Human Measurement

Dean Thomas, Project Leader

Chad Bieber

Rebecca Grier

Justin Mary

IDA

Goals

- What can surveys be used for in Operational Testing?
- What surveys are available?
- How are the surveys used?

Surveys in OT&E

- **Human factors constructs**
  - Usability
  - Workload
  - Others (trust, fatigue,…)
- **System utility**
- **Demographics**
  - Are users in the test similar to the population of users?
  - Are there differences associated with different user demographics?
- **Diagnostic information**
  - Why did performance reach or not reach satisfactory levels?
  - Will there be performance problems in untested conditions?

Performance

- **Performance is the measure of how well the task is being accomplished**
  - Accurate measurement of performance requires knowledge of ground truth for the test, which operators and maintainers typically do not have
  - Measures such as accuracy and time

- **Surveys measure the user thoughts while achieving that performance**
  - Individual assessment of own performance is conflated with other experiences and are not an accurate measurement of objective performance outcomes – See Dunning-Kruger effect, others [Kruger 1999, Reason 1990]
  - Thoughts and opinions are closely linked to performance, they need to be analyzed in conjunction with performance
  - Identify conflicts such as “good” usability, but poor performance [Grier 2015, Fracker from Endsley 2000]
    - » Users believed system was helping them, but they were performing poorly.

- **See Apache example for a comparison of time-based performance with workload**

Don’t measure performance with a survey.
Analyze survey responses with respect to performance.

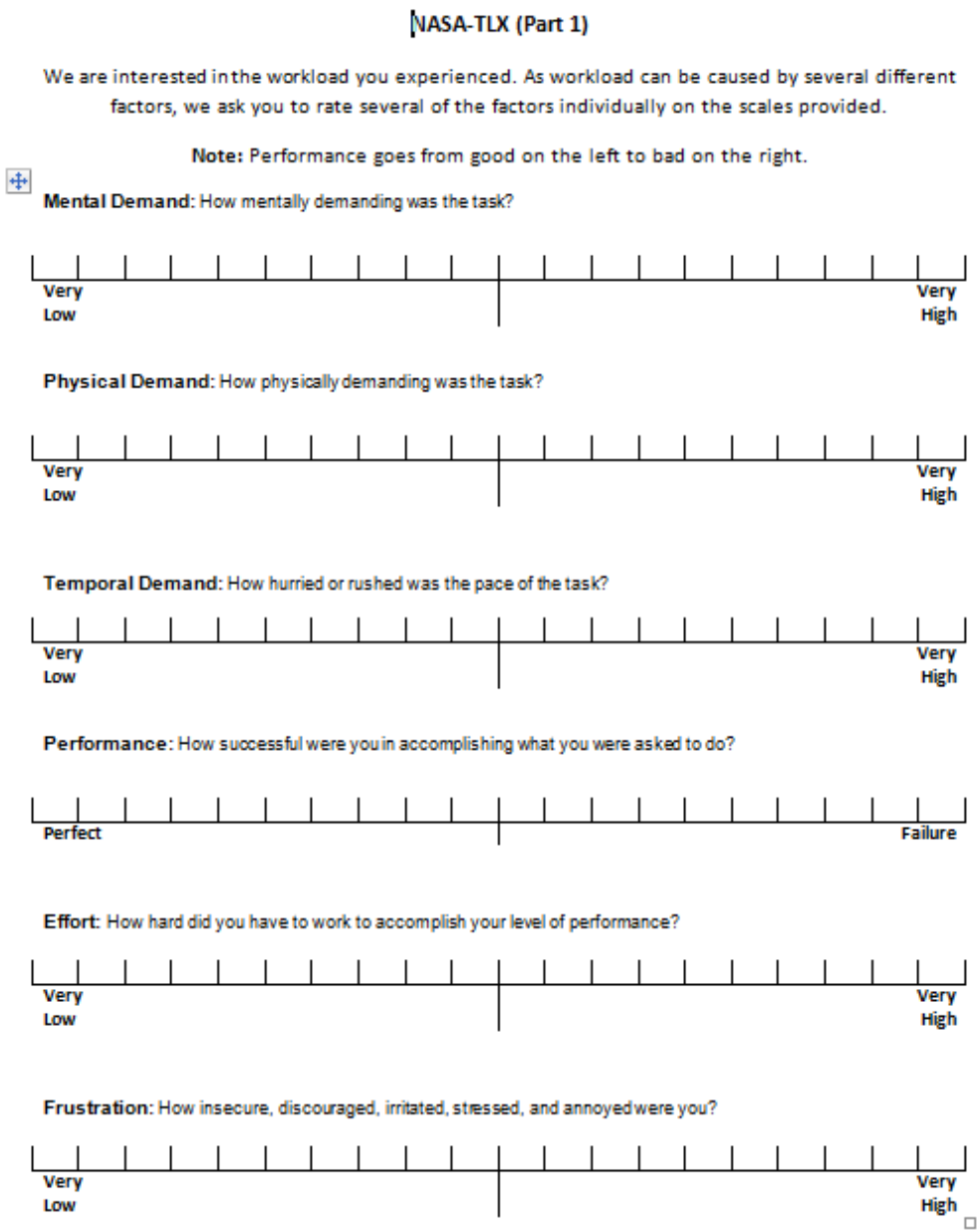
Workload

- **Workload is the demand of the task compared to resources available [Salvendy 2012]**

- **Surveys to measure workload**
  - Should be used immediately after the task of interest – within 15 minutes; intervening tasks will bias measurement
  - NASA Task Load Index (NASA-TLX) [Hart 1988]
    - » 6-question, multi-dimensional, takes 1 to 3 minutes to fill out
    - » Commonly used in research, large pool of data to compare with
    - » Proven validity in operational test environments [Hill 1992]
  - Multiple Resource Questionnaire (MRQ) [Boles 2001]
    - » 17-questions, multidimensional survey used for detailed diagnostic information
    - » Challenging analyses, small pool of data to compare with
  - Crew Status Survey (CSS)
    - » Single question, uni-dimensional, takes seconds to complete
    - » Small pool of data to compare with
    - » Proven validity in military test and evaluation [Ames 1993, George 2004]
  - Modified Cooper-Harper, Bedford
    - » Generally, not recommended for operational testing,
    - » “Acceptable” cutoff causes clustering [Linde 1988, Roscoe 1984]
    - » Poor sensitivity for high-workload tasks [Bonner 2002]

Try to use surveys that elicit more information (e.g., NASA-TLX or MRQ) and move to shorter surveys (e.g., Crew Status Survey) as test constraints demand

The NASA Task Load Index (TLX)



NASA-TLX (Part 1)

We are interested in the workload you experienced. As workload can be caused by several different factors, we ask you to rate several of the factors individually on the scales provided.

Note: Performance goes from good on the left to bad on the right.

Mental Demand: How mentally demanding was the task?

Very Low — Very High

Physical Demand: How physically demanding was the task?

Very Low — Very High

Temporal Demand: How hurried or rushed was the pace of the task?

Very Low — Very High

Performance: How successful were you in accomplishing what you were asked to do?

Perfect — Failure

Effort: How hard did you have to work to accomplish your level of performance?

Very Low — Very High

Frustration: How insecure, discouraged, irritated, stressed, and annoyed were you?

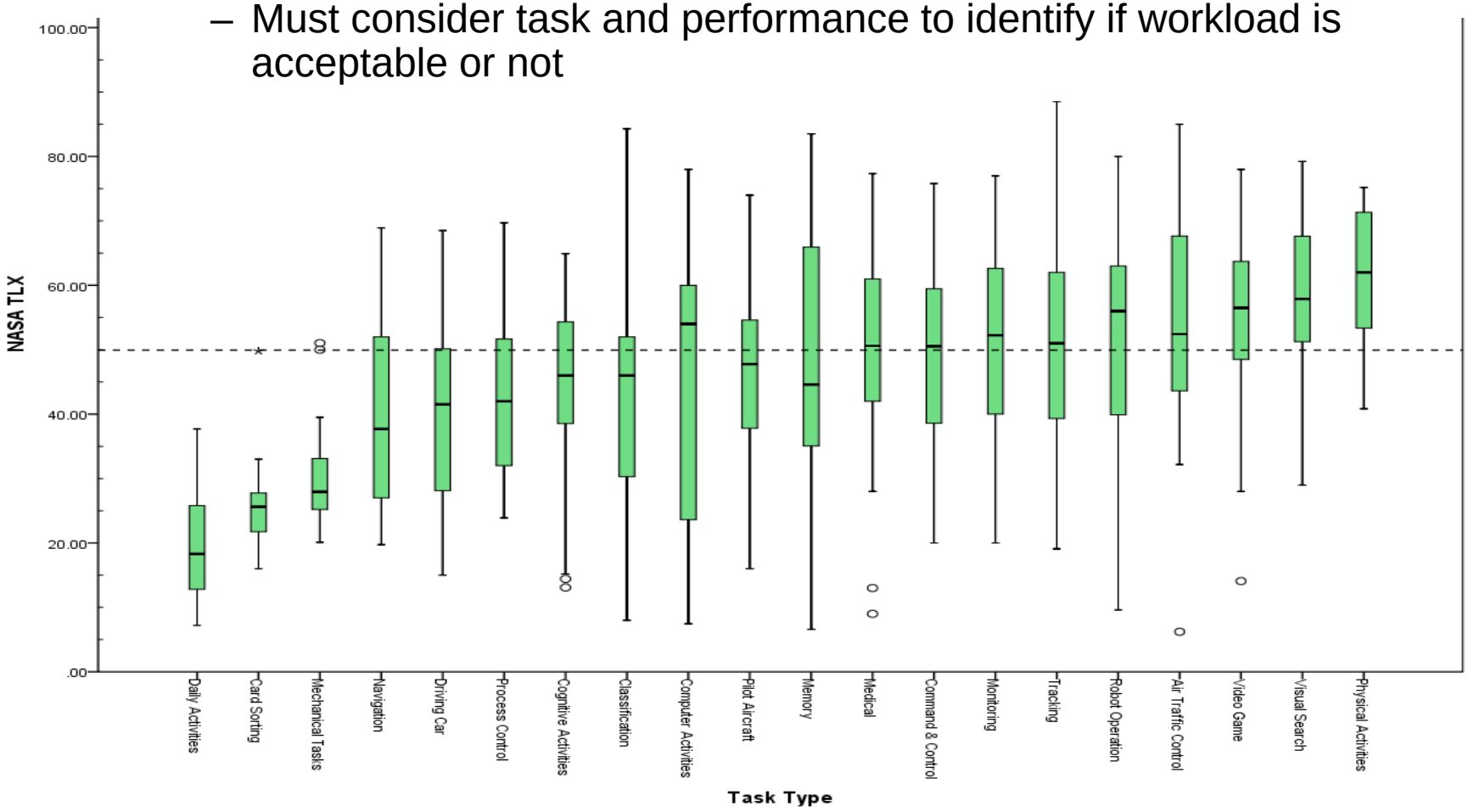
Very Low — Very High

- **21 Questions in 2 parts**
  - **Part 1:** Rate the following 6 dimensions on 100 point Scale (each box = 5 points)
    - » Mental Demand
    - » Physical Demand
    - » Temporal Demand
    - » Performance (perceived)
    - » Effort
    - » Frustration
  - **Part 2:** 15 Paired Comparisons
    - » All possible pairs of 6 dimensions
    - » Select the one that contributed more to workload in the task just completed.
- **Score:** Mean of ratings weighted by paired comparison count
  - {M(Mw)+PD(PDw)+ T(Tw)…}/15
  - 0 (Low) – 100 (High)

General Comparative Ability: NASA-TLX

[Grier 2014]

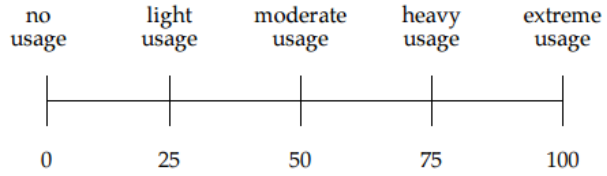
- **Range of workloads separated by task area**
  - Box plot indicates medians, quartiles, outliers
  - Must consider task and performance to identify if workload is acceptable or not



Multiple Resource Questionnaire
[Boles 2001]

MULTIPLE RESOURCES QUESTIONNAIRE for task_____

The purpose of this questionnaire is to characterize the nature of the mental processes used in the task with which you have become familiar. Below are the names and descriptions of several mental processes. Please read each carefully so that you understand the nature of the process. Then rate the task on the extent to which it uses each process, using the following scale.



Important:

All parts of a process definition should be satisfied for it to be judged as having been used. For example, recognizing geometric figures presented visually should **not** lead you to judge that the "Tactile figural" process was used, just because figures were involved. For that process to be used, figures would need to be processed tactilely (i.e., using the sense of touch).

Please judge the task as a **whole**, averaged over the time you performed it. If a certain process was used at one point in the task and not at another, your rating should **not** reflect "peak usage" but should instead reflect **average** usage over the entire length of the task.

Auditory emotional process -- Required judgments of emotion (e.g., tone of voice or musical mood) presented through the sense of hearing. ____

Auditory linguistic process -- Required recognition of words, syllables, or other verbal parts of speech presented through the sense of hearing. ____

Facial figural process -- Required recognition of faces, or of the emotions shown on faces, presented through the sense of vision. ____

Facial motive process -- Required movement of your own face muscles, unconnected to speech or the expression of emotion. ____

Manual process -- Required movement of the arms, hands, and/or fingers. ____

Short term memory process -- Required remembering of information for a period of time ranging from a couple of seconds to half a minute. ____

Spatial attentive process -- Required focusing of attention on a location, using the sense of vision. ____

Spatial categorical process -- Required judgment of simple left-versus-right or up-versus-down relationships, without consideration of precise location, using the sense of vision. ____

Spatial concentrative process -- Required judgment of how tightly spaced are numerous visual objects or forms. ____

Spatial emergent process -- Required "picking out" of a form or object from a highly cluttered or confusing background, using the sense of vision. ____

Spatial positional process -- Required recognition of a precise location as differing from other locations, using the sense of vision. ____

Spatial quantitative process -- Required judgment of numerical quantity based on a nonverbal, nondigital representation (for example, bargraphs or small clusters of items), using the sense of vision. ____

Tactile figural process -- Required recognition or judgment of shapes (figures), using the sense of touch. ____

Visual lexical process -- Required recognition of words, letters, or digits, using the sense of vision. ____

Visual phonetic process -- Required detailed analysis of the sound of words, letters, or digits, presented using the sense of vision. ____

Visual temporal process -- Required judgment of time intervals, or of the timing of events, using the sense of vision. ____

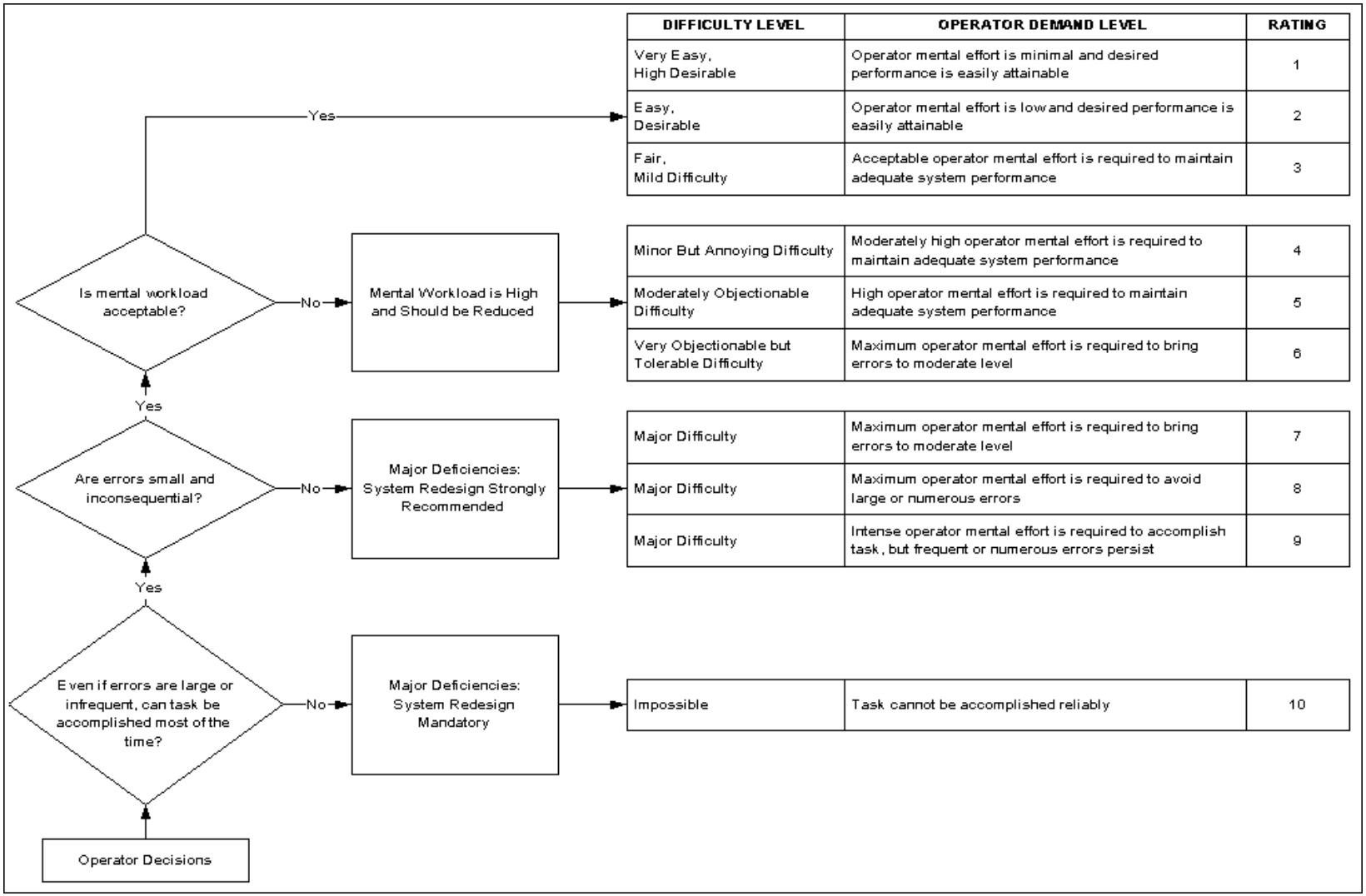
Vocal process -- Required use of your voice. ____

Crew Status Survey

From Ames and George, 1993

1) Nothing to do; No system demands.
2) Light Activity; minimal demands.
3) Moderate activity; easily managed considerable spare time.
4) Busy; Challenging but manageable; Adequate time available.
5) Very busy; Demanding to manage; Barely enough time.
6) Extremely Busy; Very difficult; Non-essential tasks postponed.
7) Overloaded; System unmanageable; Essential tasks undone; Unsafe.

Modified Cooper-Harper



DIFFICULTY LEVEL	OPERATOR DEMAND LEVEL	RATING
Very Easy, High Desirable	Operator mental effort is minimal and desired performance is easily attainable	1
Easy, Desirable	Operator mental effort is low and desired performance is easily attainable	2
Fair, Mild Difficulty	Acceptable operator mental effort is required to maintain adequate system performance	3
Minor But Annoying Difficulty	Moderately high operator mental effort is required to maintain adequate system performance	4
Moderately Objectionable Difficulty	High operator mental effort is required to maintain adequate system performance	5
Very Objectionable but Tolerable Difficulty	Maximum operator mental effort is required to bring errors to moderate level	6
Major Difficulty	Maximum operator mental effort is required to bring errors to moderate level	7
Major Difficulty	Maximum operator mental effort is required to avoid large or numerous errors	8
Major Difficulty	Intense operator mental effort is required to accomplish task, but frequent or numerous errors persist	9
Impossible	Task cannot be accomplished reliably	10

From: http://ergotmc.gtri.gatech.edu/dgt/Design_Guidelines/hndch206.htm

Usability

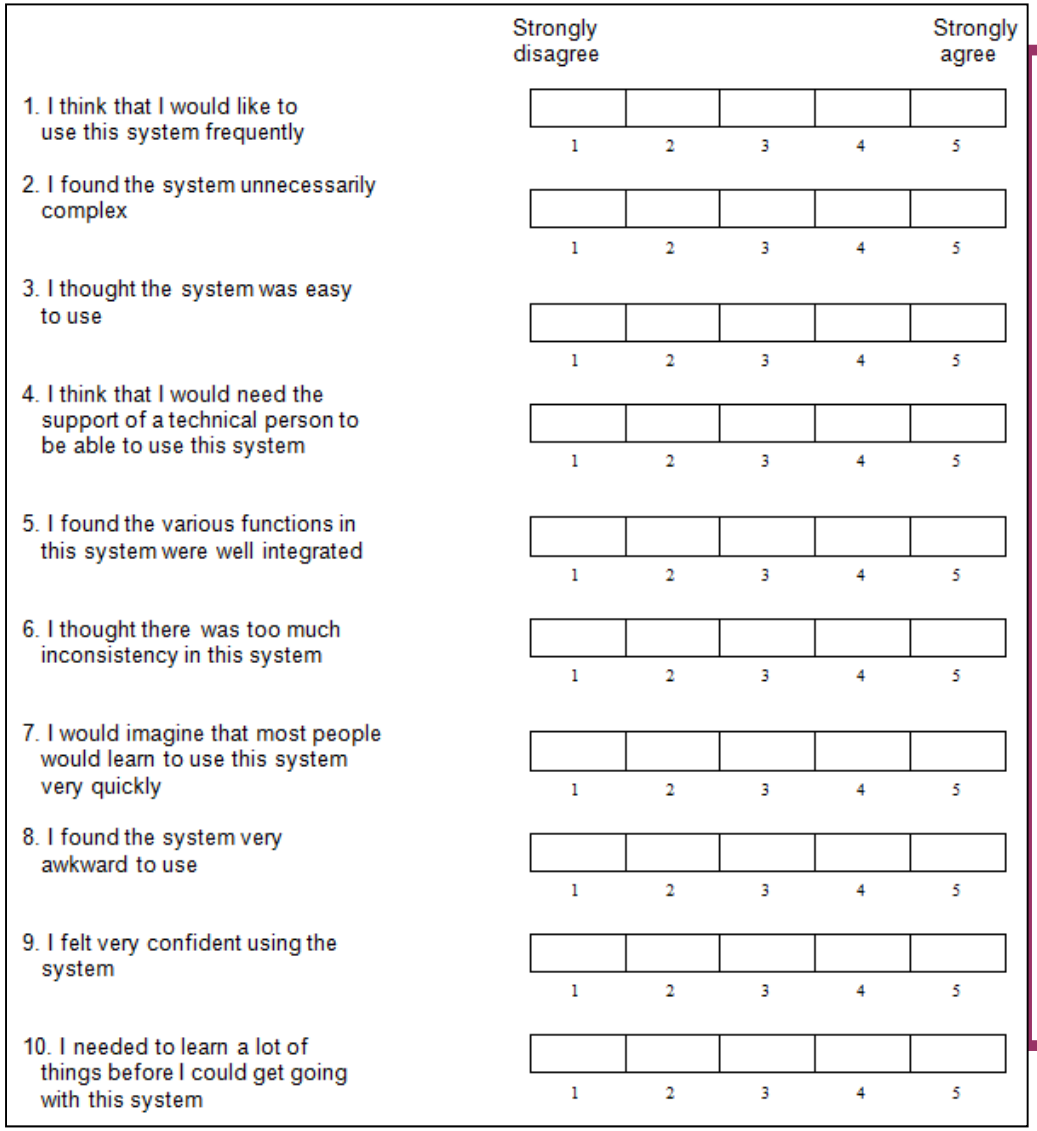
- **Usability is the effectiveness, efficiency, and satisfaction with which specified users achieve specified goals in particular environments [ISO 90241]**

- **System-level usability measurement**
  - System Usability Scale (SUS)
    - » Empirically-vetted survey
    - » Can be administered once at end of all testing or periodically to measure effect of training/experience/tasks
    - » Can measure usability between different tasks or groups of users on the same system
    - » Very broad pool of data to compare with
    - » Widely accepted ranges for Good, Fair, Poor usability

- **Component-level usability**
  - Single usability question
    - » Custom question – see other guidance on writing questions
    - » “I found the left handed torque wrench easy to use on lug nuts”

Try to use surveys that elicit more information (e.g., SUS) and move to shorter surveys as test constraints demand

System Usability Scale



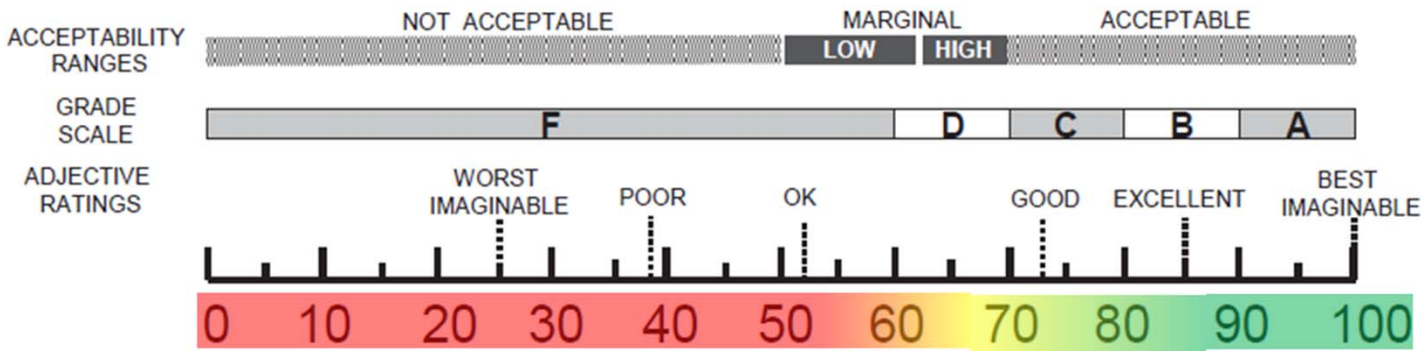
	Strongly disagree				Strongly agree
1. I think that I would like to use this system frequently	1	2	3	4	5
2. I found the system unnecessarily complex	1	2	3	4	5
3. I thought the system was easy to use	1	2	3	4	5
4. I think that I would need the support of a technical person to be able to use this system	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

Recommended Military SUS

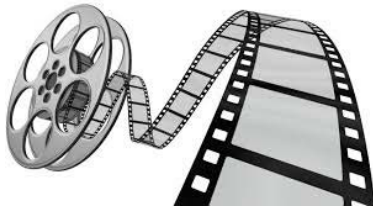
1. I think that I would like to use this system ~~frequently~~ *to accomplish the mission*.
2. I found the system unnecessarily complex
3. I thought the system was easy to use
4. I think that I would need the support of a technical person to be able to use this system
5. I found the various functions 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.

General Comparative Ability - SUS

- **Bangor, Kortum, & Miller (2008)**
  - 2,324 tests over 10 years on a wide range of systems
  - Mean = 70
  - Not Acceptable < 50 -70 < Acceptable
- **Additional Validation Studies:**
  - Brooke (1996)
  - Tullis & Stetson (2004)
  - Lewis & Sauro (2009)
  - Borsci et al (2009)
  - Sauro (2011)



Situational Awareness



- **Situational awareness:** "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of the status in the near future." [Endsley 1988]

- **Can't measure absolute SA by self-assessment**
  - Individual doesn't know ground truth to compare with self-assessment [Endsley 1994 and 1995]
  - Self-reported SA highly correlated with perception of performance [Venturino, Hamilton, Dvorachak 1989]

- **Can ask user's opinion of system support of SA**
  - Used for troubleshooting or problem detection, not measuring SA
  - Not an empirically vetted measure
  - Can be general – "The display supported my SA"
  - Or specific – "The windows were large enough to see the surroundings"

Don't ask the operator what their SA is. If SA is a critical part of the system, design a suitable test that can accurately measure SA

Measuring Situational Awareness

- **Probes (assess individual's knowledge of situation)**
  - At pre-determined points unknown to user ask pre-defined questions of
    - » Perception (how many neutral targets)
    - » Meaning (system status)
    - » Projection (where will Maverick be in 10 min)
  - Score accuracy of questions & reaction time
  - Often best to freeze scenario for question, but can be embedded in communications
  - Custom surveys – will require system SMEs and Human Factors SMEs to develop

- **Expert observers, like an evaluator pilot**
  - Observer must know ground truth & be trained in SA evaluation
  - Limited in that observer doesn't know the subject's thoughts

5 Levels of Evaluating Training Effectiveness

[Kirkpatrick 1998, Phillips 1998]

1. **Return On Investment**
   – Compare the cost of training with the value of the new mission outcome

2. **Results**
   – How did the training change the mission outcome

3. **Behavior**
   – Analysis of job performance – do the individuals use the knowledge/skills in their job?

4. **Learning**
   – Written or practical test to measure knowledge/skills gained

Good measures of training effectiveness, but not measured using surveys

5. **Reaction**
   –Student's response immediately following the course
   –"The course was well organized." "The training environment was comfortable"
   –Measures satisfaction – not how much was learned

Not a measure of training effectiveness

- **Problem discovery**
  –After the user has accomplished the tasks they were trained for can ask "I felt as if additional training was needed"
  –Not a measure of training effectiveness – but useful to find gaps

Not a measure of training effectiveness, useful in operational test

Utility

- **Utility is how useful the system is to the user**
  - A system must be usable to have utility
  - A system with poor utility may not be used

- **Overlaps some with Usability**
  - First SUS question is a utility question:

    "I think that I would like to use this system to accomplish the mission."
  - If SUS is used on the system of interest, don't need to ask a separate utility question – pull the data for the individual question out of the SUS.

- **Other Utility questions**
  - **"I would take (this system) to war"**
  - **"Are there any improvements that you would make to the system?"**
  - **"Do you have any additional comments about the system?"**

- **An effective system may have poor utility scores**
  - Not tested in proper conditions
  - Not trained properly

Other Empirically Vetted Surveys

- **Stress, Fatigue**
  - Two constructs that affect many other areas, including performance, situational awareness, workload (through resources available)
- **Trust in system**
  - System trust plays an important role – especially as systems become more complex and ‘smart’
  - Not always correlated with actual system performance
    - » High trust in a poor performing system can be dangerous
    - » Low trust in good system loses value
    - » Misplaced trust can lead to errors
- **The list continues…**
  - Many more surveys exist

Many of these areas have not typically been examined in operational testing

Demographic & Diagnostic

- **Custom made surveys**
- **Demographics (self perception)**
  - Important to understanding system use
  - MOS/Rate, training, Age, Role, etc…
- **Diagnostic**
  - System specific questions
  - Ask if you will report regardless of responses (All positive, all negative, or mixed)
  - Examples:
    - » The missile was easy to unpack.
    - » I could easily adjust the radio to an acceptable volume.
    - » Overall, the ship's living spaces are comfortable.

Surveys in Operational Testing

- **Human factors constructs**
  - Usability
  - Workload
  - Others (trust, fatigue,…)
- **System utility**
- **Demographics**
  - Are users in test, similar to the population of users?
  - Are there differences associated with different user demographics?
- **Diagnostic information**
  - Why did performance reach or not reach satisfactory levels?
  - Will there be performance problems in untested conditions?
- **Not Surveys**
  - Performance
  - Situational Awareness
  - Training Effectiveness

Up Next

- **Selecting Empirically Vetted Surveys**
- **Custom-Made Surveys**
- **ABIS Case Study**
- **Administration & Analysis**
- **Air Force DCGS Case Study**