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
About This Publication

This presentation, an outgrowth of work conducted under the DOT&E Test Science Project BD-8-229990, 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.

Copyright Notice

© 2015 Institute for Defense Analyses
4850 Mark Center Drive, Alexandria, Virginia 22311-1882 • (703) 845-2000.

This material may be reproduced by or for the U.S. Government pursuant to the copyright license under the clause at DFARS 252.227-7013 (a)(16) [Jun 2013].
Surveys in Operational Test & Evaluation

Rebecca A. Grier, Ph.D.
Institute for Defense Analyses
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.
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?
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.
"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 Time:

• Not Accuracy:

<table>
<thead>
<tr>
<th>Truth</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Success</td>
<td>🎈</td>
</tr>
<tr>
<td>Failure</td>
<td>😞</td>
</tr>
</tbody>
</table>

Belief

Success

Failure

Bad Design = Mismatch Between Truth & Belief

• 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

User Interviews
Why: non-specific thoughts
- Questions in response to rare or unexpected test events
- Infinite number of possible responses
- Possible responses are long

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

Performance & Function: 35%
Training: 15%
Usability: 19%
Documentation: 4%
Safety: 4%
Logistics: 7%
Manpower: 2%
Global Assessment: 2%
Maintainability: 3%
Workload: 2%
Other: 7%

Ex: The GPS was accurate.
Ex: I felt as if I needed more training.
Ex: I would like to use this system to accomplish the mission.
Ex: Rate the adequacy of land-based administrative support for supply support.
Review of OT&E Surveys: Percentage Appropriate Questions

Compatibility + Interoperability + Availability + Personnel = 5%

Surveys are not the best measurement method

49%
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**

#### 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

---

“A good plan is like a road map: it shows the final destination and usually the best way to get there.”

*H. Stanley Judd*
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 \[20+Q1 + Q3 + Q5 + Q7 + Q9 - Q2 - Q4 - Q6 - Q8 – Q10\]
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

  - 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

95% Success in the Lab
90% Install Ethernet Card
New Modems Introduced

# Some Common Self Report Workload Measures

<table>
<thead>
<tr>
<th>Measure</th>
<th>Published</th>
<th>Citations</th>
<th>Description</th>
</tr>
</thead>
</table>
| Cooper Harper & Variants                     | 1969      | 2036      | 1-3 Questions  
Score: (good) 1-10 (bad)  
High workload: 4  
One-dimensional/Not Diagnostic  
Task Relative  
No Theory |
| - Modified Cooper Harper (1992)              |           |           |                                                                            |
| - Bedford (1990)                             |           |           |                                                                            |
| 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                                     | 1988      | 7020      | 6 or 21 Questions  
Score: (good) 0 -100 (bad)  
High workload: ??????  
Multi-dimensional/ Diagnostic  
Task Agnostic  
Resource Pool Theory |
| - Original/Weighted                          |           |           |                                                                            |
| - RawTLX (RTLX)/ Unweighted                  |           |           |                                                                            |
| 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 C²

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?