OFFICE OF THE SECRETARY OF DEFENSE
1700 DEFENSE PENTAGON
WASHINGTON, DC 20301-1700

OPERATIONAL TEST
AND EVALUATION

JUN 23 2014

MEMORANDUM FOR COMMANDING GENERAL, ARMY TEST AND EVALUATION CENTER
DIRECTOR, MARINE CORPS OPERATIONAL TEST AND EVALUATION ACTIVITY
COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE
COMMANDER, AIR FORCE OPERATIONAL TEST AND EVALUATION COMMAND
COMMANDER, JOINT INTEROPERABILITY TEST COMMAND
DIRECTOR, MISSILE DEFENSE AGENCY

SUBJECT: Guidance on the Use and Design of Surveys in Operational Test and Evaluation (OT&E)

Operational tests are designed to collect a variety of quantitative and qualitative data to enable a robust and defensible determination of mission capability. Surveys are a key mechanism to obtain needed data to aid the operational evaluation. Properly designed surveys, which measure the thoughts and opinions of operators and maintainers, are, therefore, essential elements in the evaluation of a system's operational effectiveness and suitability. A substantial body of scientific research exists on survey design, analysis, and administration that we should leverage in OT&E. I have noted in my review of operational test plans, however, that the OTAs are not consistently applying best practices for survey design and use. This memo and attachment outlines my expectations for using surveys and interviews in OT&E. I expect all TEMPs and Test Plans to be written consistent with this guidance.

Surveys should be used to provide quantitative data as well as qualitative information for determining (1) the *usability* of the system for actual operators and maintainers (a component of human system integration assessments), (2) the operators' perceptions of the system's *utility* including their opinions on whether the system aided or hindered mission accomplishment, (3) maintainers' perceptions of the system's *maintainability,* and (4) the effects of system design on *workload.* Surveys are also used to help *diagnose* why certain performance goals were not met (e.g., training, system design). Interviews/focus groups are beneficial for collecting detailed diagnostic information, to help explain trends in survey responses, and for providing specific feedback to system developers. Since the goal of operational testing is typically to characterize system performance across the operational envelope, surveys are also used to examine how user responses might change under the variety of conditions in which a system might be employed (e.g., workload may change as a function of mission type).

In operational testing, survey responses can be used either for *diagnostic* purposes or as a *response variable* in a test design. Response variables are the key quantitative metrics used in the test to assess the system. For example, if the goal of the test is to determine if an upgraded

display resulted in lower operator workload, then the use of a survey in this case indicates that workload is the primary response variable. The corresponding analysis would compare workload levels between new and legacy displays, and might also examine how workload changes as a function of other factors (e.g., mission type or user demographics). In contrast, a diagnostic survey might be constructed to help determine what about the system design influenced the observed performance. Common questions for diagnostic purposes would be to ask about the usability/utility/accessibility of specific system components.

The goal of the survey (diagnostic or primary response variable) will guide the selection of either a pre-designed, academically-established survey or the construction of a tailor-made survey. Custom-made surveys are designed by the test team to measure thoughts specific to the system and goals of the current test. Custom-made surveys are generally diagnostic in nature. In contrast, pre-designed surveys are published surveys that have been academically verified to be accurate and good measures (or response variables) of a specific human-factors attribute, regardless of the system. Most commonly, these surveys measure attributes that have a scientifically agreed upon definition, such as usability and workload. Examples of these types of surveys are discussed in the attached guidance. Academically-established surveys should be used whenever possible and testers should avoid constructing a system-specific survey to measure constructs such as usability or workload, since proven surveys already exist for those attributes. For some systems, it might be appropriate to use both an academically-established survey as a response variable combined with a custom-made survey or interviews for diagnostic purposes. In other cases, the pre-designed survey may alone be sufficient for the system under test, or might not be used at all. OTAs should therefore clearly identify in test plans the purpose of each survey and how it will be used in the analysis of system effectiveness and suitability.

Surveys should not be used to measure system performance or the system's technical or mission capability. Surveys do provide information regarding the degree of challenge operators and maintainers experienced while achieving the observed level of performance. However, one of the most common mistakes I have observed in surveys is the inclusion of questions that ask whether the system's performance was accurate, timely, or precise enough to complete the mission (e.g., rate the accuracy of the system's onboard GPS). Accurate measurement of performance requires knowledge of ground truth for the test, which operators and maintainers typically do not have. Surveys are measures of thoughts that are highly affected by context, whereas requirements and performance are absolute, and are better measured by the test team and test instrumentation. Therefore, well-designed surveys should always be paired with other quantitative metrics of the system's performance, effectiveness, and suitability to provide a holistic evaluation of system effectiveness and suitability.

Similarly, surveys cannot be used as absolute measures of situational awareness. Situational awareness evaluation must incorporate both the accuracy of the source and the user's interpretation of that information. The tester should not ask an operator to rate his situational awareness as an absolute measure, nor should the tester ask whether the system displayed all the threats and threat positions accurately. Instead, targeted tests can be constructed to determine the operator's knowledge of ground truth, and the operator's perceptions should then be compared to the ground truth for the assessment of situational awareness.

It is important to distinguish between data sheets, interviews and focus groups, and surveys. It is sometimes appropriate to use the operator as an extension of the test team to record observable information about the system such as time to complete actions or the number of actions necessary to reboot the system. This type of questionnaire or data collection sheet is not the focus of this memo. Surveys, as discussed above, should be focused on determining the thoughts and experiences of the system after participating in a test. Extremely lengthy operator data sheets should be avoided to allow operators to spend their time and energy on the actual survey questions.

In addition to the selection or construction of surveys, it is important to consider the administration of the surveys, the number and selection of the respondents, and how the data will be analyzed and presented in test reports. The attached guidelines highlight best practices for these important issues as well as key principles to adhere to in the design of the survey questions and responses. My review of test plans and data management and analysis plans will based on these guidelines and I look forward to working with you as we hone our use and analysis of survey data. My point of contact on this matter is Dr. V. Bram Lillard; he can be reached at (703) 697-3655.

J. Michael Gilmore
Director

Attachment:
As stated

cc:
DUSA(TE)
Deputy, DoN T&E Exec
Director, T&E HQ, USAF
Director, DISA

Attachment: Best Practices of Survey Design, Administration & Analysis

In order to obtain accurate information from surveys the analyst should ensure that the survey is well written, ensure that adequate respondents are available, be mindful of the context in which the survey is administered, and determine what method will be used to analyze the survey data. Best practices for each of these are described in the following paragraphs.

1. Writing Surveys that Collect Accurate Data

Custom-made surveys are useful in OT&E because they allow the test team to measure user thoughts specific to the system/goals of the current test. When drafting survey questions, there are five golden rules to follow to prevent error in the collected data. OTAs should employ these guiding principles when writing survey questions:

- ***Neutrality*** in questions asked and administration: The goal of the survey is to obtain the respondent's thoughts without unduly biasing them. Questions should be phrased in an unbiased manner and not lead a respondent towards any particular answer.

  Bad: "Do you agree that the display is improved?"
  Good: "Rate the degree you agree/disagree with the statement: The display is easy to use."

  The word *improved* implies that the test team believes the display is better. Also by asking "do you agree," the question implies that agreement is the desired answer. Conversely, asking individuals to rate agree/disagree does not imply a correct answer.

- ***Knowledge liability***: Surveys should not ask questions the respondents cannot answer due to limitations in their knowledge.

  Bad: "The training prepared me to use all of the functions."

  Good: "I felt as if I needed more training."

  It is not possible for individuals to know if it was the training, the system design, or their own ingenuity that led to success. They may have failed to accomplish the mission, but think they succeeded. They only have knowledge about the tasks they completed in the test; not all possible tasks. For these reasons the first question can lead to inaccurate data. Conversely, the second question provides accurate data to the analyst.

  Similarly, users should not be asked whether they were successful or the degree to which they would rate their mission accomplishment. Not only is there a knowledge liability, but the question is not helpful in assessing the system under test. If a mission-focused question is desired, the tester may elect to ask whether the user found the system contributed to or hindered their ability to accomplish the mission (a question of utility). Such questions should

be paired with other, more specific, questions to enable a more detailed diagnosis of what contributed to their response.[1]

Finally, operators should never be asked to rate the precision or the accuracy of the system's technical capabilities.

Bad: "Rate the accuracy of the tracking information."

Good: Do not use a survey. Examine instrumented test data combined with truth information.

- ***User friendly***: Survey questions should be brief, clear, and not require a lot of thought or interpretation. Additionally, the ordering of the questions should be logical to the respondent.

Bad: "Rate the effectiveness of the filter configuration interface. The filter interface is Completely Effective if (a) filters are easily configurable supporting operational mission, and (b) no problems were observed. The filter interface is Completely Ineffective if (a) filters cannot be properly configured through the interface, (b) no feasible workaround exists and/or (c) the limitations inhibit your ability to accomplish your mission under operational conditions and time constraints."

Good: "The filter configuration interface is easy to use."

The first question is too long. Respondents are unlikely to read the whole question, and different respondents may weigh the various portions of the question differently. Is (a) more important than (b), or is (b) more important than (a)? The second question is much shorter and therefore more likely to be read in its entirety and is subject to less interpretation.

- ***Singularity***: Each question should ask about one and only one topic to avoid ambiguities in the responses.

Bad: "It was easy to create overlays and publish them to COP."

Good: Two questions: (1) "It was easy to create overlays." (2) "It was easy to publish overlays to COP."

[1] For example, a user may reply that the system greatly hindered his ability to accomplish the mission. Additional diagnostic questions could reveal that the documentation or training were insufficient and were the primary causes of his dissatisfaction. Employing an academic survey, in conjunction with these questions, could reveal that the system's usability was poor.

- ***Independence***: The response to each question should not affect the responses to other questions.

  Bad: "Based on your answers above, rate the acceptability of the system"

  Good: "Rate the acceptability of the system."

  The responses to "roll up" questions, like the bad example, may be unreliable, because some respondents will consider the question as written and base the response on previous responses. Other respondents may treat this question as if it is a general question and ignore the caveat. The discrepancy in how respondents read the question makes it challenging to interpret the collected data.

  Bad: "If you answered not adequate, please provide comments."

  Good: Interview the operators or have a general comments question at the end of a section.

  When respondents are asked to provide additional information depending on how they answer the question, they might be motivated to change their response in order to reduce the amount of effort required to complete the survey. These type of "branching" questions should be kept to a minimum. Follow ups through interviews or a via a general comments question at the end of a section of questions provide a means to obtain additional information for interpreting the responses.

Additional considerations that OTAs should keep in mind when constructing analytical surveys are:

- ***Minimize length***: The perceived length of the survey as well as the actual time taken to complete the survey affects the data accuracy. Ask the minimum number of questions needed for the goal of the test. Grouping questions by response type (e.g., multiple choice, response continuum, open ended) and grouping by topic reduces the perceived length of the survey.
- ***Confidentiality***: When respondents believe that their data will be kept confidential, they are more likely to provide their true thoughts. Names and other personally identifiable information should be kept separate from the actual survey.
- ***Response Types:*** Selecting the appropriate response type is as important as the wording of the question. For example, a seemingly biased question can become a neutral question by providing response options that allow for both negative and positive outcomes. Responses to survey questions can be open (e.g., fill in the blank, comment section) or closed. Closed response types include: binary (yes/no), multiple choice, ranking, and response continuums (e.g., Likert-like scales). OTAs should ensure that the responses are appropriate for the question. The selection of the response type for any particular

question must consider the goals of the test, the analyses to be performed, and how the respondent will want to answer the question. Binary (also called dichotomous) responses are typically inappropriate for survey responses. Humans tend to think in terms of a continuum (e.g., somewhat disagree). If the respondent is asked to put the response into only one of two acceptable answers, then the respondent is being asked to draw a line on the continuum in his mind. This line may or may not be where the analyst would have put the line or where other respondents would put the line. Furthermore, just as in performance measurements (e.g., hit/miss versus miss distance), the analysis of binary responses is more limited and less powerful than that of data from a response continuum.

2. Use the Appropriate Survey Based on the Objective

The objective of the test should guide the choice of generating a custom-made survey or a pre-designed academically-established survey. When the goal is to measure usability or workload the following table provides additional information on the most commonly used surveys. This is not a comprehensive list.

Measurement Construct	Survey	Brief Description	Useful for
Workload	NASA TLX	Two-part survey in which respondents rate the experience of 6 workload drivers on a scale of 0-100 and then establish weights through 15 paired comparisons. The weighted average of the ratings provides a workload score.	For any task, the NASA-TLX is a highly sensitive measure of workload. It also provides diagnostic information about workload drivers. It should be administered immediately after completion of a task.
	Cooper Harper – including Modified Cooper Harper & Bedford Workload Scale	Flow-chart for pilots to provide a rating workload on a scale of 0-10, in which a 4 is considered the cut-off for acceptable workload.	The Cooper Harper is designed for pilots. It can be administered mid-flight. While optimized for quick administration, it is not as sensitive to workload changes or determining what is driving the workload level as other surveys.
	Multiple Resources Questionnaire (MRQ)	A 17 item survey based on the multiple resource model of workload. It provides an overall workload score as well as diagnostic information.	Although it provides an overall workload score, the MRQ is primarily used to provide diagnostic information as to how a design can be improved. It is administered after respondents have competed a task no longer than 90 minutes.
Usability	System Usability Scale	A 10 question survey that provides a measure of usability from 0-100. 70 is generally considered the cut-off between an acceptable and an unacceptable system.	Administer immediately after participants have completed at least one task with the system.

3. Ensure Adequate Respondents to Support Evaluation Goals (Sample Sizes)

No matter the goal of the survey, the test team should ensure that the sample of users and maintainers adequately represent the larger population of users. Inherent to any group of people is variability, particularly as it relates to demographic factors, such as experience level and age. These data should be collected when possible to enable analysis of the survey results across these factors. Demographic data from the test participants can be systematically varied (a controlled factor), or an uncontrolled/recordable factor. Even when uncontrolled, a sufficient number of users should be selected to span the operational user population. If not, the survey results could suffer from unintended selection bias, making the survey results unrepresentative and unreliable. The number of respondents required depends on the size and diversity of the population. Typically larger populations are more diverse and need a larger sample size to adequately represent the diversity.

When the survey will be used as a primary response variable, a statistical power analysis *is needed* in order to appropriately size the test and ensure the data will adequately characterize survey responses across the relevant conditions. For the example of comparing workload between a baseline and upgraded display, the test team should use power calculations to ensure that the number of different users in the test is sufficient to show a difference in workload if a difference truly exists for the factors being examined (e.g., experience level, mission type, legacy vs. new). For diagnostic surveys, there is typically *no need* to calculate statistical power; reporting the confidence interval (if possible) after the fact is typically sufficient. In these cases, we do not need to drive the size of the test (e.g., number mission replications) based on power calculations for survey responses; instead, an adequate number of users should be surveyed to ensure the variety of operators are captured.

4. Administer Surveys in a Timely and Appropriate Fashion

Surveys measure the thoughts and opinions of human beings, which can be altered by the context of the survey: how it is administered, how the questions are phrased, and the environment that the survey is delivered. OTAs should therefore ensure that the survey is administered in a neutral fashion. To obtain the most accurate and useful data, the survey should consider the respondents' motivation for completing the survey accurately.

Surveys can be administered at the end of the test (i.e., post-test survey), at natural break points (e.g., after a mission, at the end of a day), or in response to critical or uncommon events (e.g., system crashes, safety issues). OTAs should consider the goals of the survey when determining when it should be administered. Post-test surveys, because they are administered only once, can be longer than the other types of surveys. The questions in a post-test survey should be about thoughts that will not change based on controlled task factors or time (e.g., overall satisfaction). Alternatively, workload surveys are especially important to conduct as close to the end of a task as possible, as feelings of high workload fade over time.

Surveys that are administered at natural break points during a mission are utilized when there is a desire to compare responses after different missions or over time. Examples include comparing

usability or workload of different systems, evaluating learning curves, or collecting diagnostic information specific to a mission. As these surveys are administered multiple times, the number of questions should be kept to a minimum for the goals of the test. Otherwise survey fatigue may occur and the validity of the responses is likely to deteriorate over the course of the test.

Finally, event-driven surveys should be the shortest (i.e., one or 2 questions). Event-driven surveys can interrupt the flow of the test if they are too long. Furthermore, respondents may not notify the test team of an event, if they are asked many questions.

Regardless of the type of survey, the administrator should ensure that the thoughts of the respondent, and only the respondent, are collected. This is accomplished by emphasizing that the respondent's thoughts of the system are important to the evaluation of the system. The administrator should not share his or her opinion of the system. The administrator should not provide strategies for completing the survey. If the administrator is asked for clarification, he or she should only define words and not provide guidance on the appropriate answer.

Focus groups and interviews are often beneficial, but must be carefully administered, and adhere to the same best practices as written surveys. It should be noted that focus groups have some limitations that can be exacerbated by the context of the interview. For example, if one operator had a differing opinion than his peer or his supervisor, but is only allowed to express his opinion in the context of a group interview, he may not be as willing to provide his full opinion (or may keep quiet). Similarly, some respondents might feel uncomfortable being vocal about system problems if the group is videotaped (loss of confidentiality or nervousness), but would provide valuable information in a one-on-one interview or paper survey. Focus groups with a large number of observers, or specific audience members (such as VIPs), have the potential to skew the responses as well. The value of focus groups/group interviews lies in their ability to encourage discussion about specific system problems; they are by definition, diagnostic in nature. Testers should ensure that conducting the focus group is in concert with the goal of the test and specific goal of the survey, and in general, should pair these interviews with individual paper surveys to obtain both open and closed responses and ensure the most accurate data are collected across the spectrum of users.

5. Analyze Survey Data Appropriately

As discussed above, the responses to all surveys (answers to the questions) fall into two categories: closed responses and open responses. Closed responses are limited to a finite set determined by the survey designer (e.g., multiple choice, rating scale: 0-7). Closed responses, provided they are constructed correctly, benefit from the ability to apply inferential statistical analysis, such as analysis of variance or statistical regression methods. Open responses do not lend well to such quantitative analysis; instead, open responses enable collection of useful information for diagnosing problems. Where possible, prose and interview (open) responses should be categorized and summarized to capture common themes and enable useful feedback to the system developer.

The type of closed response (i.e., binary, multiple choice, ranking, response continuum), determines the appropriate statistical analysis. Binary responses (e.g., yes/no, agree/disagree) are

the most limited in terms of statistical analysis; continuous responses support the widest range of statistical analyses. Empirical surveys can be treated as continuous measures. Response-continuum responses (e.g., Likert-like scales) can typically be treated as ordinal; treating them as continuous will achieve similar results. At a minimum, closed responses should be summarized using descriptive statistics (e.g., mean, median, mode, range) and histograms of response outcomes. Confidence intervals should be provided for closed responses using the appropriate statistical methodology. Ideally, more extensive statistical analyses should be conducted to obtain the most information from the data. In particular, factors that might affect the survey responses, whether controlled or uncontrolled, should be examined in the analysis. Analyses that examine the effects of factors, such as user demographics or mission conditions, provide the most valuable insight into system capability and have the potential to greatly enhance the rigor of the operational evaluation. Table 1 below summarizes potential statistical analyses based on the response type and factors.

			Factors		
			None (One-sample analysis)	Two Groups (One factor, two levels)	Multiple Factors
Measures	Categorical	Nominal	Percents Chi Square Test Fisher Exact Test	Contingency Table Analysis	Contingency Table Analysis
		Ordinal (Yes/No)	Binomial Test of One Proportion	Test of Two Proportions	Logistic Regression
		Ordinal	Percents Chi-Squared Test	Sign test K-S test Correlation	Ordinal Regression
	Continuous	Interval/ Ratio	Mean, Variance T-test	Means, Variances Paired t-test Correlation tests	ANOVA Regression General Linear Models