INSTITUTE FOR DEFENSE ANALYSES

Surveys in Operational Test and Evaluation

Laura J. Freeman, *Project Leader*
Rebecca Grier, *Principal Author*

February 2015

Approved for public release;
distribution is unlimited.

IDA Document NS D-5410
Log: H 15-000034

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

Approved for public release; distribution is unlimited.

The Institute for Defense Analyses is a non-profit corporation that operates three federally funded research and development centers to provide objective analyses of national security issues, particularly those requiring scientific and technical expertise, and conduct related research on other national challenges.

About This Publication

This presentation, an outgrowth of work conducted under the DOT&E Test Science Project BD-9-229990, is intended to communicate the contents of Dr. James M. Gilmore's OT&E Survey memo to the Human Systems Integration (HSI) community. By engaging the HSI community, we hope to improve measurement of HSI during operational test and evaluation. More specifically, the presentation covers the following four points: (1) an overview of the OT&E Survey memo; (2) the relevance of the memo to the HSI community; (3) the capabilities and limitations of surveys as measures; and (4) the availability of survey-based HSI measures.

Acknowledgments

The IDA Technical Review Committee was chaired by Mr. Robert R. Soule.

Approved for public release; distribution is unlimited.

Surveys in Operational Test & Evaluation

Rebecca A. Grier, Ph.D.
Institute for Defense Analyses

Approved for public release; distribution is unlimited.

Abstract

Recently Dr. Gilmore signed out a memo providing Guidance on the Use and Design of Surveys in Operational Test and Evaluation. This guidance memo helps the HSI community to ensure that useful and accurate HSI data are collected. Information about how HSI experts can leverage the guidance will be presented. Specifically, the presentation will cover what HSI metrics can and cannot be answered by surveys.

Approved for public release; distribution is unlimited.

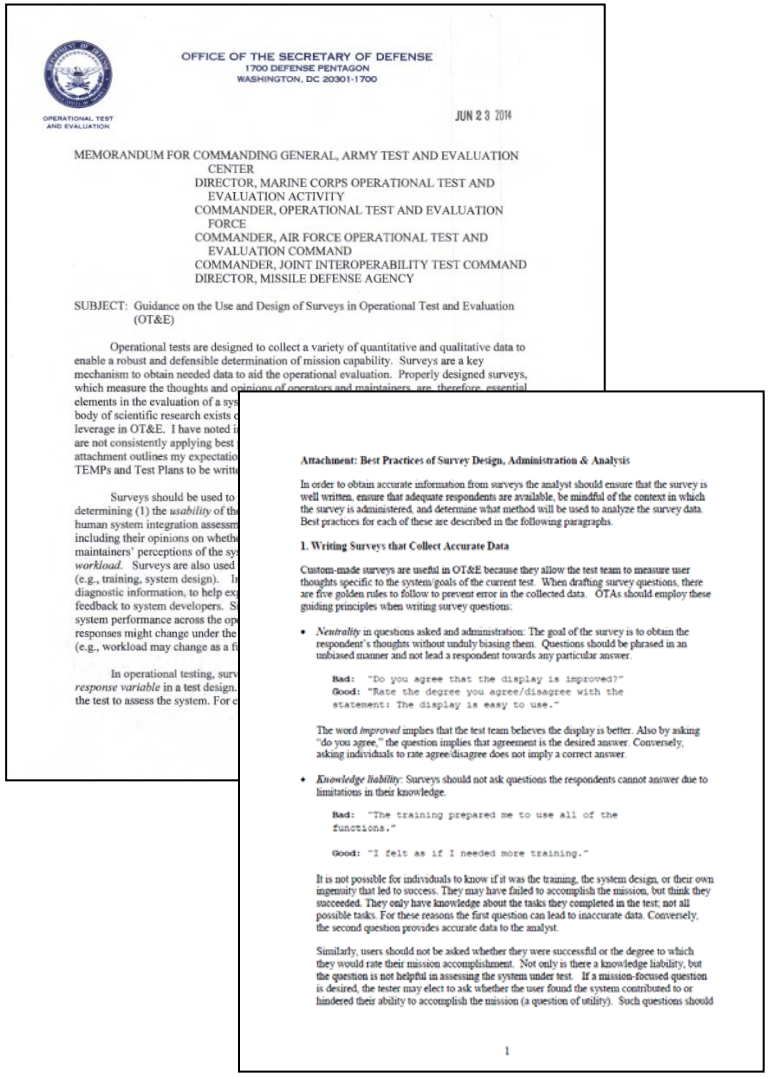
Goals

- What is in the Survey Guidance Memo to OT&E?
- How can we leverage memo to improve HSI measurement?
- What can surveys measure and what can't they measure?
- What survey based human factors measures are available?

Approved for public release; distribution is unlimited.

DOT&E Guidance on Surveys June 2014

- **Surveys are an important aspect of DOT&E evaluation**
- **Surveys should be used to (determine)**
  - the ***usability*** of the system
  - the operators' thoughts of the system's *utility*
  - maintainers' thoughts of the system's *maintainability*
  - the effects of system design on ***workload***
- **Academically-established surveys should be used for *human factors* constructs**
- **Use surveys only when appropriate**
- **It is essential to understand the goal of why you are conducting the survey**
- **Employ best practices for writing and administering surveys**
  - Memo provides a best practices guide attachment



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 sys
body of scientific research exists c
leverage in OT&E. I have noted i
are not consistently applying best
attachment outlines my expectatio
TEMPs and Test Plans to be writte

Surveys should be used to
determining (1) the *usability* of the
human system integration assessm
including their opinions on whethe
maintainers' perceptions of the sy
workload. Surveys are also used
(e.g., training, system design). I
diagnostic information, to help ex
feedback to system developers. S
system performance across the op
responses might change under the
(e.g., workload may change as a fi

In operational testing, surv
response variable in a test design.
the test to assess the system. For e

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

1

Surveys Measure *Thoughts* about Performance Only

- **Not Time:**

"Put your hand on a hot stove for a minute, & it seems like an hour. Sit with a pretty girl for an hour, & it seems like a minute."

- Albert Einstein

- **Not Accuracy:**

		Truth	
		Success	Failure
Belief	Success	☺	!!!
	Failure	☹	😐

Bad Design = Mismatch Between Truth & Belief

3 Mile Island

Vincennes Incident

- **Not Situation Awareness:**

"….There are things we do not know we don't know." - Donald Rumsfeld

Surveys Are An Important Aspect of DOT&E

Performance Data

What: time & accuracy

Subject Matter Expert Observation

How: actions taken, moments of frustration, etc.

User Surveys

Why: usability, workload, thoughts about specific design features, etc.

- Questions known ahead to be appropriate for test
- Finite set of concise responses possible

Effectiveness & Suitability

User Interviews

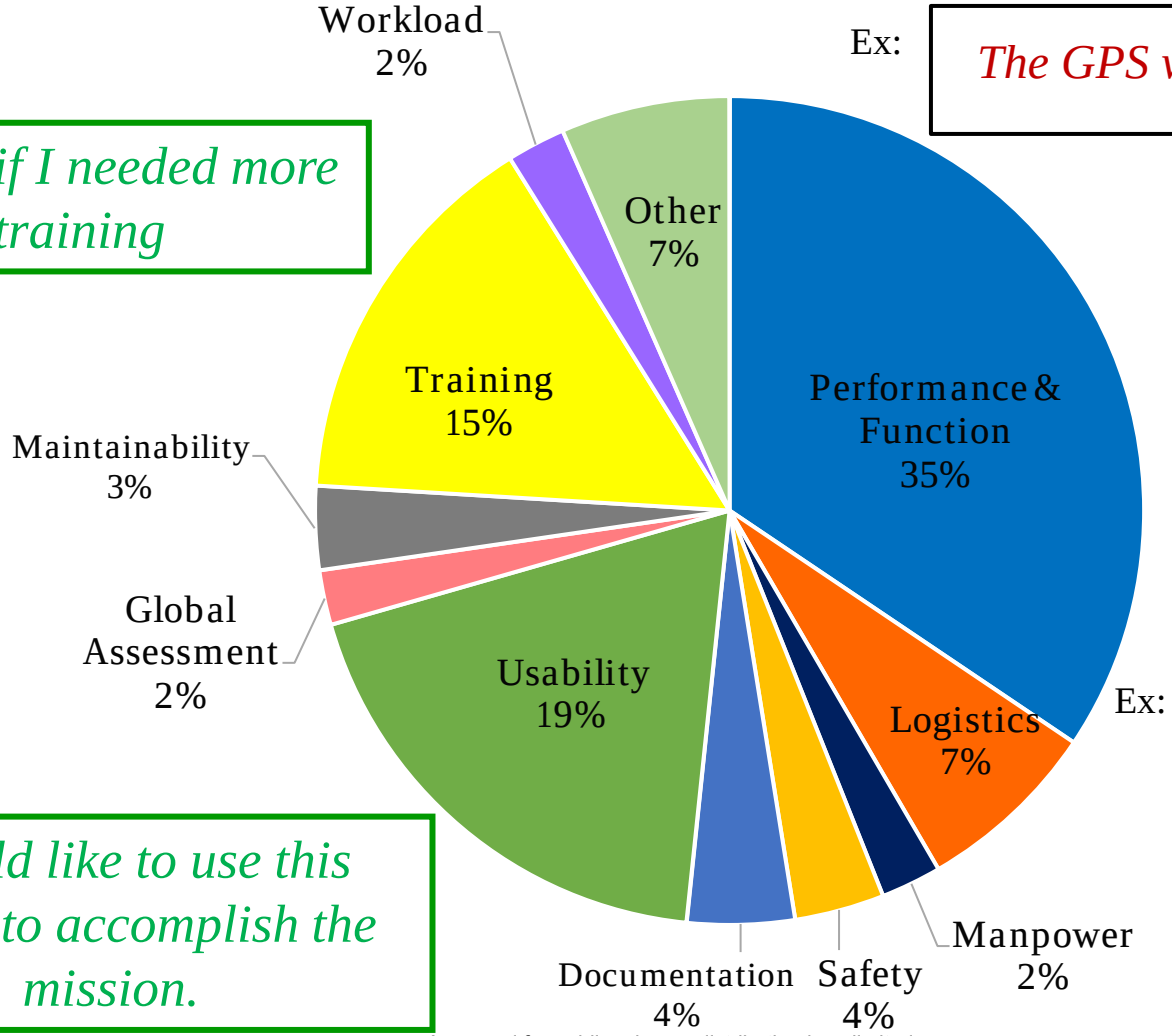
Why: non-specific thoughts

- Questions in response to rare or unexpected test events
- Infinite number of possible responses
- Possible responses are long

Review of OT&E Surveys:
Percentage of Questions for each Topic

Ex: *The GPS was accurate.*

Ex: *I felt as if I needed more training*

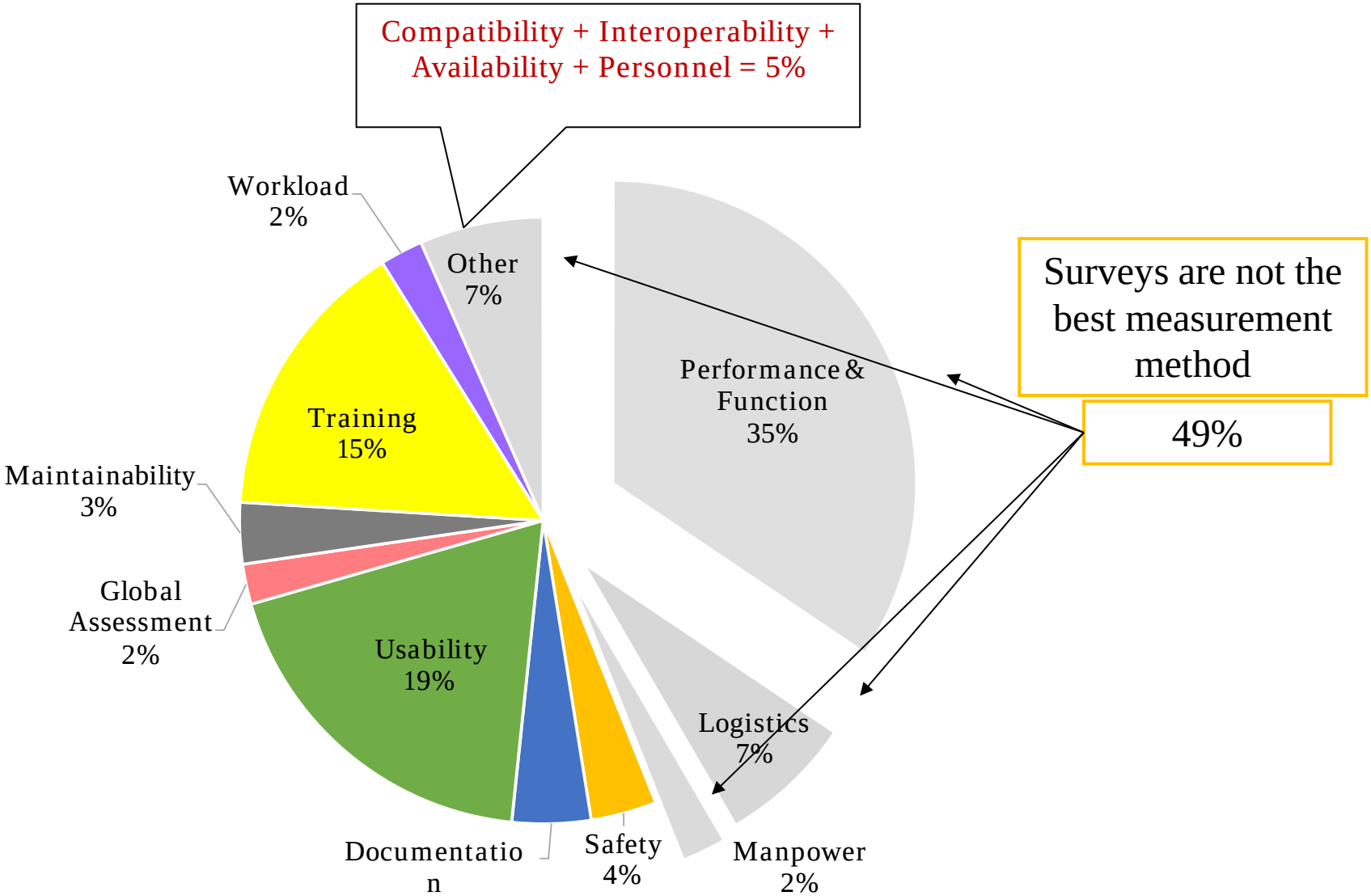


Ex: Rate the adequacy of land-based administrative support for supply support

Ex: *I would like to use this system to accomplish the mission.*

Approved for public release; distribution is unlimited.

Review of OT&E Surveys: Percentage Appropriate Questions



DOT&E Vetted Example Questions

- I would like to use this system to accomplish the mission.
- The instructor presented the material clearly.
- I feel as though additional training is needed.
- The _(e.g., work station, cockpit)_ is well organized.
- I did not have the information needed to __(e.g., execute the mission, perform a specific task)__.
- It was difficult to _(e.g., perform a specific task)_.
- _(e.g., Equipment, Controls, Information, Features, Applications)_ are easily accessible.
- Are there any improvements that you would make to the system?
- Please comment on any safety concerns that you have.

When to Design A Survey

Appropriate

1. **There Isn't an Appropriate Academically-Established Survey**

2. **Measure Specific User/Maintainer Thoughts**
   - Utility/Ease
   - Specific features/ components
   - Specific issues with regard to CONOPS

3. **Quantify Observer Ratings**

"A good plan is like a road map: it shows the final destination and usually the best way to get there."
H. Stanely Judd

Not Appropriate

1. **Obtain Random Thoughts of Respondents**
   - Interview

2. **Measure Performance**
   - Time
   - Accuracy via Appropriate Physical Measure
   - Observers

3. **Measure Requirements**
   - Appropriate Physical Measure
   - See e.g., MIL-STD-1472G

4. **Measure Situation Awareness**
   - Numerous techniques in Human Factors Literature
   - Salmon et al (2006) for review

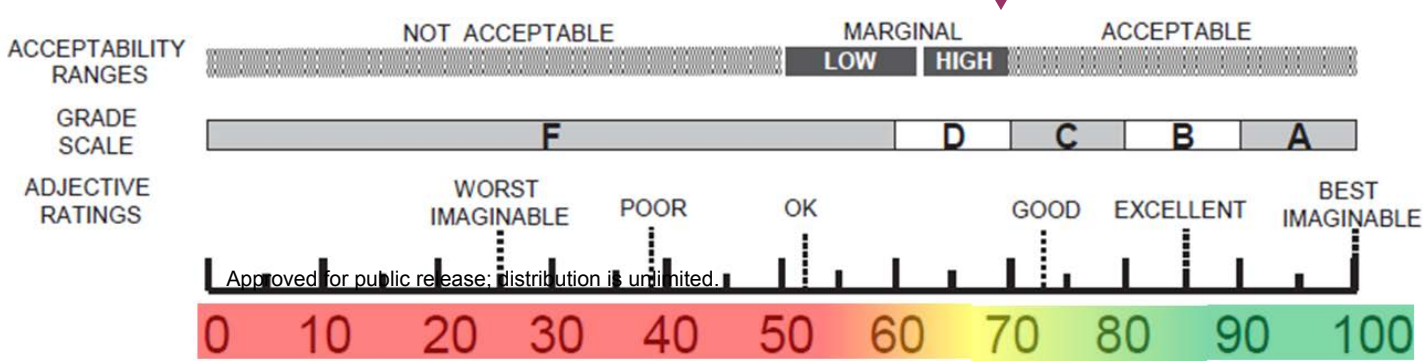
System Usability Scale (SUS)

- **Most Used Usability Survey**
  - 43% of usability studies
  - Sauro & Lewis (2009)
- **10 Questions**
  - 5 point alternating Likert response
  - Administered immediately after user completes tasks
- **Score: (bad)0 – 100(good)**
  - Subtract 1 from each odd question
  - Subtract each even question from 5
  - Multiply the sum of above by 2.5
  - 2.5 [20+Q1 + Q3 + Q5 + Q7 + Q9 - Q2 - Q4 - Q6 - Q8 – Q10]

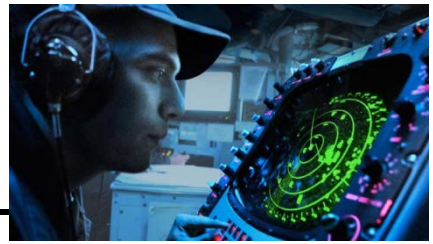
	Strongly disagree				Strongly agree
1. I think that I would like to use this system frequently	1	2	3	4	5
2. I found the system unnecessarily complex	1	2	3	4	5
3. I thought the system was easy to use	1	2	3	4	5
4. I think that I would need the support of a technical person to be able to use this system	1	2	3	4	5
5. I found the various functions in this system were well integrated	1	2	3	4	5
6. I thought there was too much inconsistency in this system	1	2	3	4	5
7. I would imagine that most people would learn to use this system very quickly	1	2	3	4	5
8. I found the system very awkward to use	1	2	3	4	5
9. I felt very confident using the system	1	2	3	4	5
10. I needed to learn a lot of things before I could get going with this system	1	2	3	4	5

Further Reliability & Validity Assessments of SUS

- **Tullis & Stetson (2004)**
  - Compared SUS to other usability surveys
  - More accurate conclusions with smaller sample sizes
- **Bangor, Kortum, & Miller (2008)**
  - 2324 tests over 10 years wide range of systems
  - High internal consistency (r = 0.91)
  - Correlated to user-friendliness rating (r = 0.806)
  - Sensitive to usability differences
- **Lewis & Sauro (2009) & Borsci et al (2009)**
  - Two Interdependent Factors
    - » Usability (Items 1, 2, 3, 5, 6, 7, 8, & 9)
    - » Learnability (Items 4 & 10)



Recommended Modifications to SUS



- **Learnability** (items 4 & 10)
  - Key Component of HSI
  - Key Component of Effectiveness
  - Key Component of Suitability
- **Slight Modifications to Text Suggested for Military Operators**
  - Item 1: Military missions are not frequent
  - Item 7: Clarify baseline
- **User Sophistication is a Test Design Issue**

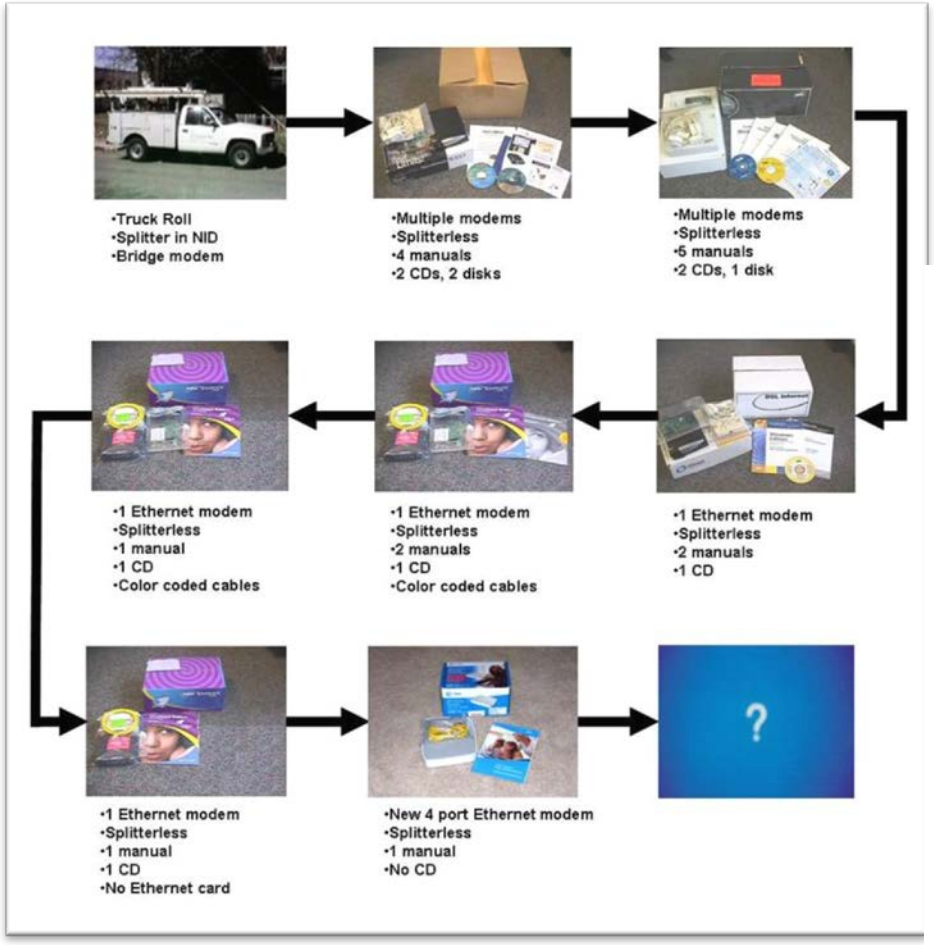
ISO: "The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency & satisfaction in a specified context of use."

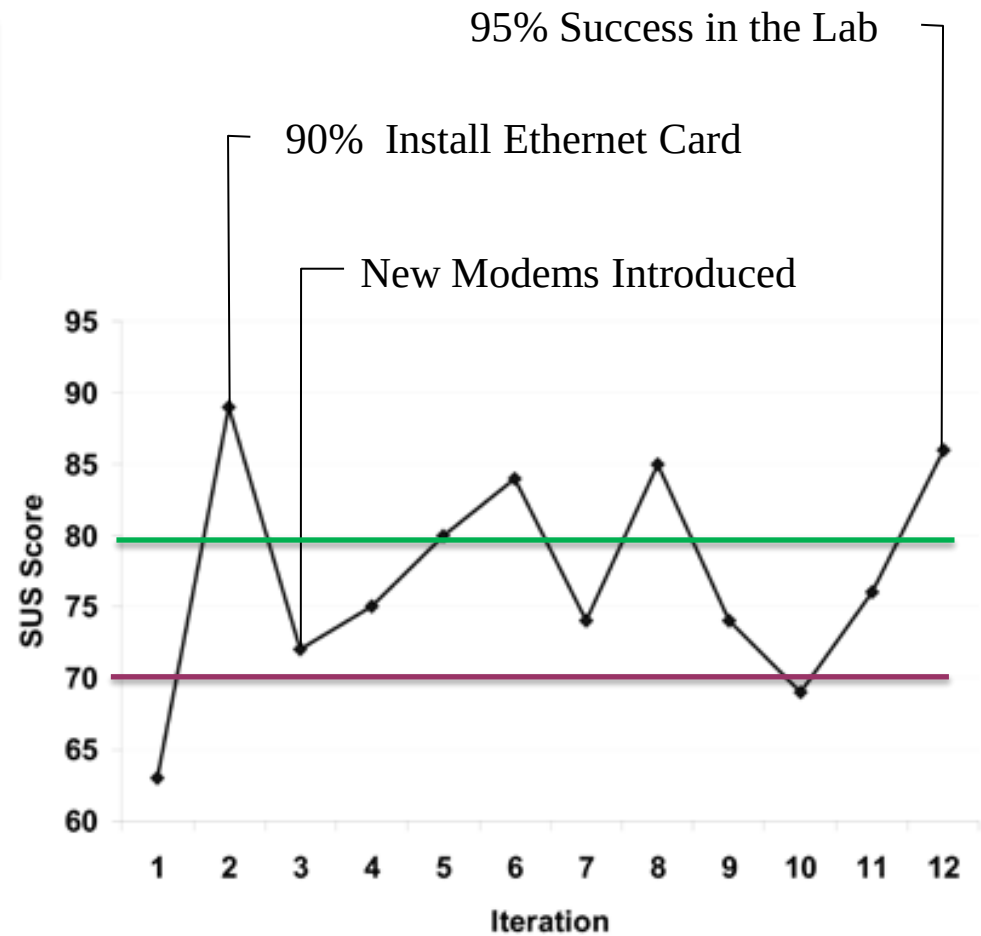
Effective: "mission accomplishment when used by representative personnel in the (expected environment) …considering organization, training…"

Recommended Military SUS

1. I think that I would like to use this system ~~frequently~~ *to accomplish the mission*.
2. I found the system unnecessarily complex
3. I thought the system was easy to use
4. I think that I would need the support of a technical person to be able to use this system
5. I found the various functions in this system were well integrated
6. I thought there was too much inconsistency in this system
7. I would imagine that most people *with my MOS* would learn to use this system very quickly
8. I found the system very awkward to use
9. I felt very confident using the system
10. I needed to learn a lot of things before I could get going with this system.

Case Study: DSL Self Installation



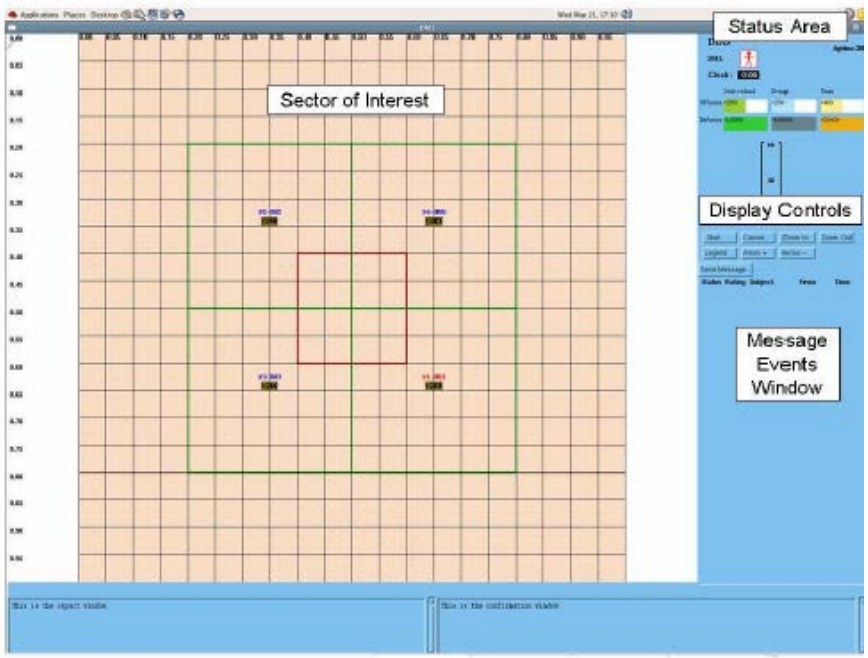


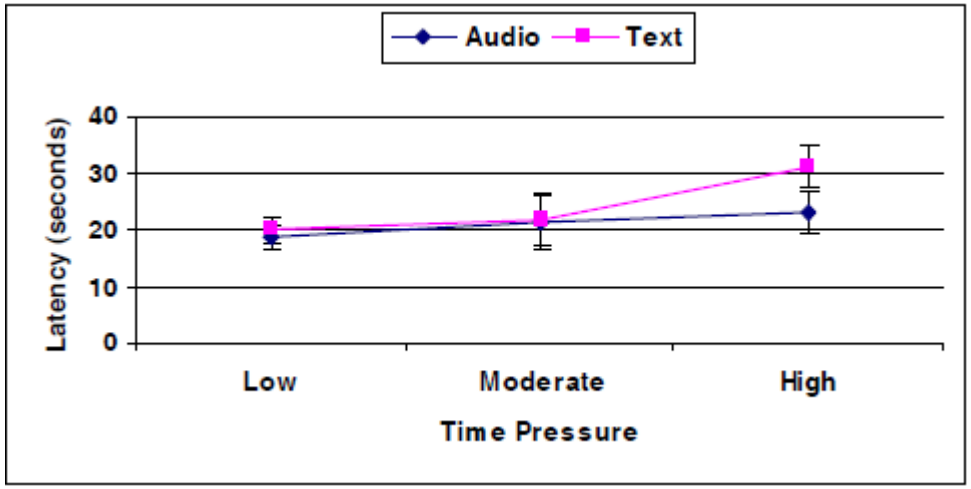
Kortum, P., Grier, R. & Sullivan, M. (2009). DSL Self-installation: From Impossibility to Ubiquity. *Interfaces, 80*, 12-14.

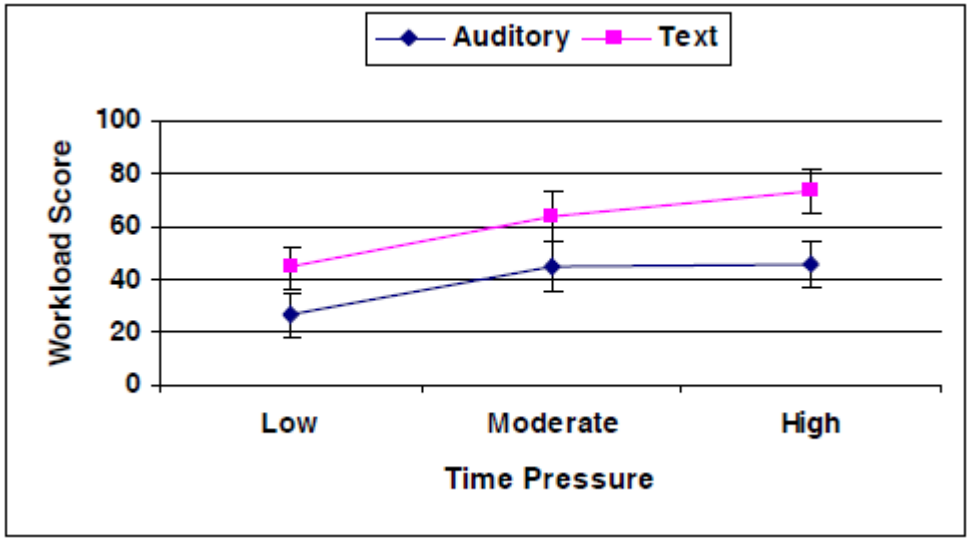
Some Common Self Report Workload Measures

Measure	Published	Citations	Description
Cooper Harper & Variants - Modified Cooper Harper (1992) - Bedford (1990)	1969	2036	1 -3 Questions Score: (good)1-10 (bad) High workload: 4 One-dimensional/Not Diagnostic Task Relative No Theory
Crew Status Survey/ Integrated Workload Scale	1993/2005	26/63	1 Question Score: (good) 0 -7/9 (bad) High Workload: ???? Uni-dimensional/Not Diagnostic Task Agnostic No Theory
NASA-TLX - Original/Weighted - RawTLX (RTLX)/ Unweighted	1988	7020	6 or 21 Questions Score: (good) 0 -100 (bad) High workload: ????? Multi-dimensional/ Diagnostic Task Agnostic Resource Pool Theory
MRQ	2001/2007	217	Up to 17 Questions Score: (good) 0 -100 (bad) High workload: ?????? Multi-dimensional/Diagnostic Task Agnostic Multiple Resource Theory

Using NASA TLX to Compare Versions: Value of Multi-Modal System to C^2







Grier, R.A., Parasuraman, R., Entin, E., Bailey, N., & Stelzer, E. (2008). A test of intra- versus inter-modality interference as a function of time pressure in a warfighting simulation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting in New York City.*

Conclusions

- **HSI is an important component of Operational Test & Evaluation**
- **All measurement should be done with a goal in mind and according to best practices**
- **Academically vetted surveys tell the test team about HSI constructs**
  - **Usability:** are there likely to be critical errors in operational context?
  - **Workload:** how much effort is required to achieve performance level?
- **Situation Awareness should not be measured via survey**





Questions?