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

OPERATIONAL TEST
AND EVALUATION

SEP 25 2019

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

SUBJECT: Guidance for Testing and Evaluating Human-System Interaction

The purpose of this memorandum is to consolidate and update the existing survey guidance, and expand it to provide a broader view of the tools available to evaluate Human-System Interaction (HSI). [1] It identifies key HSI concepts testers should measure during operational test and appropriate methods for measuring these concepts and integrating them into test plans. This memorandum supersedes the existing guidance documents on surveys dated June 23, 2014, February 24, 2015, April 2, 2015, and January 6, 2017.

Background

The primary goal of operational testing in the DoD is to evaluate the operational performance (i.e., effectiveness, suitability, and survivability) of systems under realistic conditions when employed by typical military users. The quality of HSI directly affects operational performance and is a critical part of operational testing. Over the past five years, DOT&E has provided the test community with methods for designing and administering surveys – a common tool for evaluating HSI. I applaud the prompt implementation of these policies and their integration into the community's detailed test plans. Our efforts to improve in this area clearly align with Section 227 of the FY 2019 NDAA directing the DoD to develop and provide for a variety of human factors activities.

Evaluating the Effect of HSI on Mission Accomplishment

An adequate operational test enables a fully-integrated, mission-level evaluation of how system design affects an operator's ability to accomplish his or her mission. The test must allow

[1] In this memorandum, the term HSI refers to human-system interaction and should not be confused with another common term within the acquisition community, human-system integration. Human-system integration is a process for designing and developing systems that effectively and affordably integrate human and machine capabilities and limitations. When this process is successful, effective human-system interaction results. For more information see https://www.acq.osd.mil/se/initiatives/init_hsi.html

me to evaluate the degree to which HSI affects operational performance, and the methods used should maximize test efficiency and minimize the risk of error.

HSI may degrade operational performance during a mission because the system is hard to use, the tasks are too difficult, or operators were insufficiently trained. Therefore, an adequate evaluation of HSI should include measures of (1) system usability, (2) workload during mission critical tasks, and (3) training. Some systems will require that testers consider additional HSI concepts like situation awareness or system trust. Whatever the measurement set chosen, testers should focus on measurement at the mission-level and avoid doing so at the component-level, to minimize the burden on the operators.

Quantitative methods include objective measures of human behavior and surveys of an operator's subjective[2] experience. These measures yield numeric data (e.g., time, counts, ratings) which are easily introduced into statistical models to quantify the degree to which HSI affects operational performance. Quantitative methods are described in greater detail in Attachment 1.

Qualitative methods serve an important but distinct role from quantitative methods and the two methods are not interchangeable. Qualitative methods include interviews, focus groups, comment boxes, and in many cases, test team observations. These methods are best used to explore and diagnose the causes of findings and may lead to the discovery of unanticipated problems. Qualitative methods are described in greater detail in Attachment 2.

Table 1. General Guidance for When to Apply Common HSI Measurement Methods

	<u>Quantitative</u>			<u>Qualitative</u>			
Test Team Goals	Behavior	Validated Survey	Custom Survey	Interview	Focus Group	Comment Boxes	Testers' Qualitative Observation
1. Measure operational performance	X						
2. Quantify effect of HSI on operational performance	X	X	X				
3. Quantify recognized HSI concept		X					
4. Describe operator experiences				X	X	X	
5. Problem discovery or diagnosis				X	X	X	X
6. Manage resource constraints		X			X	X	
7. Get high quality data	X	X		X			
8. Reduce operator burden	X	X		X	X		X

[2] Testers should not conflate subjective/objective and qualitative/quantitative dimensions. These can be mixed together in different combinations. In particular, "I rate this seven out of ten" is an objective measure of a subjective experience. Scales are the method to quantify subjective experiences. See Attachment 1.

Testers will gain the greatest insights into the quality of HSI by combining quantitative and qualitative methods, as the strengths of one method mitigate the weaknesses of the other. General guidance for when to apply these methods is provided in Table 1.

To properly integrate quantitative and qualitative data, testers should:

1. Measure each of the HSI concepts with both quantitative and qualitative methods.
2. Integrate measures of HSI concepts into the test design so that measures of HSI can be directly linked to measured operational performance.
3. Collect the data in a way that permits unique operator responses to be compared with specific events in test and operational conditions.

As with all measurement, testers should ensure that measures are as precise as possible and minimize measurement error. Existing validated surveys reduce error and should be used in lieu of custom-written questions whenever possible. A repository of these scales is available on the DOT&E website. However, validated measures do not yet exist for some HSI concepts and custom-written questions may be required. If testers plan to use custom questions, they must be pre-tested with the target audience. To reduce errors in qualitative measures, interviews and focus groups should be conducted by trained moderators using a common plan.

Integrating HSI Measures into Operational Tests

Testers must describe how they plan to implement HSI measures in the test plan. Because operators and maintainers should be able to employ the system effectively under all relevant conditions, testers should collect HSI data across these conditions. Testers should construct a data collection plan that summarizes (1) what HSI concepts will be measured, (2) what method(s) they will use to measure them, and (3) when and how they will capture the data.

The data collection plan must describe, for each test condition, the number of expected responses to HSI measures, and the number of different operators or maintainers who will provide these responses. Test plans should follow both the general guidance discussed in this memorandum and the best practices described in the attachments. Testers should submit HSI measures and plans to DOT&E as part of the test plan.

Robert F. Behler
Director

Attachments:
As stated

Attachment 1: Quantitative Methods

Quantitative methods are most useful for evaluating the overall quality of HSI across the operational space. These methods include recording behavior and conducting surveys and should serve as the primary means by which testers judge the degree to which HSI affects operational performance. Standardization enables testers to link operator/maintainer ratings to test design factors and demonstrated performance via statistical models.[1] Quantitative methods are good for characterizing HSI across operational conditions and linking human experience to operational performance.

Behavior & Human Performance

Evaluating the degree to which HSI affects operational performance requires measuring both how the operator interacts with the system and the resultant effect on operationally relevant outcomes. For example, an operator's ability to detect something is directly related to the usability of his or her detection devices. Measuring operator behavior (via instrumentation or systematic observation) is a way to directly, objectively, and accurately quantify this interaction. Operator behavior is complex, however, typically arising from multiple factors, including factors related to the person and his or her operational environment. Pairing a behavioral measure with a survey can help testers clarify which HSI concept is driving the behavior they observed.

Surveys

It is important for testers to minimize measurement error in the surveys they choose. They can do so by following a few key guidelines:

1. **Surveys should be scales**. Scales use multiple questions to get at the same concept from different angles. Analysts aggregate responses to these questions to get a clearer picture of the concept of interest.
2. **Testers should use scales that have undergone validation**. Validated scales for operational test generally are shorter and more accurate than custom surveys. Rigorous validation refines the items that comprise the scale, chooses only the best performing ones, and demonstrates that the scale measures what it intends.
3. **Testers should avoid overburdening respondents with surveys**. Burden decreases data quality and increases measurement error. To avoid burden, testers should consider not just the number of questions, but their quality, context, and manner of administration.

All surveys in operational test should use validated scales unless there is a compelling reason—which should be articulated—why a custom survey must be written. Since validated scales already exist for most HSI concepts relevant to operational testing, use of a custom survey should be rare. Validated scales represent HSI concepts at an appropriate level of detail in fewer

[1] Note that quantitative methods are inefficient and ineffective for discovering and diagnosing problems because the problem must be anticipated by the question writer, and its response options are pre-determined. The qualitative methods described in Attachment 2 are typically better suited to problem identification, discovery, and diagnosis.

questions and with less measurement error than custom scales. Scales, whether custom or validated, should be developed to quantify a clearly defined HSI concept rather than diagnose or discover problems.[2]

Occasionally, testers have a real need to evaluate the usability of multiple systems (e.g., if there are multiple variants/vendors or operationally distinct sub-systems). In these instances, testers should not write a different set of questions for each system. Instead, they should apply the same usability scale to each system to facilitate comparisons where appropriate.

Custom scales

If a custom survey must be used, the test plan must describe how testers will minimize measurement error. A common set of steps for this process exists for all programs. The plan must (1) explain what HSI concept the custom survey intends to measure and (2) describe how the survey will be pretested. All custom surveys in operational testing must be pretested. The main goals of pretesting are:

1. **Content validation:** an independent subject matter expert (SME) examines the questions to determine whether they sufficiently cover the concept that is to be measured.
2. **User (Face) validation:** interview representative respondents (leveraging SME involvement) and ask what they think each question means and how they would choose an answer.

Each step of pretesting should provide modifications to the questions. Pretesting can be accomplished outside of test, or before IOT&E during early Operational Assessments or Limited User Tests. When custom surveys are used, the plan for pre-testing them should be available for review. Whenever possible, testers should involve my staff in the early creation of these surveys to ensure a smooth approval process.

If writing custom questions, testers should take the same top-down approach used to pick the concepts measured in test. Testers should ask themselves what operationally relevant aspect is being evaluated (e.g., "Usability of the digital interface") and why that aspect could have a bad outcome. Scales are meant to aggregate questions into single scores of a concept. Questions should get at that concept from slightly different angles. If the scale is written correctly, it should make sense to average the responses. Testers should plan in advance how these items are going to be combined and describe that in the evaluation framework.

Testers can also reduce measurement error by writing good questions. Well-written question are at minimum:

1. **Singular.** Each question asks about a single, distinct thing.
2. **User friendly.** Write questions the way the respondent would say it.
3. **Neutral.** Questions do not lead the respondent to an answer.
4. **Knowable.** The respondents can accurately assess their experience.

[2] Focus groups and interviews are much more effective at diagnosing problems (see Attachment 2).

Attachment 1: Quantitative Methods

Avoid the use of "Not Applicable" response options. If a scale is not relevant for a group of people, then it should not be administered to them. For a well-constructed scale, no respondents should legitimately require an NA response. If the NA is truly needed, it should be spatially separated and visually distinct from the other response options.

When choosing the number of scale points (i.e., a 1–5 or a 1–7 scale), testers should use at least five and preferably seven. The evidence is clear that using rating scales with fewer than five points produces less reliable estimates. Testers can anchor scale points to self-reported internal experiences (e.g., workload, trust) or external behavior (e.g., perceptions of how frequently a behavior occurred).The U.S. Army Research Institute's Questionnaire Construction Manual (1989) contains various scale anchor descriptors that have undergone validation testing and may be useful to testers as they develop rating scales.

Additional References:

1. Kortum, P. (2016). *Usability Assessment: How to Measure the Usability of Products, Services, and Systems.* Human Factors and Ergonomics Society, Santa Monica, CA.
2. Sauro , J. & Lewis, J. (2016) *Quantifying the User Experience: Practical Statistics for User Research* (2nd Edition). Morgan Kaufmann, Cambridge, MA.
3. Fowler, F. (2014). *Survey Research Methods.* SAGE Publications, Inc., Thousand Oaks, CA.
4. Leary, M. (2014). *Introduction to Behavioral Research Methods.* Pearson Education Limited. London.

Attachment 2: Qualitative Methods

Qualitative methods serve an important but distinct role from quantitative methods and the two methods are not interchangeable. Testers should use interviews and focus groups to understand *why* something happened, rather than to establish *whether or not* it happened.[3] Testers should avoid just asking yes/no questions or getting ratings during interviews and focus groups. Instead, testers should use this opportunity to collect details that they can only obtain from a human who experienced the mission or task. In general, qualitative methods are more burdensome than quantitative methods and should only be used to mitigate the shortfalls in quantitative methods, most notably their limited capacity to diagnose and discover problems.

To independently evaluate qualitative data, analysts need access to the primary source information. Qualitative data that have been paraphrased, interpreted, or condensed no longer permit independent analysis. Interviews and focus groups must have an independent record. Testers should take the lowest effort and highest fidelity option of using video and/or audio recording followed by transcription. If a program cannot record the conversation (e.g., for security reasons), then a transcriptionist is required to accompany the moderator to all pre-scheduled interviews and focus groups. The transcriptionists should record the conversation as close to verbatim as possible.

Qualitative comments can be very illuminating, but are subject to bias. It is very easy for moderators to influence responses or fail to address important issues. Consequently, they should not be the sole method for evaluating HSI. Conducting an interview or focus group is a specific skill, and test teams should ensure that all moderators are properly trained to implement these methods prior to the test event.

Writing Qualitative Questions

There are three general approaches to interview and focus group data collection: (1) structured, (2) semi-structured, and (3) unstructured. In a fully structured interview or focus group, the phrasing, order, and specific follow-ups for questions are completely determined before the test. Structured interviews require significant planning and are inflexible, which prevents moderators from following up on topics that were not built into the interview ahead of time. Unstructured interviews and focus groups most closely resemble a regular conversation, but require extremely skilled moderators to extract meaningful or comparable data. Consequently, unstructured interviews should not be used for operational test. Semi-structured interviews blend these two approaches: the first few questions of each topic are written, the progression of topics is the same between moderators, and general follow-up questions are pre-written (e.g., "Do you think that affected your mission?", "[Follow up] Can you expand on that?"). In operational tests, a semi-structured approach is most effective. Semi-structured

[3] Human perception is inherently limited, and qualitative responses can be strongly influenced by the person asking the question. Qualitative data run the risk of either being biased in reality or of being dismissed as biased if someone does not agree with the response. Objective, quantitative data are better to show that something has happened.

interviews allow flexibility to follow up on unanticipated topics while reducing error with some standardization.

If the goal of the interview or focus group is problem diagnosis, questions should start broadly and then move to a more narrow focus. Moderators should let operators share what they feel is most important, then follow up on that. Operators and maintainers have previously reported during testing that there are problems they wanted to raise, but the specificity of the questions asked gave no opportunity to do so. The moderator's role in an interview or focus group is to keep the operator talking about relevant information. In a semi-structured environment, this requires practice.

If testers intend to use qualitative methods as part of the formal evaluation, they must submit an interview or focus group guide plan with the test plan. This plan should detail (1) the general progression of topics, (2) the question that will introduce each topic, (3) a list of areas of interest or follow-ups for each topic, and (4) phrasing of generic follow up questions. It should also describe who will take part (an estimate of how many interviews or how many focus group participants) and when. This guide should be provided as part of the test plan.

Interview vs. Focus Groups

The choice to use an interview or a focus group to obtain qualitative data will depend on individual program considerations. Focus groups tend to be less rigorous and more susceptible to biased responses than interviews. However, focus groups allow varying perspectives on an issue brought up by a single individual. Test plans should explicitly state the rationale for choosing one or the other.

Comment Boxes

Comment boxes involve different trade-offs compared to interviews or focus groups. Comment boxes are less susceptible to bias or influence from interviewers or peers, but they limit testers' ability to follow up on comments and increase operator burden. If comment boxes will be employed, testers should use them sparingly to minimize operator burden. Testers should also avoid requiring that operators complete a comment box based on how they answer rating scale questions (i.e., "If you gave a 1 or a 2, explain why"), as this can bias both responses. Because of these trade-offs, testers should use comment boxes to cue topics for interviews or focus groups rather than as a primary means of qualitative data collection.

Qualitative Test Team Observation

Qualitative data can also arise from the direct reports of third-party observers. Most frequently, these are ad-hoc test cell observations providing context to deviations or unexpected events, the practices of which are beyond the scope of this guidance. If, however, qualitative observations will directly support system evaluation, the test plan must explain the collection methodology. Adequate methods minimize the risk of bias or measurement error. Examples of techniques that reduce error include standardizing observation or data collection forms, establishing shared explicit definitions, and utilizing observers who are SMEs in the system's operational context.

Additional References:

1. Maxwell, J. (2013). *Qualitative Research Design: An Interactive Approach* (3rd Edition). SAGE Publications, Inc., Thousand Oaks, CA.
2. Miles, M., Huberman, M., & Saldaña, J. (2014). *Qualitative Data Analysis: A Methods Sourcebook.* SAGE Publications, Inc., Thousand Oaks, CA.
3. Krueger, R. & Casey, M. (2015). *Focus Groups: A Practical Guide for Applied Research.* SAGE Publications, Inc., Thousand Oaks, CA.
4. Seidman, I. (2013). *Interviewing as Qualitative Research.* Teachers College Press, New York, NY.