17 |
| VI. Workload Questionnaires and Questionnaire Tools | 24 |
| Figure 1 Questionnaire Cover Sheet | 12 |
| Figure 2 Example - Histograms | 20 |
| Figure 3 Example - Box Plot | 21 |
| Table 1 Common Questionnaire Types | 6 |
| Table 2 Recommended Descriptors Sets for OT&E | 8 |
| Table 3 Example - Frequency of Responses | 18 |
| Table 4 Example - Distribution of Responses | 19 |
| Example Questionnaires | |
| Appendix A: Sample Human Factors Questionnaires | |
| Appendix B: Generic Logistics Supportability Surveys | |

I. INTRODUCTION

A questionnaire is a structured set of questions used to obtain subjective information from a particular group of people. A great deal of operational test and evaluation (OT&E) data is routinely gathered through the use of questionnaires. Questionnaires can be used to quantify difficult to measure aspects of system and operator performance with economy and a high degree of precision. Unfortunately however, when questionnaires are improperly employed, poorly written, or misused, they can produce inaccurate, misleading, or useless information. As such, devising a good questionnaire and knowing when to use it requires careful attention to the principles of questionnaire design. The purpose of this handbook is to provide the reader with the knowledge to determine when to use questionnaires and the tools necessary to effectively develop, administer, and analyze questionnaires during operational tests.

A. Background

Questionnaire usage during operational test has run the gamut between non-use, overuse and misuse. At one time, questionnaires employed during OT&E were thrown together haphazardly and used to gather information for what were perceived to be low-priority or less important portions of an operational test. It was believed that questionnaire data could not stand alone as an effective measure and consequently became almost an afterthought during the test planning process. Sometime later however, when the test community began to feel the squeeze of the budget crunch, questionnaires became the measure of choice as testers began to realize that they were an inexpensive and easy means to collect a myriad of data during a test. Unfortunately, this also led to questionnaires being utilized in lieu of more appropriate objective data measures. The OT&E community has come to realize that effectively written and appropriately employed questionnaires can be used to augment objective measures and provide substantive information when objective data is unavailable or difficult to collect.

B. Overview

The methods used to collect questionnaire data are extremely important for assuring the usefulness of the data they produce. The following sections present fundamental issues in questionnaire development and use. Section II describes some of the relevant issues for OT&E and provides recommendations for when and what kinds of questionnaires to use. Section III provides guidance on how to construct questionnaires. Section IV presents guidelines on administration of questionnaires. Section V describes frequently used methods of analyzing questionnaire data. Section VI outlines some of the advanced techniques available for questionnaire development. Section VII identifies some special-purpose questionnaires and questionnaire tools. Section VIII contains a variety of example questionnaires for effectiveness issues. Finally, Section IX contains logistics suitability questionnaires and basic supportability surveys.

II. GENERAL GUIDELINES

A. Definition Of Questionnaire

A questionnaire is an organized set of questions that have been tailored to obtain information about a particular subject. A questionnaire is one approach to collecting subjective data from a target group of people. Subjective information relies on the judgement of the respondent. As such, questionnaires can only provide subjective opinions about a system's performance.

B. When To Use Questionnaires

1. The most common mistake an analyst can make is to use a questionnaire in situations where objective data are available and better suited to answer the issues at hand. To alleviate this common pitfall, decisions about what test data will be collected by questionnaires should be considered early in test concept development. A number of important factors should be considered before deciding to use questionnaires. First and foremost, you must determine what you want to say about the system in the test report. If you want to report subjective opinions about the system's performance which may affect overall effectiveness or suitability, then questionnaires can provide good quantitative data on who and how many people have what kind of opinions about the system. **A questionnaire will not, however, provide direct, objective data on how the system performed during the test.**

2. A second related consideration is the overall importance of the test issue to system evaluation. For example, does the issue pertain directly to the critical operational issues (COI) or task level measures of effectiveness (MOE's)? If the test issue is key to the evaluation of the system then objective measures are usually preferable, with supporting information gathered from questionnaires. Objective data on system performance are more easily interpreted and defended.

3. In summary, objective measures should always be the primary source of system information supported by subjective data collected using questionnaires. If objective data is unavailable or impractical, then questionnaires can be used to provide a subjective assessment of system performance.

C. Questionnaires As Test Criteria

We do not set rigid criteria for test measures that rely on subjective data. The reason for this is we cannot validate that the operational requirements for a system must be for example, that 80% of questionnaire respondents rate a test measure as very good versus adequate. Test measures which rely on subjective data have their evaluation criteria stated as "none; results will be reported in narrative fashion." This narrative will normally consist of descriptive statistics of the questionnaire data (median, frequency distribution, histogram, etc.) supplemented by other applicable supporting data, such as DT&E data, test team observations, test subject comments, etc.

D. Wording The Test Plan

1. If a decision is made to use a questionnaire to collect test data, care should be taken in wording the (MOE) or measure of performance (MOP). A clear concise statement is desirable, such as; "Pilot ratings of situation awareness." Examples of poorly worded measures are: a) "Adequacy of aircraft maintainability" and b) "The average adequacy rating of maintenance tasks based on questionnaires developed by the test team and administered to all level 5s and above." The thing being measured is either ambiguous as in example (a) or buried under layers of methodological detail as in example (b). Sometimes operational requirements documents (ORD's) are written in such a way that it can be difficult to envision an objective method of measuring the system parameter directly. For example, a requirement that a system must "provide effective training" does not readily lend itself to constructing an objective test measure. TSH analysts will assist test team personnel in determining the appropriate use of questionnaires.

III. CONSTRUCTION GUIDELINES

A. Creating an OT&E Questionnaire.

The questionnaire development process should begin early in test planning. The first step in the process of creating questionnaires is to make a list of all the test areas, MOE's, and MOP's for which you may use questionnaires, and think through the way you would analyze the data to support test reporting. This will enable you to select the right type of questionnaire to provide you with the data you need for the test report. There are a number of questionnaire types suitable for OT&E, they are described in the sections to follow.

1. Types of Questionnaires.

a. Questionnaire data can be gathered in a variety of ways. These include interviews and free-form responses, open-ended questions, multiple choice or multiple option questions, and rating or matrix scales. Table 1 lists these common questionnaire types along with brief descriptions of their pros and cons. The rating scale has the most utility for collecting data relevant to OT&E because it produces easily quantifiable data that can be readily integrated with other sources of data. In some situations, subjective data can also be collected by methods other than written questionnaires. For a discussion of some of these alternatives see section IV.B.

2. Response Scales.

a. An often overlooked, but very important aspect of questionnaire design is the selection of the response scale. While the wording of the questions is what prompts the respondent's answer, the response scale determines the form of the answer. The response scale defines the distribution of responses by providing the number and type of allowable answers to a question. For operational test purposes, balanced, bipolar, 5-7 point scales are preferred. A bipolar scale has both negative and positive alternatives. The scale is called balanced when there is an equal number of positive and negative alternatives. Historically, researchers prefer balanced scales because they tend to produce distributions that are more nearly normal. Unbalanced scales are typically used only when

there is reason to suspect that a tendency to select extreme responses will produce an uneven distribution.

b. The number of response alternatives is often determined on the basis of the degree of discrimination required. Sometimes greater discriminability can be obtained by more response alternatives, although there is no assurance of this. However, an increase in the number of response alternatives also tends to increase the questionnaire administration time. Perhaps the best basis for selecting the number of alternatives is to consider how easy each response is to differentiate from the others. Research shows that clear discriminability can be obtained with up to seven alternatives. More than seven alternatives increases the response variability and lowers the overall reliability of the questionnaire.

Table 1: Common Questionnaire Types

| | | | | | | | | | | | | | | | | |
|---|--|--------------------------|--------------------------|--------------------------|------|------------|------------|------------|----------|----------|----------|--------------------------|--------------------------|--------------------------|--------------------------|--|
| <p style="text-align: center;">Multiple Choice</p> <div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;"> <p>What is your duty position?</p> <p><input type="checkbox"/> Ground System Operator</p> <p><input checked="" type="checkbox"/> Satellite Operations Officer</p> <p><input type="checkbox"/> Crew Commander</p> </div> <p>Pros -Answers are easy to summarize and may be very reliable</p> <p>Cons -Cannot ask complex questions -Questions may force respondent to make a choice that is not wholly consistent with their thinking</p> | <p style="text-align: center;">Multiple Option</p> <div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;"> <p>Which sandwich toppings do you prefer? (select all that apply)</p> <p><input type="checkbox"/> Pickles</p> <p><input checked="" type="checkbox"/> Lettuce</p> <p><input checked="" type="checkbox"/> Onions</p> <p><input type="checkbox"/> Cheese</p> </div> <p>Pros -Often used for areas of concern within a rating scale question -Provides examples to prompt respondents</p> | | | | | | | | | | | | | | | |
| <p style="text-align: center;">Fill in the Blank</p> <div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;"> <p>My name is _____</p> <p>Unit _____</p> <p>Date _____</p> </div> <p>Pros -Useful for collecting biographical data</p> <p>Cons -Hard to score</p> | <p style="text-align: center;">Open Ended Question</p> <div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;"> <p>What is your overall opinion of the CETO system?</p> </div> <p>Pros -Respondent can raise issues not previously addressed in questionnaire</p> <p>Cons -Hard to score -Responses are open to misinterpretation by analyst</p> | | | | | | | | | | | | | | | |
| <p style="text-align: center;">Rating Scale</p> <div style="border: 1px solid black; padding: 5px; margin-bottom: 10px;"> <p>Rate the adequacy of the CETO display:</p> <table style="width: 100%; border-collapse: collapse;"> <tr> <td style="text-align: center;">Totally</td> <td style="text-align: center;">Very</td> <td style="text-align: center;">Barely</td> <td style="text-align: center;">Barely</td> <td style="text-align: center;">Very</td> </tr> <tr> <td style="text-align: center;">Inadequate</td> <td style="text-align: center;">inadequate</td> <td style="text-align: center;">inadequate</td> <td style="text-align: center;">adequate</td> <td style="text-align: center;">adequate</td> </tr> <tr> <td style="text-align: center;">adequate</td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> <td style="text-align: center;"><input type="checkbox"/></td> </tr> </table> </div> <p>Pros -Easily answered -Easily applied to most items -Easily scored</p> <p>Cons -Vulnerable to bias built into stem - Can be confusing</p> | Totally | Very | Barely | Barely | Very | Inadequate | inadequate | inadequate | adequate | adequate | adequate | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | |
| Totally | Very | Barely | Barely | Very | | | | | | | | | | | | |
| Inadequate | inadequate | inadequate | adequate | adequate | | | | | | | | | | | | |
| adequate | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | | | | | | | | | | | | |

c. The most contentious aspect of choosing a response scale is whether to use a neutral midpoint. The presence or absence of a neutral midpoint does not inherently affect a scale's balance, however, it may affect the response distribution. Denying a neutral midpoint tends to increase the variability about the theoretical center and thus reduces the discriminability near the center. In addition, some respondents resent being forced to select a choice that departs from true neutrality. This occasionally results in the omission of responses to some questions. On the other hand, there may be reason to believe that respondents will be unwilling to provide anything but a neutral response. In this case the midpoint can be dropped from the response scale. The consequences of forcing the respondents to make a choice must be carefully weighed against the potential benefit of obtaining non-neutral responses.

3. Descriptor Sets.

a. Response alternatives that accompany response scales, often called descriptors or semantic anchors, are critical considerations in scale construction. Descriptors must be chosen for consistency, discriminability, and comprehensibility in order to be effective in avoiding response bias. Several recommended sets of response alternatives are presented below in Table 2. These examples ensure that the phrases in each set have means at least one standard deviation away from each other, parallel wording, and extreme endpoints. Response alternatives should be ordered from "low" to "high". In general, unless there is some reason to believe that the order of response alternatives may make a difference in the response selected, the best practice is to retain the same directional order for all items in a questionnaire.

b. Descriptive anchors should be selected to be consistent with or match the wording of the MOE, user requirement/test criterion, or objective. If the MOE asks for ratings of effectiveness, anchors such as "effective," "ineffective," "very effective," etc., should be used. Select a descriptor set from Table 2 that best matches the wording and intent of the MOE. If you are working with a MOE that does not match any of the example descriptor sets, you may consider changing the wording of the MOE.

Table 2 Recommended Descriptor Sets for OT&E

| | | | | | |
|-------------------------|------------------------|-----------------------|----------------------|-----------------------|-----------------------|
| Totally Inadequate | Somewhat Inadequate | Borderline | Somewhat Adequate | Totally Adequate | |
| Completely Unacceptable | Somewhat Unacceptable | Borderline | Somewhat Acceptable | Completely Acceptable | |
| Completely Ineffective | Somewhat Ineffective | Borderline | Somewhat Effective | Completely Effective | |
| Extremely Difficult | Somewhat Difficult | Borderline | Somewhat Easy | Extremely Easy | |
| Completely Disagree | Substantially Disagree | Borderline | Substantially Agree | Completely Agree | |
| Extremely Unimportant | Moderately Unimportant | Borderline | Moderately Important | Extremely Important | |
| Completely Useless | Somewhat Useless | Borderline | Somewhat Useful | Completely Useful | |
| Undoubtedly Worse | Moderately Worse | The Same | Moderately Better | Undoubtedly Better | |
| Never | Rarely | Now and Then | Often | Always | |
| Totally Inadequate | Very Inadequate | Somewhat Inadequate | Somewhat Adequate | Very Adequate | Totally Adequate |
| Completely Unacceptable | Largely Unacceptable | Somewhat Unacceptable | Somewhat Acceptable | Largely Acceptable | Completely Acceptable |
| Completely Ineffective | Largely Ineffective | Somewhat Ineffective | Somewhat Effective | Largely Effective | Completely Effective |
| Extremely Difficult | Moderately Difficult | Somewhat Difficult | Somewhat Easy | Moderately Easy | Extremely Easy |
| Completely Disagree | Substantially Disagree | Slightly Disagree | Slightly Agree | Substantially Agree | Completely Agree |
| Extremely Unimportant | Moderately Unimportant | Barely Unimportant | Barely Important | Moderately Important | Extremely Important |
| Completely Useless | Largely Useless | Somewhat Useless | Somewhat Useful | Largely Useful | Completely Useful |
| Undoubtedly Worse | Moderately Worse | Slightly Worse | Slightly Better | Moderately Better | Undoubtedly Better |
| Never | Very Rarely | Somewhat Rarely | Somewhat Often | Very Often | Always |

| | | | | | | |
|-------------------------|------------------------|-----------------------|------------|---------------------|----------------------|-----------------------|
| Totally Inadequate | Very Inadequate | Somewhat Inadequate | Borderline | Somewhat Adequate | Very Adequate | Totally Adequate |
| Completely Unacceptable | Largely Unacceptable | Somewhat Unacceptable | Borderline | Somewhat Acceptable | Largely Acceptable | Completely Acceptable |
| Completely Ineffective | Largely Ineffective | Somewhat Ineffective | Borderline | Somewhat Effective | Largely Effective | Completely Effective |
| Extremely Difficult | Moderately Difficult | Somewhat Difficult | Borderline | Somewhat Easy | Moderately Easy | Extremely Easy |
| Completely Disagree | Substantially Disagree | Slightly Disagree | Borderline | Slightly Agree | Substantially Agree | Completely Agree |
| Extremely Unimportant | Moderately Unimportant | Barely Unimportant | Borderline | Barely Important | Moderately Important | Extremely Important |
| Completely Useless | Largely Useless | Somewhat Useless | Borderline | Somewhat Useful | Largely Useful | Completely Useful |
| Undoubtedly Worse | Moderately Worse | Slightly Worse | The Same | Slightly Better | Moderately Better | Undoubtedly Better |
| Never | Very Rarely | Somewhat Rarely | Borderline | Somewhat Often | Very Often | Always |

4. Question Wording. There are no hard and fast rules for stringing words together into effective questions. As mentioned previously, there are a number of example questionnaires at the back of this manual and you are encouraged to make use of these where possible. Some rules of thumb to follow in wording your questions are as follows:

a. Avoid Difficult Vocabulary. It is important to speak to the level of the individuals who will be answering the questionnaire. Avoid using jargon, acronyms, or overly technical terms that may be misunderstood by the respondents.

b. Avoid Negatives. Use neutral phrases whenever possible. The use of "not", "no", "un" or other negatives should be avoided. Negative questions such as: "Rate the degree to which the system possesses no voids or gaps" may not only bias the respondent but may also be misread or misunderstood. The word "no" can be eliminated from the above example to make the question more easily understood. Finally, double negatives should never be used.

c. Avoid Positives. Use neutral phrases whenever possible. Positive questions such as: "Do you agree that the BPOS is an adequate system?" may bias the respondent.

d. Avoid Double-Barreled Questions. Double-barreled questions pose two questions simultaneously, such as: "Rate the responsiveness and reliability of the system." If the responsiveness is

good but the reliability is poor, the respondent will have great difficulty answering the question. In the above example, two separate questions should be asked, one for responsiveness and one for reliability.

e. Avoid Leading/Loaded Questions. Leading questions presuppose some event or state. Questions like: "Rate the lack of responsiveness of the system" presume that the system is unresponsive. Loaded questions, like leading questions, presume some state but also carry with them a charged emotional content as in the following example: "Rate your lack of ability with respect to the duties you carried out today." Leading and loaded questions may produce biased data for that question and should be avoided.

f. Avoid Emotionality. Related to the issue of loaded questions described above, questions containing emotional or sensitive words have the potential for invalidating the data for the entire questionnaire. Questions perceived as self-incriminating, emotional, or sensitive frequently involve the personal qualities, capabilities, and knowledge of the person completing the questionnaire. Since it is unreasonable to expect people to objectively evaluate their own performance, questionnaire items should be directed to the adequacy of the system, rather than the user.

g. Be Brief. Keep your questions short. A single sentence is best. Fewer words = better questions. The more words it takes to ask a question, the more complicated it is to understand and the greater the opportunity for misunderstanding.

h. Focus on Relevant Topics. Only ask what is necessary. Once embarked on a questionnaire creation effort, it is all too easy to add more and more "nice to know" sorts of questions. If the questions are irrelevant, the result is an unnecessary burden on the respondents and the data analysts.

5. Questionnaire Assembly. While individual questions may have balanced response scales, good descriptor sets, and appropriate wording, they still must be assembled as a complete package: the questionnaire. Good questionnaires will include a cover sheet and instructions, appropriate question sequence, brevity, and an appropriate questionnaire medium. All of these areas require attention in order to ensure the quality of the questionnaire data. Each area is described below.

a. Cover Sheet & Instructions. The questionnaire cover sheet is a key ingredient in obtaining the respondents' cooperation. A good cover sheet should include the title of the questionnaire, the purpose (including any information regarding the use of the data and assurance of confidentiality (if needed)), instructions, an example question, and space for demographic information. It may also contain methods of tracking the data (space for operator ID, position, location, and/or date/time group). An example cover sheet for an OT&E questionnaire is presented in Figure 2.

b. Question Sequencing. The questions should flow in a logical order as much as possible. Questions may be organized from the most general topics to the most specific, from the most specific to the most general, from the most frequent/common events to the rare or unusual, or grouped by particular areas of interest.

c. Questionnaire Length. Questionnaires should be as brief and to the point as possible.

One of the dangers of overly long questionnaires is the tendency for respondents to rush through the questions and not give appropriate attention to what is being asked. Carelessness is a frequent source of error in lengthy questionnaires. These biases may be detected by scrutinizing the performance or response patterns of individual subjects. Carelessness may be detected in unusually short amounts of time taken to complete the questionnaire, as well as a tendency to give answers that consistently deviate from the norm. A simple rule of thumb: 20 questions may take anywhere from 15 to 40 minutes for a respondent to complete.

d. Questionnaire Medium.

(1) Paper Questionnaires. A popular medium for questionnaires is paper and pencil. The paper and pencil method is quite flexible, simple to prepare, and easily administered. However, there are important considerations in its use. First and foremost, consider whether adequate time and manpower will be available to manually code and enter the data into a database or statistical analysis package. Manual data entry required for paper questionnaires may introduce a risk of data translation errors during analysis. Another consideration in using paper questionnaires is the difficulty in transporting large stacks of paper to and from the test site. Additionally, if you decide to use paper questionnaires be sure to leave adequate room to respond, particularly in the "comments" area. The amount of space you provide will determine the amount of detail you receive in the answers. Conversely, don't leave an excessive amount of space since some subjects feel compelled to fill all the available space. Readability of the questionnaire is another important consideration, particularly in situations where illumination levels are low and distractions are high (cockpit or nighttime situations). Consider also the requirements for the size and stock of paper where handling considerations and field use make compact or semi-rigid forms necessary.

(2) Computer-based Questionnaires. Computer-based questionnaires can be an attractive alternative to paper and pencil questionnaires for a variety of reasons. The major advantage to questionnaires presented on a laptop or other computer is that the data will not require later key entry by a data clerk or analyst. For complex questionnaire formats, the computer can present the questions in a clear and easy to follow sequence. Laptop questionnaire administration also reduces test team workload and may allow for an on-site "quick look" review of the results. The obvious disadvantage is the hardware and logistics requirements of acquiring and transporting a laptop computer. As computer resources are often limited, using computer-based questionnaires can be difficult with larger groups. See Section VII for details on the computer tools available.

Figure 1: EXAMPLE QUESTIONNAIRE COVER SHEET
JPOS Effectiveness Questionnaire

DIRECTIONS:

1. Your personal responses to this questionnaire are very important in helping to evaluate the performance of XXXX. There are XXquestions for which you will be asked to provide a rating response on a X-point scale.
2. Please select and mark one rating on the scale which best corresponds to your response. If after providing a rating, there are any problem areas you wish to identify, select ALL of the areas that apply by circling the number. If you select "Other/Comments", please go to the lines immediately following the question and write your response (as in the example below). Your comments are encouraged and will be valuable to the success of this questionnaire.
3. If you have any questions, please ask a test team member.
4. Please write down the name of your duty title/position in the space provided below:

THANK YOU FOR YOUR PARTICIPATION

- Rate the acceptability of the

| | | | | | |
|----------------------------|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|
| <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> | <input type="checkbox"/> |
| Completely Unacceptable | Largely Unacceptable | Somewhat Unacceptable | Somewhat Acceptable | Largely Acceptable | Completely Acceptable |

PROBLEM AREAS: 1. Problem Area X

(Select all 2. Problem Area Y
that apply 3. Problem Area Z
and explain)

Other/Comments (space provided below)

Comments: _____

Position/AFSC: _____

Months of experience with this equipment: _____

6. Quality Check and Pretest. The final step in the construction of a questionnaire is the quality check or pretest. The questionnaire must, of course, be reviewed for grammatical and typographical errors, but it must also be reviewed for accuracy and clarity of content. A quality check of the questionnaire is imperative to avoid hours of wasted time and energy collecting invalid and useless data. Further, a poor quality questionnaire will tend to increase resistance on the part of respondents such that you may not be welcomed back for a second visit. Quality checks can be performed either by interviews (table-top reviews) with subject matter experts (SMEs) or by actually administering the questionnaire to a sample of system operators as a pretest. A pretest should involve respondents from the same population as the actual test, and the interviewers/administrators should also be those who will be implementing the questionnaire. As such, it is important to keep in mind that a portion of the test population will be familiarized with the questionnaire prior to test. This could introduce bias if these same respondents are part of the actual test subject pool. It is often possible to conduct detailed question-by-question table-top reviews with system SMEs drawn from the operational squadron who would not form part of the test questionnaire sample. A pretest or quality check interview serves the following purposes:

- a. It may identify unforeseen problems in question wording, such as use of unfamiliar or inappropriate terminology and ambiguously phrased questions. These problems can be corrected prior to test.
- b. It may identify problems in instructions, format, question order, or questionnaire administration. Again, these problems can be corrected before test.
- c. It may indicate the need for additional questions on some topics or the elimination of others.
- d. The length of the questionnaire can be determined, such as the possible necessity to shorten it.
- e. Open-ended responses can be collected to permit the phrasing of closed-ended response alternatives for the final questionnaire.
- f. A pretest questionnaire may constitute part of the administrator's training.

B. Review and Approval Process.

1. Prior to their use, questionnaires must be approved by HQ AFOTEC at the director level. If at all possible questionnaires should be submitted to TS 60 days prior to test start to ensure adequate time for revision and review.

2. TSH Review Method. The following paragraphs present a brief overview of the questionnaire elements examined by TSH in their questionnaire review process. A review of your own questionnaire in accordance with the standards described below prior to submitting them for approval can greatly speed the questionnaire development and approval process.

a. Question-MOE Match. The first feature of the questionnaire to be examined is the correspondence between the questionnaire and the MOE's and MOP's identified in the test plan. Question relevance is as important a consideration as correspondence to the wording and intent of the MOE's/MOP's. Irrelevant questions or questions that do not address the scope of the test objective or MOE/MOP are identified at this point and excluded from subsequent review steps.

b. Question Wording. The next area to be examined is the wording of the individual questions. The questions are examined to ensure that there are no double-barreled, leading/loaded, or poorly worded questions, and that the questions match the descriptors used on the response scale. Spelling, typographical, and grammatical errors are also identified at this time.

c. Questionnaire Format. The final area examined is the overall questionnaire format, the clarity of instructions, adequacy of any comment fields, and the overall length of the questionnaire. This is typically accomplished by attempting to complete the questionnaire from the perspective of the average respondent.

IV. DATA COLLECTION AND ADMINISTRATION GUIDELINES

A. Steps In Administering A Questionnaire .

Typically, test considerations such as the situation, amount of time, and number of subjects available are driven by factors other than questionnaire administration. There are, however, a number of data collection considerations that can be planned for in advance that will greatly increase the probability that OT&E questionnaires will yield useful data.

1. When to Administer a Questionnaire.

a. Mission/Scenario/Task-Based Administration. If questionnaires are being administered in order to collect data for separate and unique scenarios (such as maintenance tasks), each administration should be scheduled to occur when the questionnaire topic is still fresh in the respondent's memory. Questionnaires should be administered during or just after the mission debriefing. In order to ease the strain on the respondents and increase compliance, try to make these task-based questionnaires as short as possible.

b. Total Test-Based Administration. Questionnaires can also be administered in order to collect an overall opinion of the system during the whole test. The worst time to administer a questionnaire is at the end of the day or duty shift. Most respondents are tired, eager to go home, and will spend the minimum amount of time they possibly can to complete a questionnaire. A better time to administer a questionnaire is at the start of the duty day or during a break in the respondents' shift. The best of all possible conditions is to designate a period of time during the shift when the respondents complete the OT&E questionnaires as part of the day's activities such that it doesn't "cost" them any of their off-duty or break time.

2. Instructions to Respondents. Written instructions are an absolute necessity and should be provided on the cover page and throughout the questionnaire as appropriate. An additional technique that frequently increases compliance and the quality of the answers is an oral introduction prior to the first administration of the questionnaire. Although much of the information presented orally will be the same as the information provided on the cover sheet, it allows the respondents to associate the questionnaire with a particular person. This association has two benefits; it prompts respondents to give more thoughtful consideration to their answers and it provides them with a point of contact if they have a concern or don't understand a question.

3. Administration Protocol. Another key to successful use of questionnaires is to be involved and available during the time(s) questionnaires are being filled out. A questionnaire that is simply distributed or left on a desk in the break room will be perceived as unimportant and will be given little effort or even ignored altogether. However, your involvement should not be so overwhelming as to be a nuisance or bias the respondents' answers. The number and type of respondents in the test effort should guide the amount and type of interaction you have with the respondents during questionnaire administration. In cases where multiple questionnaire administrators are used, written instructions and answers to frequently asked questions should be provided to the administrators to ensure uniformity in

how the administrators interact with the respondents.

4. Administration Frequency. One final issue in the subject of questionnaire administration is how often to administer questionnaires to the test participants. A good rule of thumb is to administer the questionnaire only as often as is absolutely necessary. Repeatedly administering an OT&E questionnaire to the same subjects does not increase the sample size for data analysis purposes. Sample size calculations are based on numbers of respondents, not volume of responses. Respondents who are presented with the same questionnaire over and over again will quickly stop putting their time and energy into answering the same questions repeatedly. The data analyst is also confronted with the problem of deciding how to reduce and present the large volume of completed questionnaires. If data are needed for each of five test scenarios, then questionnaires should be prepared for each of the scenarios and administered once during each scenario. General questions on topics that are not scenario dependent should only be administered once or twice. One very successful practice is to administer the questionnaire to each test participant twice: once at the very beginning of OT&E and once at the end. The first administration informs the respondents of the kinds of test issues you are interested in and the second time provides the analyst with good data based on their experiences with the system during the OT&E. Only the data from the end of test are used in the analysis, data from the initial administration are discarded. The initial administration at the start of OT&E is also useful in finding questions that need revision or rewording.

B. Alternative Data Collection Methods.

A questionnaire is not universally suited for all subjective assessment data collection efforts. There are some situations (based on cost, logistics, or phase of test) which may be better served by an alternative data collection methodology. Two common alternatives to questionnaire administration include interviews and archival data methods.

1. Interviews.

a. When the number of respondents is small, or the scope of the assessment area is quite limited, an interview may be able to provide the information needed. Interviews used to collect OT&E data can range from the general question "How'd it fly?" to a more complex and structured approach. A good structured interview is typically arranged hierarchically such that general questions act as filters for more detailed or specific questions. The principles of questionnaire design also pertain to structured interviews. Questions should be relevant, should not be phrased in an ambiguous or leading fashion, and should be as brief as possible. Interviews can be time-consuming and the data collected may be difficult to analyze; yet despite these concerns, interviews often yield valuable data. The primary difference between questionnaire and interview methodologies is the role of the human data collector.

b. The immediate and central involvement of the data collector in interviews has two distinct effects. First, it allows the interview to be more flexible and comprehensive than a written questionnaire. An interviewer can take the time to explore unanticipated events and aspects of testing or system functioning that could not have been foreseen during the preparation of a written

questionnaire.

c. Secondly, data collector involvement may produce an undesirable effect -- the interviewer may bias the responses. Responses may be affected, either consciously or unconsciously, in the way the questions are asked or recorded. Care must be taken to present the questions impartially to all of the respondents in exactly the same way such that the personal biases or opinions of the interviewer do not affect the data. Additionally, some respondents may feel intimidated by the interviewer. This may lead to shorter or less substantive responses. If the above concerns are addressed, interviews can provide a very flexible and cost effective method for OT&E data collection.

2. Archival Data Collection.

Archival data collection refers to the examination of various sources of historical data or physical records. Historical data can take the form of written records (such as crew information files or duty logs) or can even be the physical evidence of some activity such as wear patterns on consoles and controls. There is no need, for example, to ask a respondent how many messages he/she sent if there is a communications log available for you to examine. Similarly, messages or workarounds posted on a bulletin board can provide information about system problems and potential fixes. Archival data sources can often be used in the early stages of test to direct or focus subsequent data collection activities and later to supplement and amplify data collected by other means. All that is required to make use of archival data is the willingness to spend some time looking. The payoff is frequently a lead or indication of a problem area that can direct your energies to the most efficient and significant areas of operational testing.

V. QUESTIONNAIRE ANALYSIS

A. Descriptive Analysis.

1. Scoring questionnaire responses is not difficult. For most types of rating scales, the analyst first assigns numerical values to the descriptors (e.g. 1 through 6 for a six-point scale). This scoring simplifies the analysis and presentation of questionnaire data. It should be noted that such numbers should not appear with the rating scales of the questionnaire to be administered. These numbers are used for data analysis purposes only.

2. There are several ways to analyze and present questionnaire data – measures of central tendency, frequencies, and percentages. However, before selecting a particular technique, there are some factors to consider:

a. Questionnaire data represent an ordinal (or in a very few cases with specialized response scales, interval) measurement scale. Because of this, use of the sample mean or average is inappropriate. Instead, a median or mode statistic should be used to summarize questionnaire ratings.

b. Operational test and evaluation often relies on a small number of participants. The data collected by questionnaires may reflect the opinions of only a few people. In such a case, avoid

the use of percentages (e.g. 75% rated display legibility as “largely acceptable”) to prevent misrepresenting the data. Frequency of response will provide a clearer picture of the findings (e.g. 3 of the 4 operators felt display legibility was “largely acceptable”).

c. Carefully review the data to highlight possible trends or discrepancies before formal analysis. Table 3 shows an example using data from a 6-point scale ranging from “Completely Unacceptable” to Completely Acceptable.” The questionnaire data is organized by the number of responses for each rating. A glance at the table can quickly highlight potential trends. Another approach involves computing percentages of ratings for each question. These percentages provide the basis for comparison and will provide insight into any problems experienced during the test. Table 4 provides an example. The distribution of response percentages in the table readily shows the ratings of workload, and workspace are generally high; displays received a “borderline” rating; and there was wide disagreement on the subject of climate control. With this insight, the analyst can then examine the questionnaires to see if different crew positions, duties, training, or environmental factors (night vs. day shifts) led to the differences in ratings.

Table 3. Example – Frequency of Responses

| Evaluation Area | No. responses <i>Completely Unacceptable</i> 1 | No. responses <i>Largely Unacceptable</i> 2 | No. responses <i>Somewhat Unacceptable</i> 3 | No. responses <i>Somewhat Acceptable</i> 4 | No. responses <i>Largely Acceptable</i> 5 | No. responses <i>Completely Acceptable</i> 6 |
|-----------------|---|--|---|---|--|---|
| Workload | 0 | 0 | 0 | 12 | 3 | 0 |
| Workspace | 0 | 0 | 0 | 3 | 11 | 1 |
| Display | 0 | 1 | 9 | 3 | 2 | 0 |
| Climate Control | 1 | 4 | 2 | 0 | 5 | 3 |

Table 4. Example – Distribution of response percentages

| Evaluation Area | % responses <i>Completely Unacceptable</i> 1 | % responses <i>Largely Unacceptable</i> 2 | % responses <i>Somewhat Unacceptable</i> 3 | % responses <i>Somewhat Acceptable</i> 4 | % responses <i>Largely Acceptable</i> 5 | % responses <i>Completely Acceptable</i> 6 |
|-----------------|---|--|---|---|--|---|
| Workload | 0 | 0 | 0 | 80 | 20 | 0 |
| Workspace | 0 | 0 | 0 | 20 | 73 | 7 |
| Displays | 0 | 7 | 60 | 20 | 13 | 0 |
| Climate Control | 7 | 27 | 13 | 0 | 33 | 20 |

3. Measures of central tendency are very useful in presenting questionnaire data. As mentioned earlier, the measures most appropriate for this application are the median and mode. Do not use the sample mean or average.

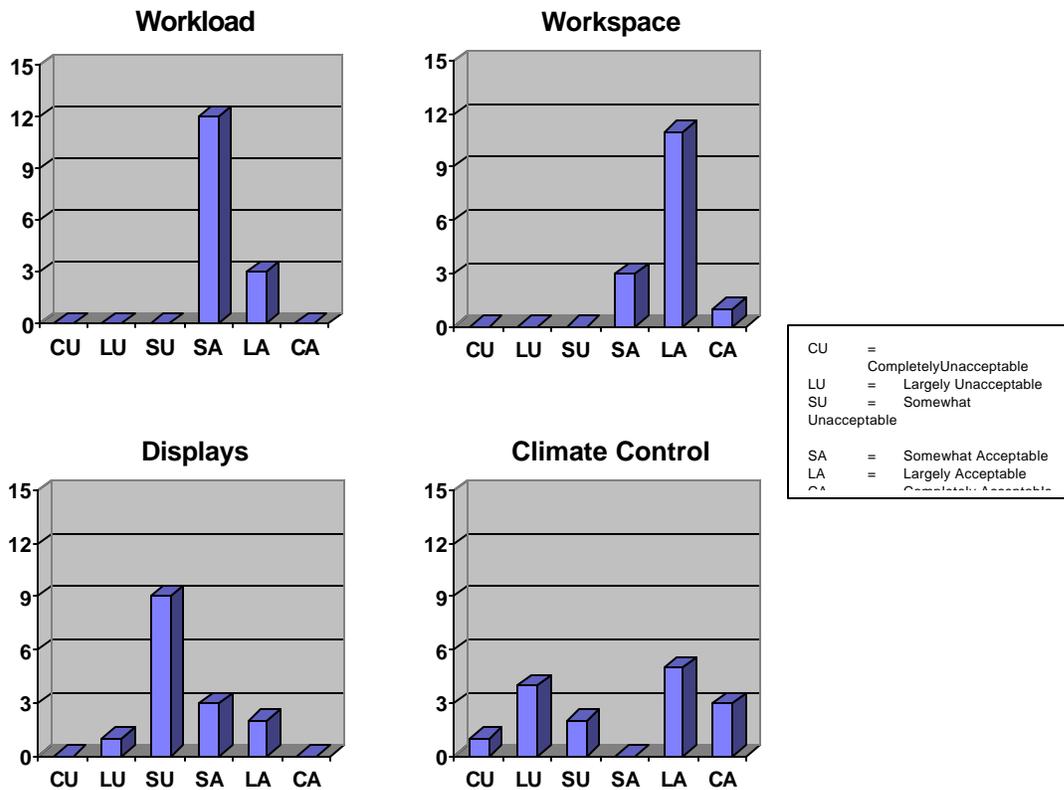
a. The definition of a *median* is the 50th percentile. With the scores arranged highest to lowest, 50% of the responses lie above this midpoint, and 50% lie below. For the data in Table 3, the analyst can easily calculate the median for each area: 5 for and workspace; 4 for workload; 3 for displays. It is possible to calculate a median value for climate control (e.g 5), but it would provide an incomplete picture. The responses represent a special kind of pattern known as a bimodal distribution where the ratings fall into two distinct groups. When using a median to summarize questionnaire data, bimodal distributions need to be explicitly described and investigated. In actual OT&E settings, the two groups or distributions of ratings could be the result of different room temperatures associated with different shifts. The report would need to discuss the ratings of climate control separately for each shift. Most statistics programs will calculate the median for you from raw data, so it is important to always examine the response distributions looking for bimodal distributions or extreme ratings rather than simply rely on the median values reported by a statistics program.

b. The *mode* is defined as the most frequently occurring score. In the data presented in Table 3, the mode for workspace is 5 (largely acceptable), 4 (somewhat acceptable) for workload, and 3 (somewhat unacceptable) for displays. With the ratings for climate control it again becomes necessary to 1) investigate the potential influences creating this condition, and 2) report the results in light of your findings.

4. Figure 1 shows an example of another method for analyzing and presenting test data. The histogram is a graphical representation of the frequency of response for each rating. The principle that “a picture is worth a thousand words” describes the advantage of using histograms to summarize data. They quickly show the overall pattern of responses including bimodal distributions and the amount of agreement overall. In general, for briefings, small numbers of questions, or for detailed reports, histograms present the data very nicely. Where page space is limited or where the number of questions

is quite large, tables or simple narrative descriptions may be required to present questionnaire data.

Figure 1. Example – Histograms for Questionnaire Data

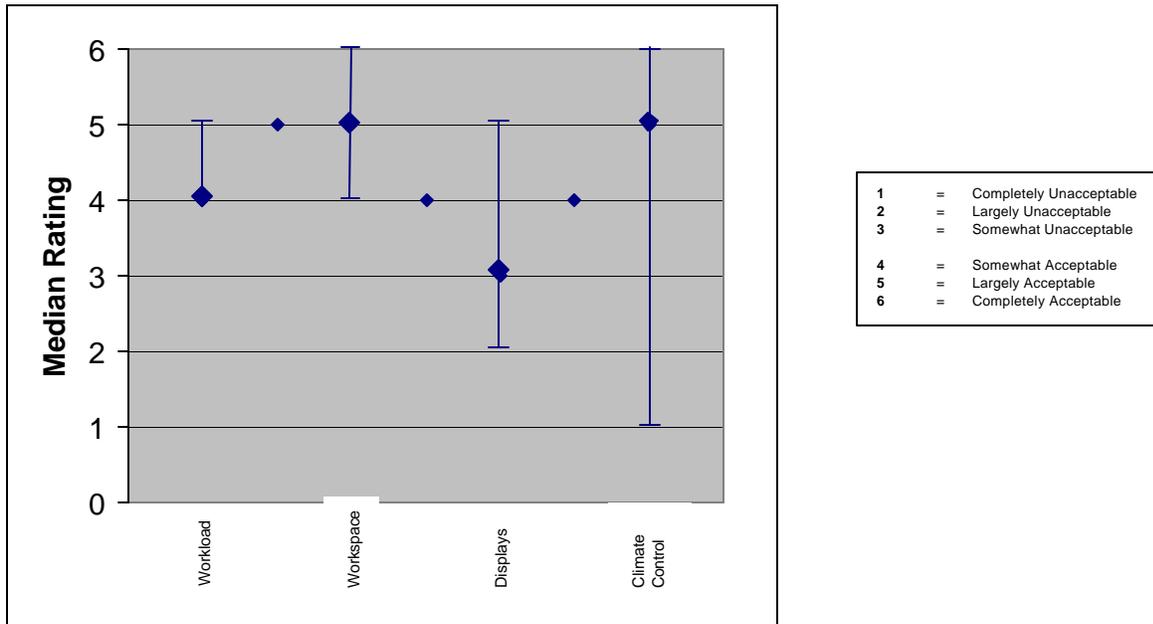


5. When summarizing questionnaire data, using percentages can also help frame the test results. For example, “82% of the maintainers rated TOD usability as ‘Largely Acceptable’ or better.” As mentioned earlier, percentages allow for quick comparison of the ratings when used appropriately. Avoid percentages when the number of respondents is small (e.g. 30 or less).

6. Another way to view these sorts of test data is to describe the amount of agreement in the questionnaire responses by examining the variability of the ratings. This is especially important for results like those obtained for the questions on climate control. One way to capture both the variability and central tendency data in a single format is to use "box plots" as shown in Table 5. The diamonds show the median value and the lines extending on either side indicate the range of responses. When you obtain a wide range of responses to a question, this indicates that the subjects disagreed in their ratings of the system's performance. In these cases the analyst should examine the raw data or a frequency distribution of the data to determine the reason for the disagreement. Sometimes different groups of people or different test conditions can produce differing opinions on a question, as in the case of the bimodal distribution of ratings of climate control described in the example. Once again, where significant disagreement is observed in the results, the analyst may need to report the results separately

for different test scenarios, duty shifts, or personnel categories.

Figure1. Example – Box Plot for Questionnaire Data



7. After you have summarized the data by one of the techniques described above, the next step is to identify the specific problem areas and relate them to the rest of the test data. The specific problem areas can be identified by reading the comments written for the questions (or by examining the problem areas checked on a hierarchical questionnaire). Comments and problem areas identified by the respondents should be categorized and tabulated for those questions receiving negative ratings. Additional information about specific problems can sometimes be collected after the fact by interviewing the respondents 1 or 2 days after questionnaire administration. Delays greater than 2 days will usually result in the respondents being unable to accurately recall the reasons for their negative ratings for a specific question. Once the problems have been identified they should be related back to system performance and user requirements. In the example data presented above, the specific problem associated with noise levels may trace back to excessive noise from air circulation equipment. The high noise levels may interfere with communication between operator. The displays could have problems with poorly organized and formatted information. The operators may have difficulty in locating and reading the displays. The analyst would then use this information to tie both of these areas to their mission impact.

B. Statistical Analysis Considerations.

1. Questionnaire data are different from the type of data that analysts and statisticians usually encounter. For example, questionnaire data do not represent a ratio measurement scale in the same way that distance measures like meters or feet do. Instead, questionnaire data are often ordinal measures, "very acceptable" is better than "effective." Under the best cases, where response scales and

descriptor sets are based on normative data (as in the scales and descriptors recommended in section 2), questionnaire data will approximate an interval scale. Interval scales allow an analyst to state that the difference between adjacent ratings is equal. In other words, we know that the psychological difference between "very ineffective" and "ineffective" is the same as the difference between "ineffective" and "borderline" which is the same as the difference between "borderline" and "effective" and so on. At their worst, questionnaires based on scales and descriptors with unknown properties represent qualitative, or categorical scales; nothing can be said about the order of merit or degree of difference between various ratings, just as "right" is not better than "left" and apples are not better than oranges.

2. With well-developed questionnaires and carefully selected descriptor sets, an analyst can be assured of a relatively greater degree of power in his or her analysis. Questionnaire data, however, still should not be subjected to statistical analyses without careful consideration of the way in which the data are distributed. As is the case with most OT&E data, questionnaire ratings typically do not represent a statistical "normal distribution." Test data, which are not normally distributed, should not be subjected to higher order tests of significance such as analysis of variance. If statistical comparisons between groups, scenarios, or procedures are required for a particular test design, questionnaire data are more appropriately analyzed using non-parametric tests such as the Mann-Whitney, Chi-square, squared ranks, or median tests. In cases where you find bimodal distributions, even non-parametric tests can be distorted by the extreme variability of the data. For these situations, descriptive analyses are better suited to identify the respondent characteristics associated with each cluster of ratings.

C. Validity and Reliability.

Validity and reliability may be achieved when questionnaires are developed in accordance with widely accepted practices in questionnaire construction. These practices and principles include: the wording of items, the length of questions, the length of questionnaires, and the general ability of the respondents to understand the requirements of the test situation. Wording of the "stem" of the question should not favor any of the response alternatives to the question. Moreover, the wording of any one item should not influence the response to any other item. Other questionnaire development practices that apply to the issues of validity and reliability include the "representativeness" of the respondents to the population of system users and the relevance of the questions to the system area under assessment. The individual topics of validity and reliability are described in greater detail below for individuals with a particular interest in these areas.

1. Questionnaire Validity. A questionnaire has validity if the component items measure the variable that was intended, and not some other variable. In general terms, there are two methods of determining a questionnaire's validity. The first and most rigorous method involves repeated administration of the questionnaire accompanied by statistical tests of validity. Unfortunately, in most OT&E situations one rarely has the opportunity to conduct empirical investigations and statistical tests to establish questionnaire validity. The second method of validity is called 'face validity', the extent to which a questionnaire measures its intended subject according to the subjective judgement of the analyst and the readers of the test report. That is, a questionnaire will have good face validity if it appears to clearly and logically quantify the opinions and judgements it is designed to measure. By following the recommendations presented in this handbook, and through careful attention to the organization and coverage of the issues presented in the questionnaire, the face validity of your questionnaire will be

enhanced. The important lesson for the OT&E planner who designs questionnaires is that the validity of questionnaires is directly and inevitably dependent upon the thoroughness and quality of the planning that goes into their creation.

2. Questionnaire Reliability. Reliability refers to the extent to which the same results can be obtained with the same questionnaire when repeatedly applied to the same group of raters. Momentary changes in a respondent's mood, individual differences in question interpretation, and variability in the testing conditions may all decrease the reliability of the response to a question. Reliability is normally evaluated by means of statistical analyses. These analyses consider such issues as (1) the consistency or stability of responses over time, (2) the extent to which the measures are free from sampling error (error resulting from unwanted influences like personal bias or poor wording), (3) the extent to which different approaches to measuring the same thing differ in the responses obtained, and (4) the interpretation of the response by the test team itself. As with validity, the best way to ensure the reliability of a questionnaire in the absence of repeated administrations and statistical testing is to develop the items according to approved practices in questionnaire construction. Controlling for unwanted variability in the responses will improve the reliability of your questionnaires.

VII. WORKLOAD QUESTIONNAIRES & QUESTIONNAIRE TOOLS

A. Workload Questionnaires.

There are a wide variety of questionnaires and tools specifically designed to assess cognitive and physical workload. This section will provide a brief description of several of the most popular tools. For more detailed information contact your TSH analyst assigned to your program or the TSH division chief.

1. Crew Status Survey. The Crew Status Survey (CSS) workload and fatigue questionnaire was developed by the School of Aerospace Medicine and has many years worth of testing and validation data behind it. AFOTEC employs a variant of the CSS (AFOTEC Form TSH20) developed by test engineers at the Air Force Flight Test Center (AFFTC). This variant contains minor revisions to the CSS descriptors that make the scale almost perfectly equal-interval. AFFTC and AFOTEC experience indicate this survey is practical for flight test situations, as well as C4I and other ground-based test situations. The CSS consists of three questions regarding an individual's fatigue, maximum workload, and average workload. The chief advantages of the CSS are its simplicity and ease of administration. Operators can complete the CSS very quickly (usually 30 seconds or less), allowing frequent administration of the survey without adversely affecting system operations. The CSS is typically administered to each operator after each data collection or mission phase to provide a dynamic representation of how operator fatigue and workload change over time. The simplicity of the CSS (as compared to SWAT and other workload indices) also makes for straightforward interpretation and presentation of the results.

2. Modified Cooper-Harper Scale. As the name implies, the Modified Cooper-Harper (MCH) Scale is a derivative of the Cooper-Harper aircraft handling characteristics scale. Investigators developed the MCH to extend the tool's capabilities outside the aircraft domain. The resulting scale provides a sensitive measure of overall workload for a wide variety of operator tasks. Operators typically provide a rating at the end of a test event such as a flight, task sequence, or duty shift. Thus, MCH provides a single value representing the required workload of an entire mission/task or segment. The scale is not meant to identify the specific tasks contributing to the workload, but to identify those operator positions and mission scenarios in which mental workload may be excessive. Once the high workload positions and scenarios are identified, other methods such as SWAT, TLX, etc., are more appropriate to determine specific tasks or functions contributing to the workload. When used appropriately, the MCH provides quality data for many different test situations.

3. Subjective Workload Assessment Technique. The Subjective Workload Assessment Technique (SWAT) was designed by the Armstrong Aeromedical Research Laboratory (AAMRL, now Air Force Research Lab or AFRL) to quantify the workload associated with various events in cockpit and other operator stations. Workload, as defined for SWAT, consists of three dimensions – time load, mental effort load, and stress load. *Time load* refers to the amount of time available for an operator to perform a task. This includes both overall time and the rate at which the person must work to keep up. *Mental effort load* refers to the amount of attentional capacity or effort required without regard to time. This includes such functions as retrieving information from memory, performing

calculations, and decision making. *Stress load* refers to anything that makes the task more difficult by producing anxiety, frustration and/or confusion. This includes factors such as fatigue, vibration, G-loading, and heat. SWAT is widely used because it is generally considered a well-developed, reliable, and valid workload metric. However, this power comes with a cost. SWAT requires a significant amount of preparation and extensive training of subjects prior to use. Also, the data does not lend itself to “quick look” analysis and results. Test planners should carefully weigh these factors before adopting this technique.

4. **NASA Task Load Index.** The NASA Task Load Index (TLX) is a multidimensional rating procedure that provides an overall workload score based on a weighted average of ratings on six subscales: Mental Demands, Physical Demands, Temporal Demands, Own Performance, Effort, and Frustration. The degree to which each of the six factors contribute to the workload of a specific task to be evaluated, from the raters' perspectives, is determined by their responses to pair-wise comparisons among the six factors. Magnitude ratings on each subscale are obtained after each performance of a task or task segment. Ratings of factors deemed most important in creating the workload of a task are given more weight in computing the overall workload score, thereby enhancing the sensitivity of the scale. Like SWAT, the administration and analysis of TLX is considerably more complex than other workload metrics. These factors may limit its usefulness for operational test.

B. Questionnaire Tools

1. Automated Rating Tool. The automated Rating Tool (ART) is a computer-based tool for the generation, administration, analysis, and reporting of rating scale and free-form questionnaires. ART was developed by AFOTEC for the assessment of operational effectiveness and suitability of systems during operational test and evaluation. However, ART is sufficiently flexible to be used for any type of subjective assessment where fixed response alternatives are appropriate.

Analysts can create standard or compressed rating-scale questions through the selection of pre-approved 5, 6 or 7 point rating scales. Free-form and Yes/No questions can also be chosen.

The ART software and any ART questionnaires can be saved to floppy disks for installation on laptop computers or PCs at a test site. Multiple computers can be used to collect data for one questionnaire through the consolidation of data after administration.

For more information of the Automated Rating Tool, analysts contact your assigned TSH analyst or division chief.

2. Computer Usability Evaluator (CUE). CUE is a software package which runs on a Windows based PC and aids in the development, administration, and analysis of computer usability questionnaires. The current CUE questionnaire tool contains a database of 203 questions, from which, only the questions that best apply to the system under test are chosen. Typical CUE questionnaires range from 50-90 questions and require approximately 25-45 minutes to complete. The questions in the database address the five software usability characteristics of descriptiveness, responsiveness,

consistency, simplicity, and error abatement. Each of the questions is cross-referenced with one of these characteristics and the appropriate software design guidelines drawn from MIL-STD-1801, MIL-STD-1472D, and other software usability standards. It is up to the analyst, with possible assistance from subject matter experts, to select relevant questions from the database, provide notes and examples relevant to the system under test, and produce either a paper questionnaire or a file for administering the questionnaire on a laptop computer. The CUE tool also reduces the questionnaire results and produces several standard reports.

More information about the CUE 3.0 software, can be found in the CUE User's Guide. If you need additional information about AFOTEC policy on software usability, sample CUE questionnaires, sample CUE "write-ups" and how each CUE question item is cross referenced to software usability design standards, documents can be obtained from your assigned TSH analyst or division chief.