Survey Design and Best Practices in Operational Test & Evaluation

Dr. Chad Bieber
Dr. Justin Mary
Mr. Jonathan Snavely
Dr. Heather Wojton

Operational Test & Evaluation

Goal: Determine the effectiveness and suitability of military systems for use by **military users**.

Military users make military systems function.

We want to understand:

- System capabilities
- Users' experience operating the system

Outline

- **Basics of Human Measurement**
- **Selecting a measurement method**
- **Survey Basics**
- **Types of surveys & how to construct them**
  - Empirical Surveys
  - Custom-Made Surveys
  - Demographics Surveys
- **Survey Administration & Data Collection**
- **Data Analysis**

The Human Factor

Human factors can be measured by physical and survey measures.

Physical Measures
- Physiological Responses
- Behavioral Responses
- Performance Measures

Survey Measures
- Thoughts
- Feelings
- Opinions

Survey Measures

Goal: apply measurement units to thoughts, feelings, and opinions.

Unique Challenges

– Thoughts, feelings, and opinions are context dependent.

https://www.youtube.com/watch?v=FWSxSQsspiQ

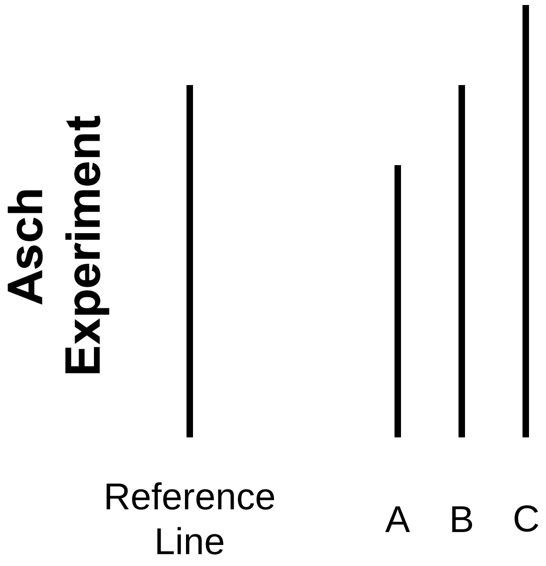
https://www.youtube.com/watch?v=IGQmdoK_ZfY

Survey Measures

Goal: apply measurement units to thoughts, feelings, and opinions.

Unique Challenges

- Thoughts, feelings, and opinions are context dependent.
- Subject to biases like demand characteristics and social desirability.
  - Experimenter biases responses through explicit and implicit cues.
  - Respondents' desire to be viewed positively



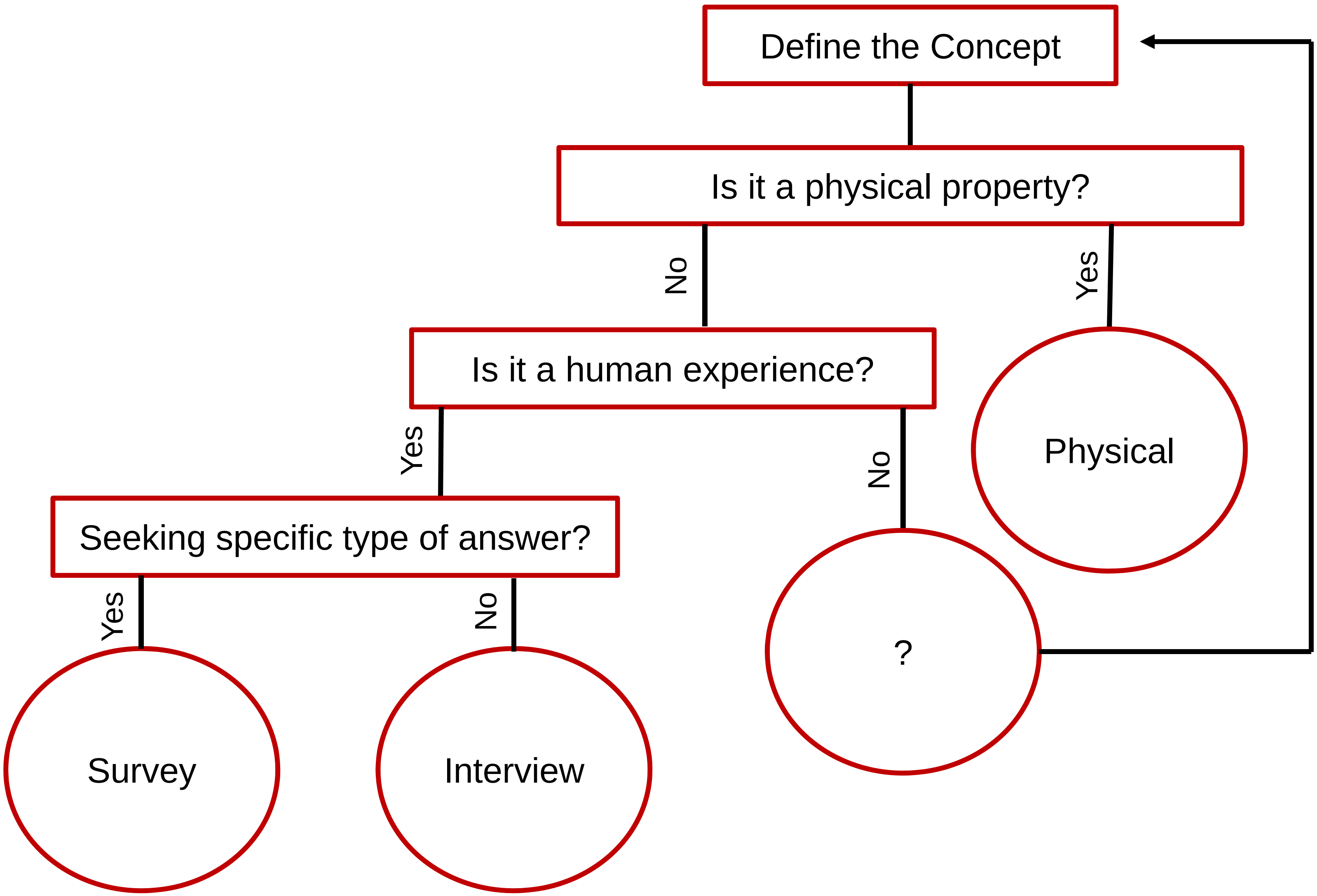
Reduce error by considering these issues when constructing and administering survey measures.

Outline

- **Basics of Human Measurement**
- **Selecting a measurement method**
- **Survey Basics**
- **Types of surveys & how to construct them**
  - Empirical Surveys
  - Custom-Made Surveys
  - Demographics Surveys
- **Survey Administration & Data Collection**
- **Data Analysis**

Selecting a Measurement Method

Measurement Method Selection Process



Define the Concept

Is it a physical property?

No

Yes

Is it a human experience?

Physical

Yes

No

Seeking specific type of answer?

?

Yes

No

Survey

Interview

Define the Concept

- Identify the effectiveness / suitability concepts you want to measure.
- Start with system CONOPS, COIs, and MOEs.

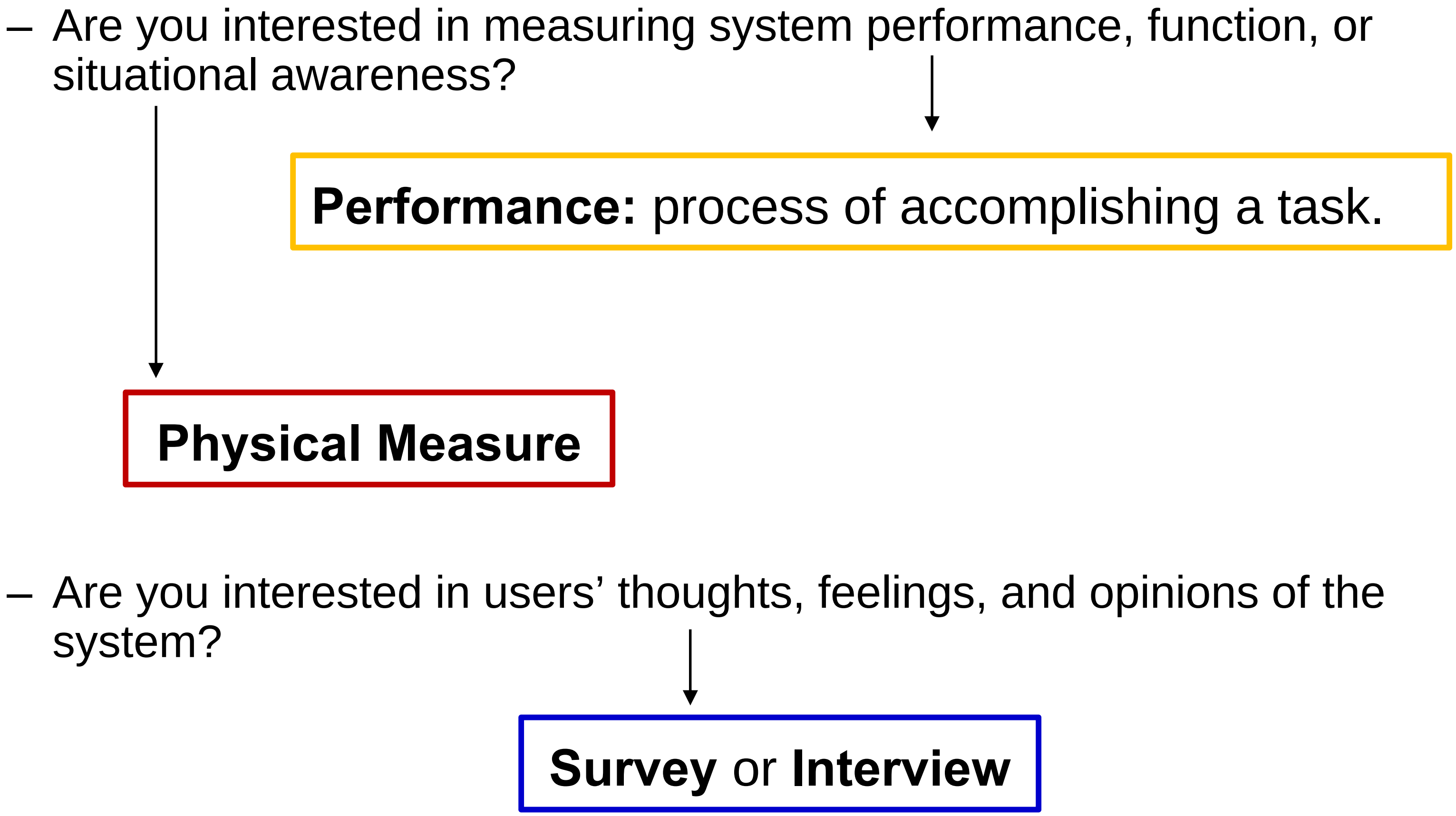
Do NOT simply copy document language into surveys.

- Operationally define each concept.

Translate concepts into **concrete**, **measurable events**.

Define the Concept

- **For each concept:**



  - Are you interested in measuring system performance, function, or situational awareness?

Performance: process of accomplishing a task.

Physical Measure

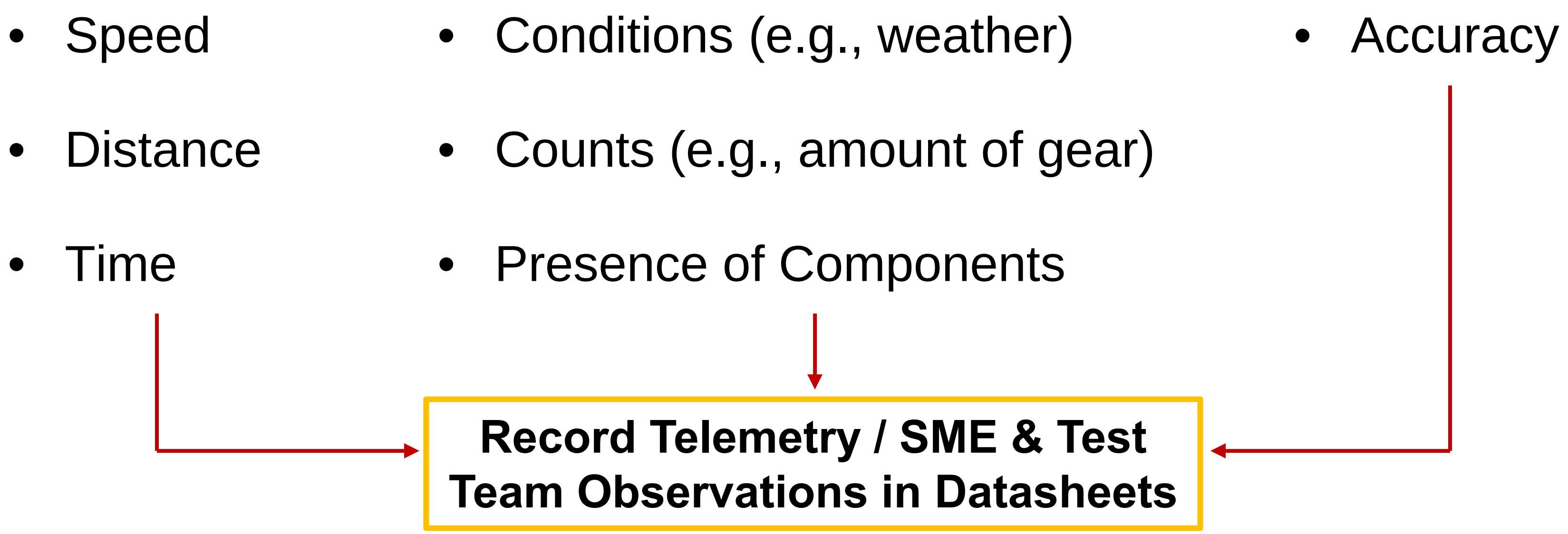
  - Are you interested in users’ thoughts, feelings, and opinions of the system?

Survey or **Interview**

Define the Concept

Be careful! Thoughts, feelings, and opinions can impact performance, but they are NOT measures of performance.

Physical Measures



- Speed
- Distance
- Time

- Conditions (e.g., weather)
- Counts (e.g., amount of gear)
- Presence of Components

- Accuracy

Record Telemetry / SME & Test Team Observations in Datasheets

	Start Time	End Time	Weather	Safety Gear
Session 1				
Session 2				
Session 3				

Surveys & Interviews

Thoughts, feelings, and opinions.

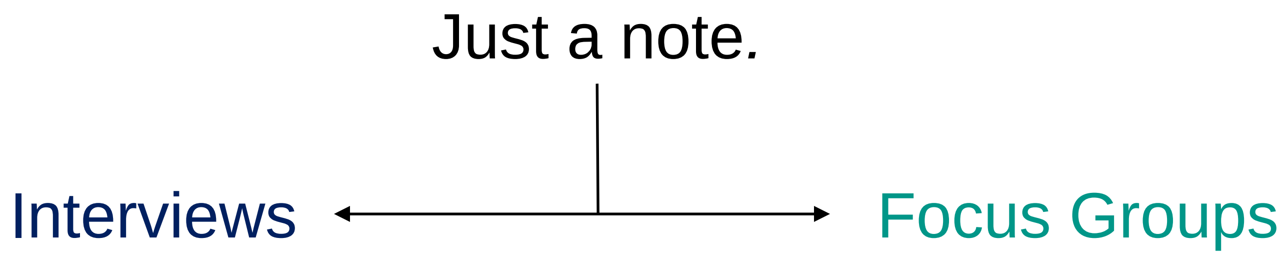
Survey

- Elicit specific information
  - Usability, workload
  - Perceptions of design features
- Finite set of responses
  - Closed response sets
  - Short open-ended response sets
- Planned events
- Quantitative analysis

Interview

- Elicit non-specific information
  - Useful for understanding problems highlighted in survey responses
- Infinite possible responses
- Unplanned events
- Limited analysis options
  - Qualitative analysis
  - Frequencies possible

Interviews & Focus Groups



Just a note.

Interviews

Focus Groups

Interviews

- Elicit non-specific information
  - Useful for understanding problems highlighted in survey responses
- Infinite possible responses
- Unplanned events
- Limited analysis options
  - Qualitative analysis
  - Frequencies possible

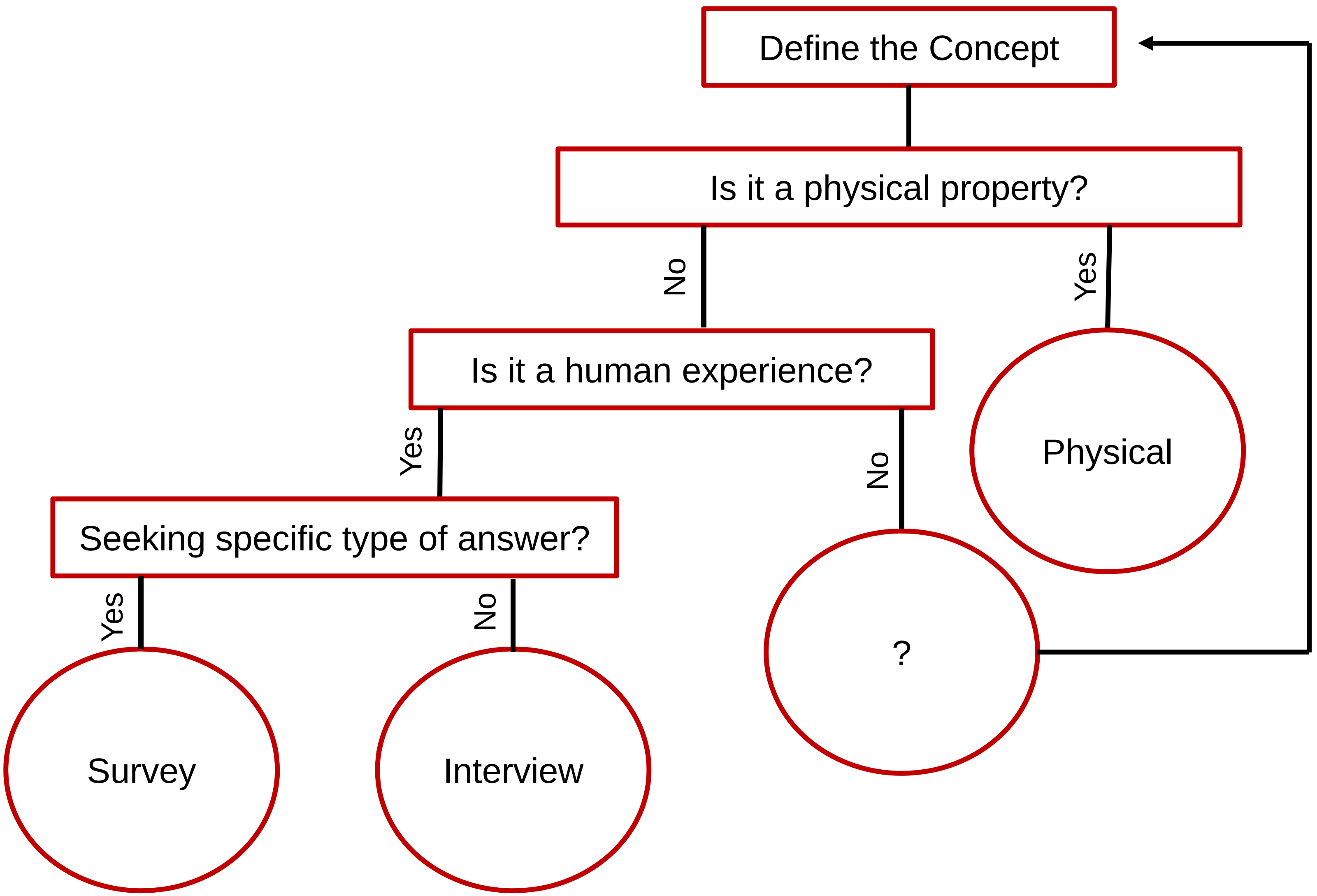
Focus Groups

- Elicit non-specific information
- Group dynamics shape responses
  - May bias responding
- Data obtained at group-level
- Creating solutions
- Useful for generating quotes

DOT&E Guidance

http://www.dote.osd.mil/pub/policies/2015/2-24-15_Discussion_on_the_Use_and_Design_of_Surveys(8944).pdf

Measurement Method Selection Process


Define the Concept
Is it a physical property?
No
Yes
Is it a human experience?
Physical
Yes
No
Seeking specific type of answer?
?
Yes
No
Survey
Interview

Your Turn!

Activity: Part 1

Conceptual Clarity and Measurement Approach

Concepts

Imagine your team must choose which smart phone the DoD will acquire for all of you.

- What capabilities and characteristics should the phone have?
- Write a few criteria you would want to test before recommending the phone.

Conceptual Clarity

Which criteria would you include in an operational test?

- Design/build characteristic?
  (e.g., 4G LTE network compatible, display size)
- Operational performance or characteristic?
  (e.g., Time to load email, portability)
- Human-system interaction?
  (e.g., Operator experience, User opinions)

Identify additional factors for proper testing

- Operator characteristics
- Use scenarios
  - Professional vs. personal capabilities vary?

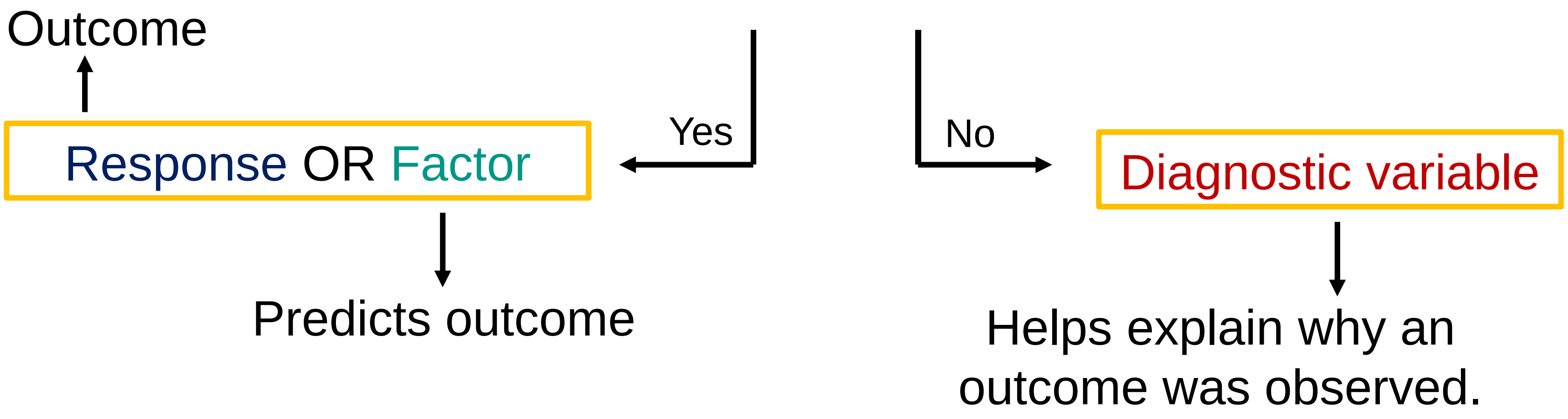
Outline

- Basics of Human Measurement
- Selecting a measurement method
- Survey Basics
- Types of surveys & how to construct them
  - Empirical Surveys
  - Custom-Made Surveys
  - Demographics Surveys
- Survey Administration & Data Collection
- Data Analysis

Survey Basics

Survey Purpose

- A **systematic** measure of people’s thoughts, feelings, and opinions.
  - Apply measurement units to subjective experiences.
- Collect data for a **defined** purpose.
  - The fact that you can’t measure it another way doesn’t mean a survey is appropriate.
- Surveys can be **factors**, **response**, or **diagnostic** variables.
  - Did you design the test around this variable?



Outcome

Yes

No

Response OR Factor

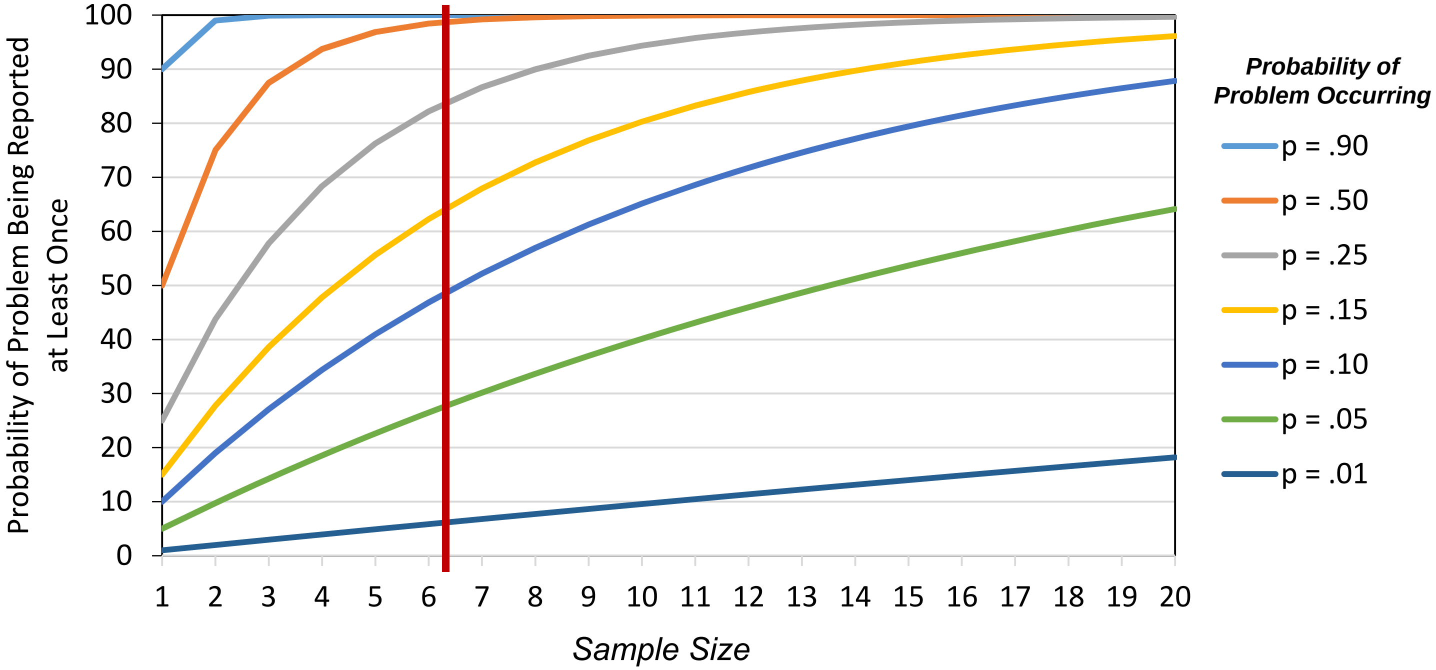
Diagnostic variable

Predicts outcome

Helps explain why an outcome was observed.

Sample Size

- **Most surveys in OT&E are diagnostic variables**
  - No requirement to create a DOE for these surveys
  - Sample size is a minimum of **6 participants per condition**

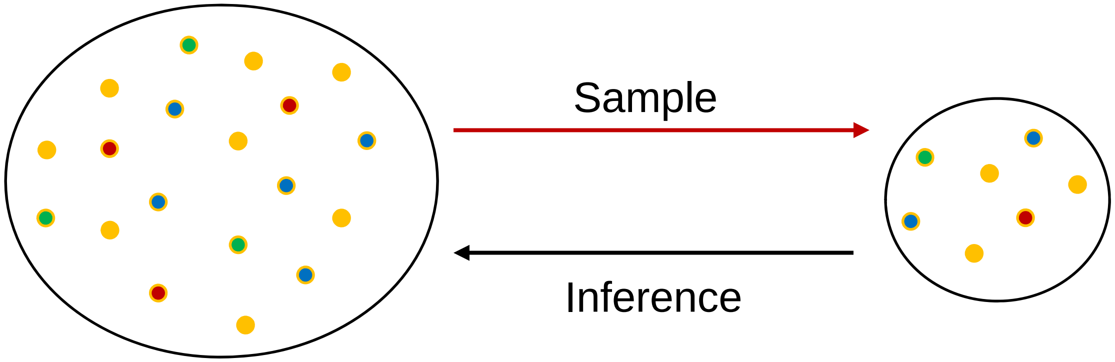


This does **NOT** apply to factors or response variables.

Power analysis & DOE are required for response variables and design factors

Sampling Considerations

Sample: a subset of the population participating in the test



Goal: infer from the sample to the population.

Randomly selecting individuals from the population is ideal.

Random selection is rarely possible in OT&E. In these cases, strive to select a representative sample.

Survey Design

Surveys are comprised of several parts.

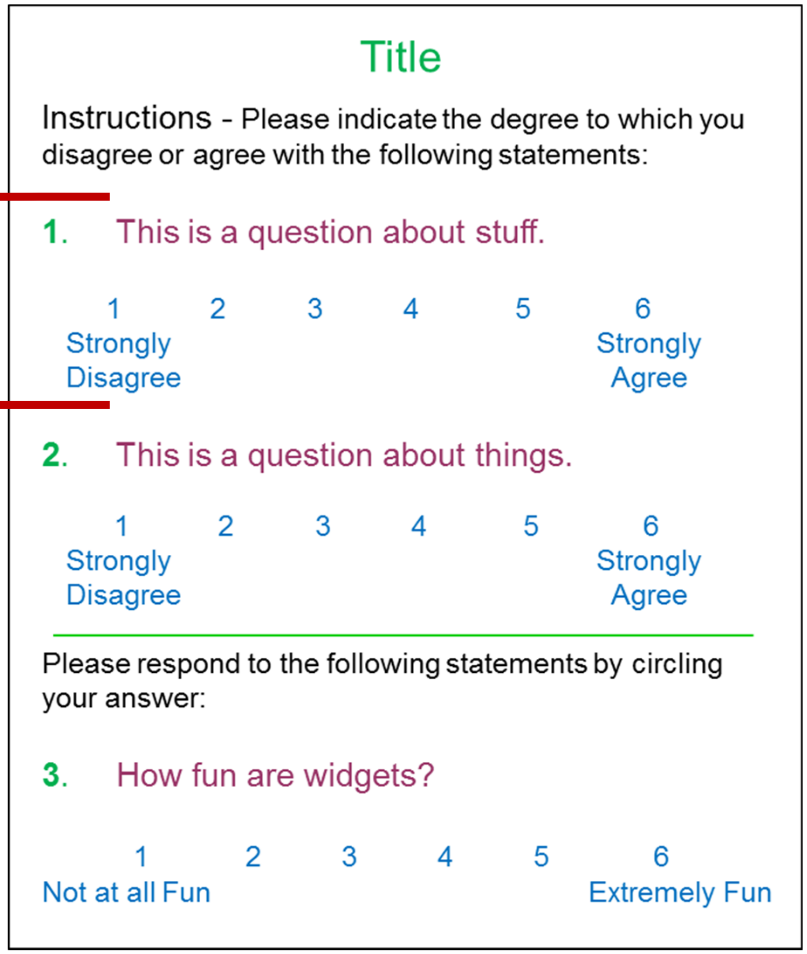
Survey: a collection of questions

Question: item and response option

Item: words a respondent addresses

Response Options: how the respondent provides an answer

Identifier/Formatting: symbols and layout to assist in organizing the survey



Title

Instructions - Please indicate the degree to which you disagree or agree with the following statements:

1. This is a question about stuff.

1	2	3	4	5	6
Strongly Disagree					Strongly Agree

2. This is a question about things.

1	2	3	4	5	6
Strongly Disagree					Strongly Agree

Please respond to the following statements by circling your answer:

3. How fun are widgets?

1	2	3	4	5	6
Not at all Fun					Extremely Fun

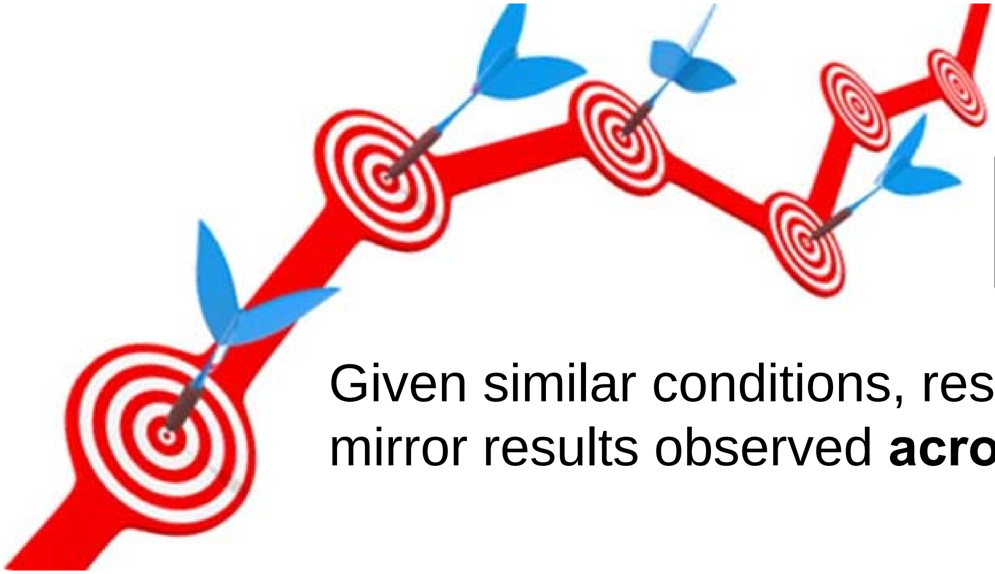
Data Quality

Survey design affects data quality.

- It impacts the **reliability** and **validity** of data collected.

↓

The consistency of a measure.



Reliable ≠ valid.

Given similar conditions, results obtained today should mirror results observed **across time** and **raters**.

Data Quality

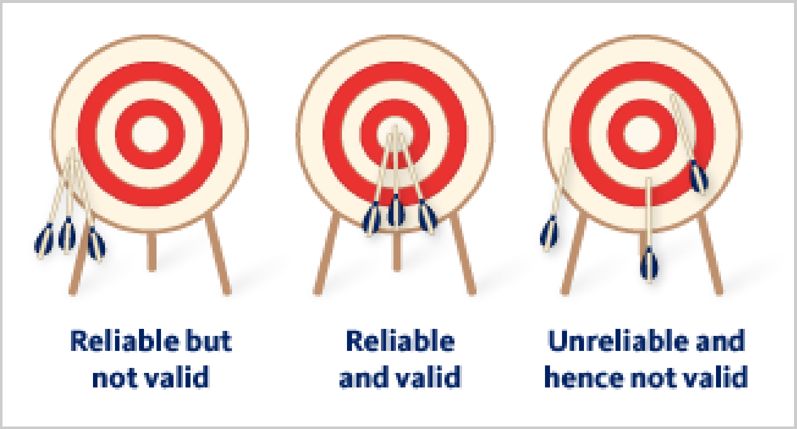
Survey design affects data quality.

- It impacts the **reliability** and **validity** of data collected.

Degree to which the survey is a good measure of the concept it's intended to measure.

- Determined by examining relations **1)** among survey items and **2)** between the survey and measures of related concepts.

Valid measures are reliable.



Data Quality

Survey design affects data quality.

- It impacts the **reliability** and **validity** of data collected.
- Limited resources can make reliability and validity testing unrealistic.
- Surveys that undergo testing are differentiated from those that don’t.

↓

Empirically-Vetted vs. Custom-Made Surveys

OT&E Concepts Measured by Surveys

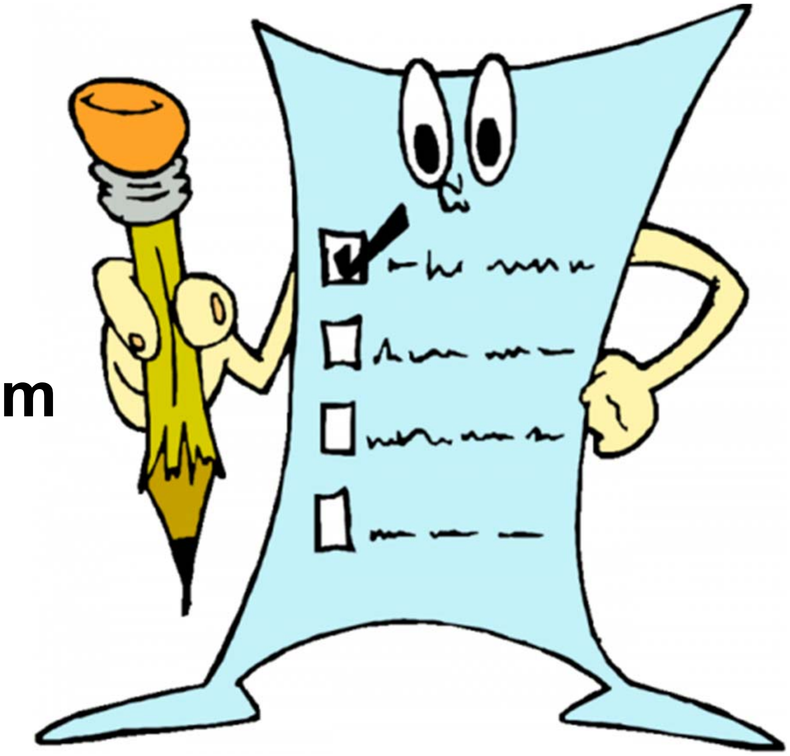
Many OT&E concepts are appropriate to measure via survey.

- Usability
- Workload
- Trust
- Training
- Safety
- Stress/Fatigue
- Utility
- Efficacy

Outline

- **Measurement Basics**
- **Selecting a measurement method**
- **Survey Basics**
- **Types of surveys & how to construct them**
  - Empirical Surveys
  - Custom-Made Surveys
  - Demographics Surveys
- **Survey Administration & Data Collection**
- **Data Analysis**

Survey Types & How to Construct Them



Survey Types

Surveys differ by level of reliability and validity testing.

- Empirically-Vetted Surveys: undergone reliability and validity testing.

Most rigorous type of survey measure

- Custom-Made Surveys: not undergone reliability and validity testing.

Empirically-Vetted Surveys

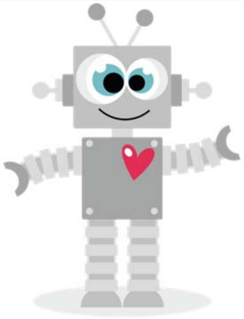
Strive to use empirically-vetted surveys.

- The survey’s reliability and validity are known.
- Effect sizes and variances are available to aid in power analyses.
- Average scores can be used as a standard for comparison.

Currently, empirical surveys available for **workload**, **usability**, **trust**, **fatigue**, and **stress**.

NASA-TLX

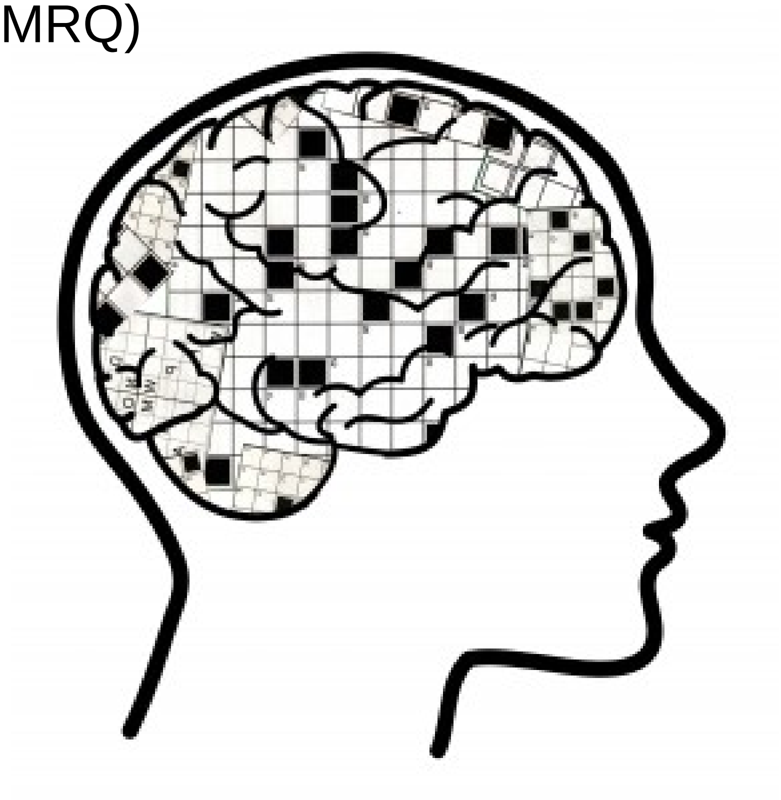
SUS



Workload

Task demand vs. available resources

- NASA Task Load Index (NASA-TLX)
- Multiple Resource Questionnaire (MRQ)
- Crew Status Survey (CSS)



The NASA Task Load Index (TLX)

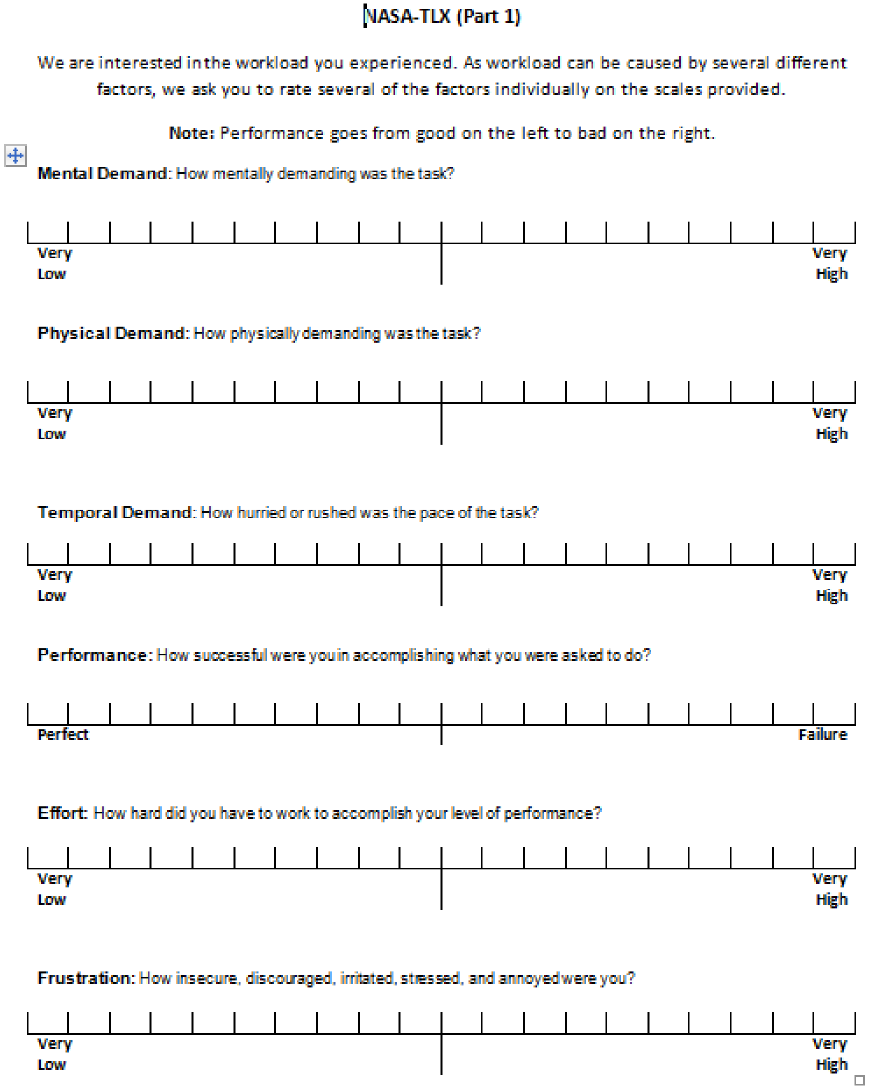
1. Rate workload along 6 dimensions
   - Mental, physical, and temporal demand
   - Perceived performance, effort, and frustration

Its okay to use Part 1 only!

2. Select dimensions that contributed most to workload
   - 15 paired comparisons

Score: Mean workload ratings weighted by paired comparisons

How each dimension contributes to overall workload during the task.



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?

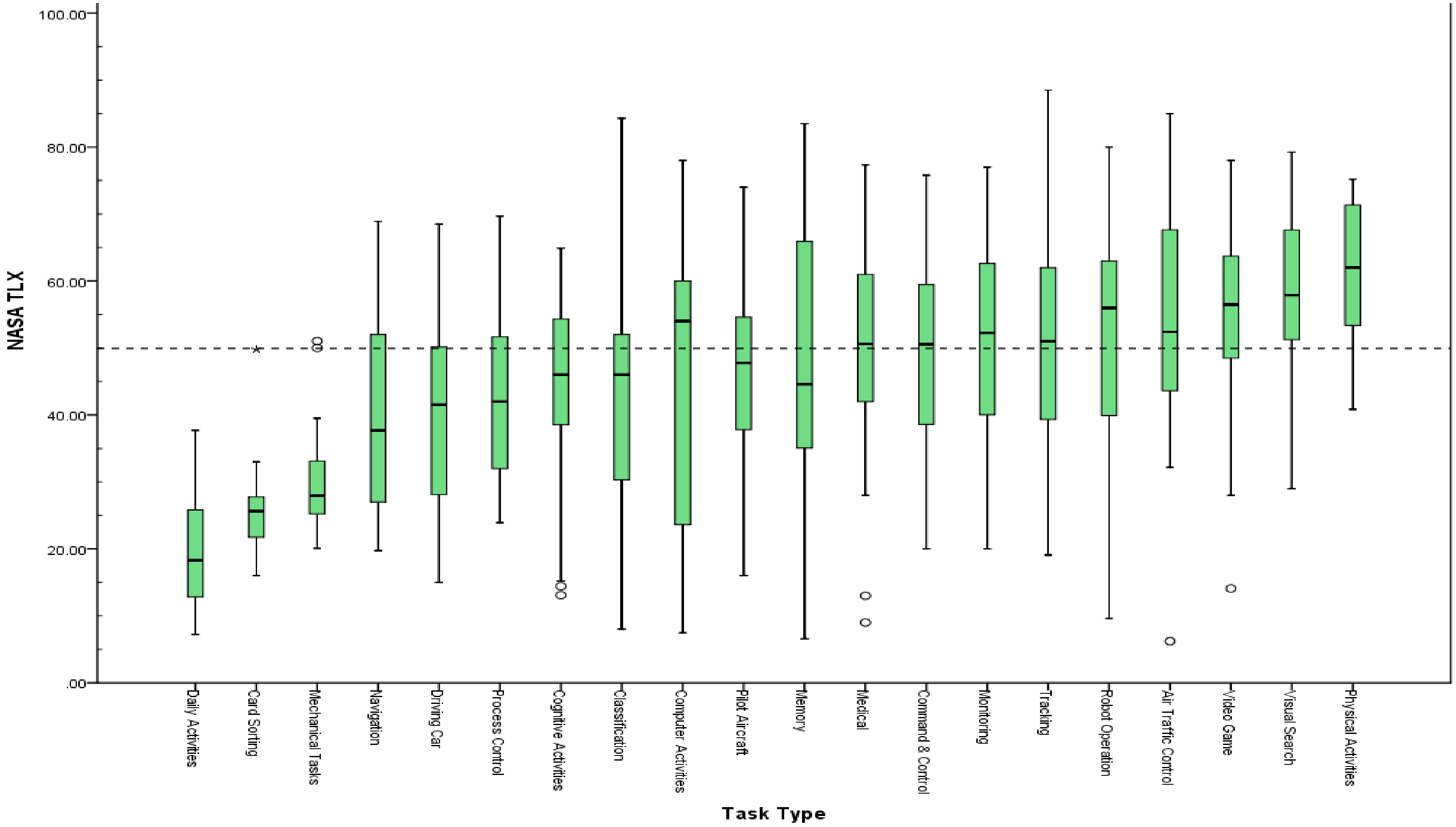
Very Low — Very High

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

Very Low — Very High

NASA-TLX Comparison Standard

Workload ranges separated by task area

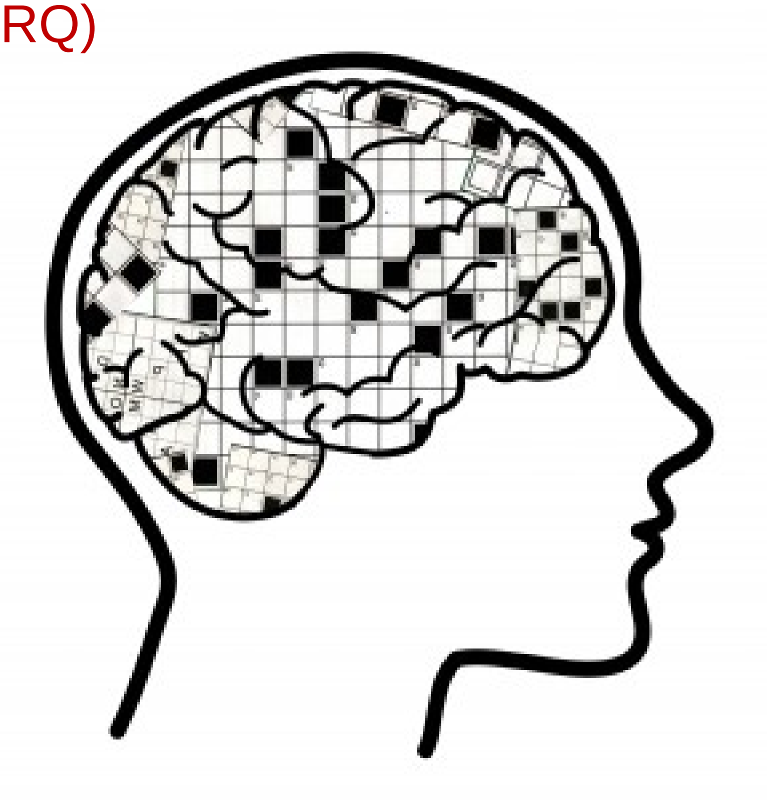


(**Source:** Grier, 2014)

Workload

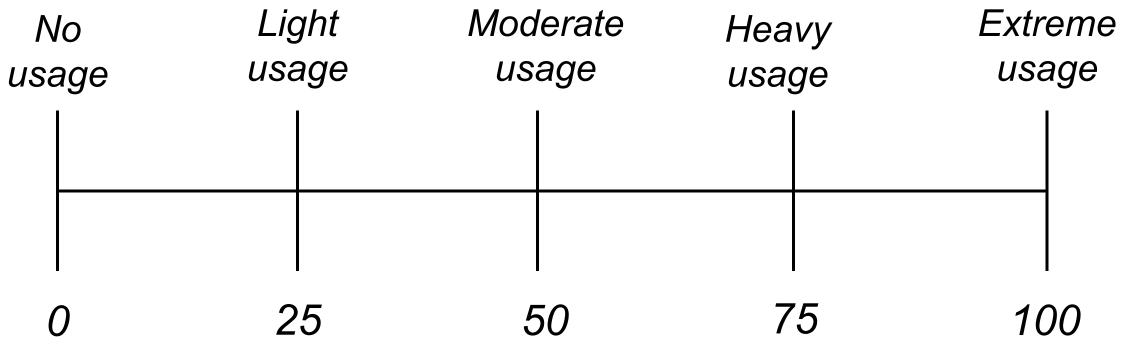
Task demand vs. available resources

- NASA Task Load Index (NASA-TLX)
- Multiple Resource Questionnaire (MRQ)
- Crew Status Survey (CSS)



Multiple Resources Questionnaire (MRQ)

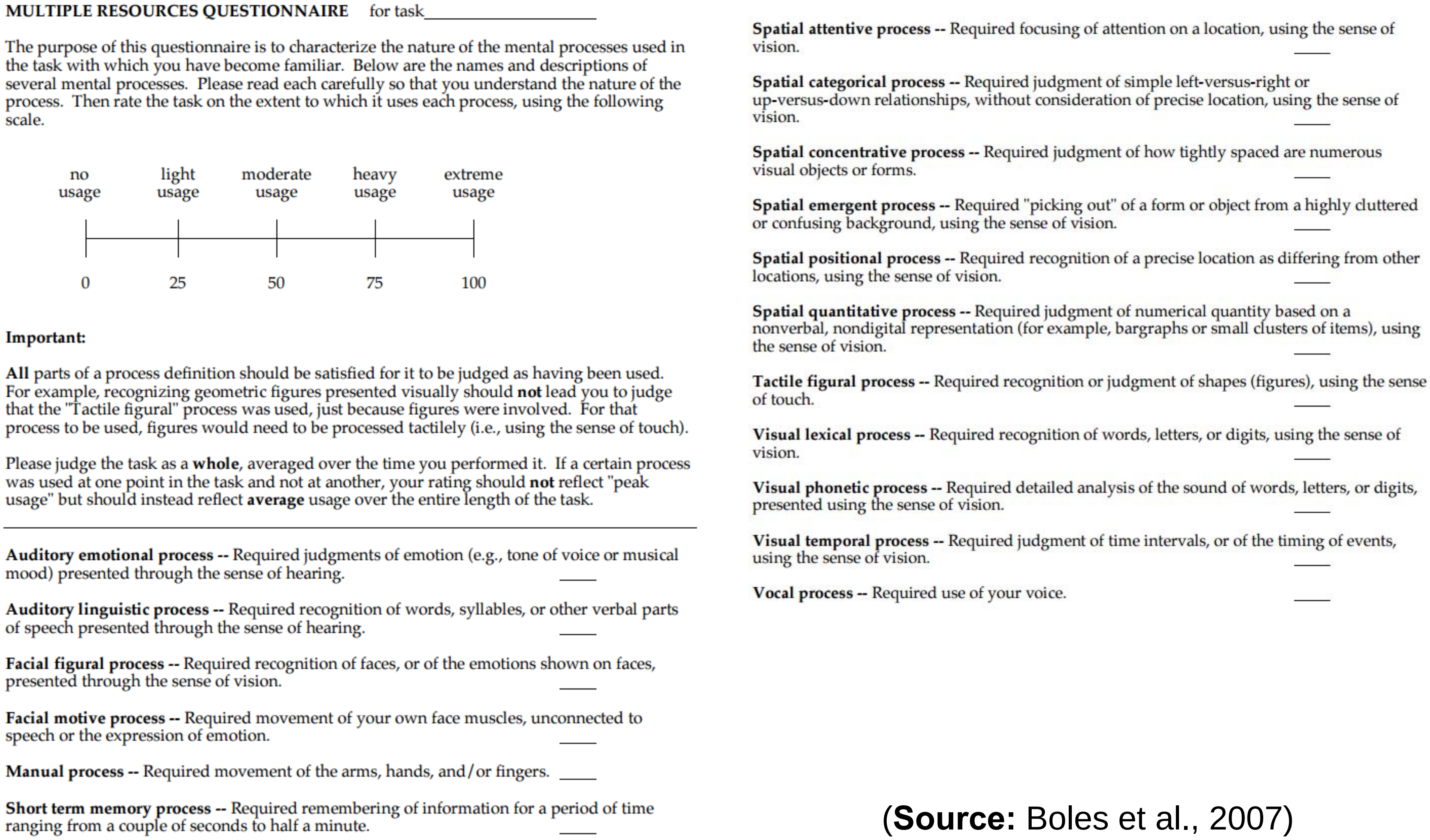
Rate extent to which 17 mental processes were used during a task.



Score: sum across the 17 items

Provides overall workload score and is capable of identifying the mental processes that contribute most to this score.

Multiple Resource Questionnaire (MRQ)



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.

no usage	light usage	moderate usage	heavy usage	extreme usage
0	25	50	75	100

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. ____

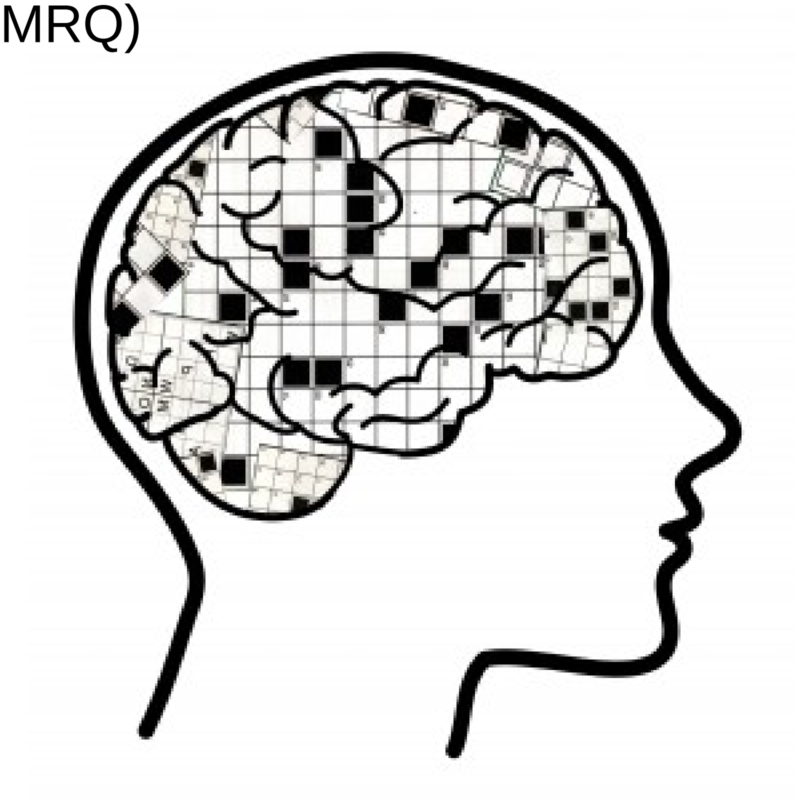
Vocal process -- Required use of your voice. ____

(**Source:** Boles et al., 2007)

Workload

Task demand vs. available resources

- NASA Task Load Index (NASA-TLX)
- Multiple Resource Questionnaire (MRQ)
- Crew Status Survey (CSS)



Crew Status Survey (CSS)

Single item measure of workload during a task

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.

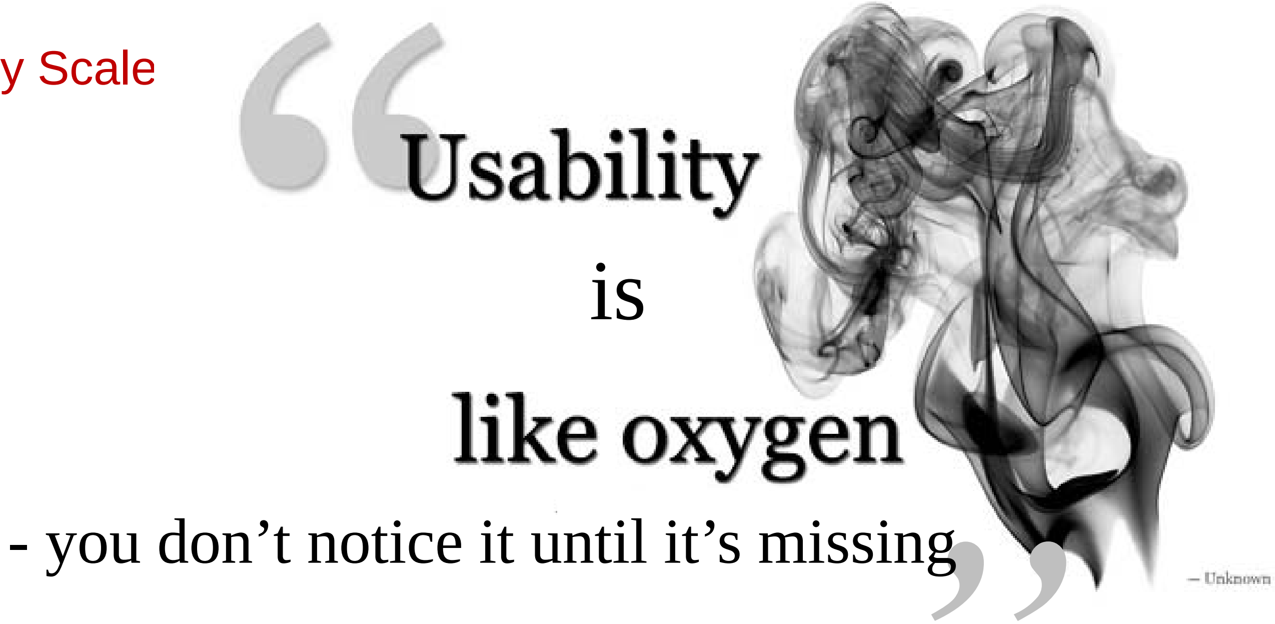
Ideal for giving at regular intervals throughout task because it’s short and produces minimal interference.

Usability

The **effectiveness**, **efficiency** and **satisfaction** with which specified *users* achieve specified *goals* in particular *environments*.

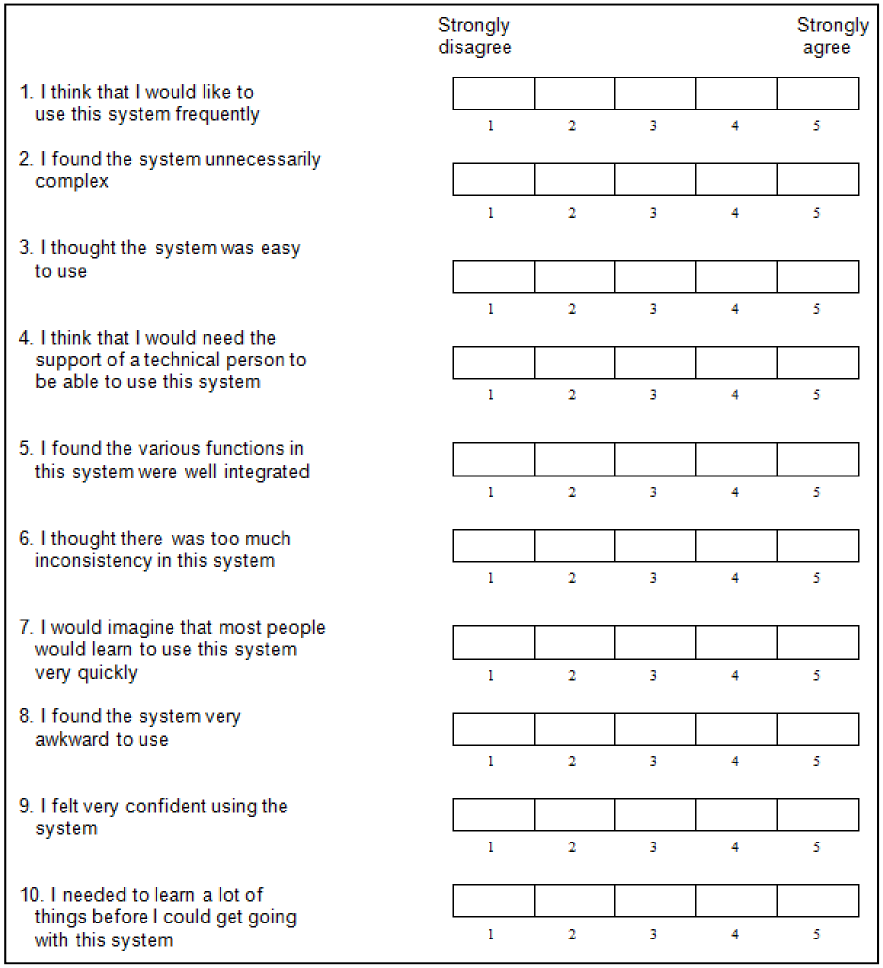
(ISO, https://www.w3.org/2002/Talks/0104-usabilityprocess/slide3-0.html)

- System Usability Scale



System Usability Scale (SUS)

10-item measure of system usability.



	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

1. Some items reverse scored.
2. Normalize scores.

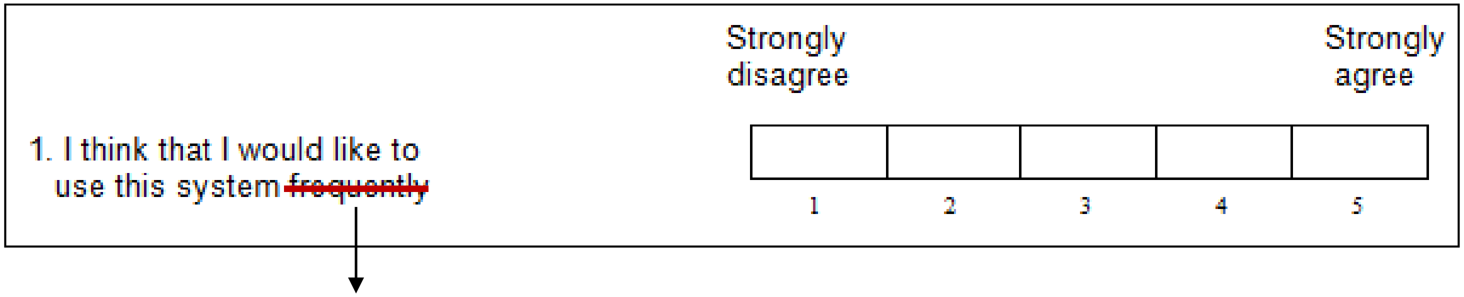
↓

Scoring Guide Available

Produces a score from 0-100

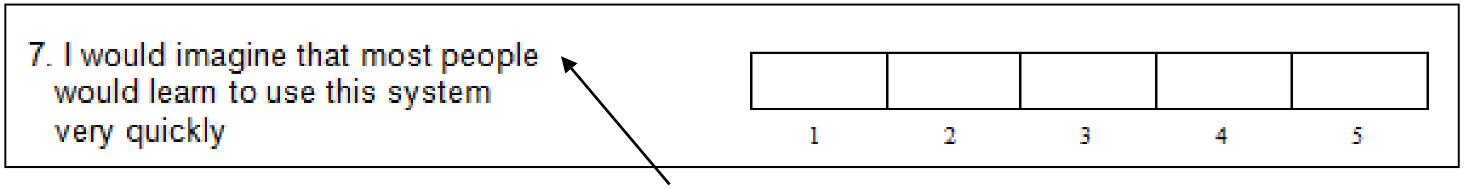
System Usability Scale (SUS)

10-item measure of system usability.



	Strongly disagree				Strongly agree
1. I think that I would like to use this system ~~frequently~~	□	□	□	□	□
	1	2	3	4	5

to accomplish the mission

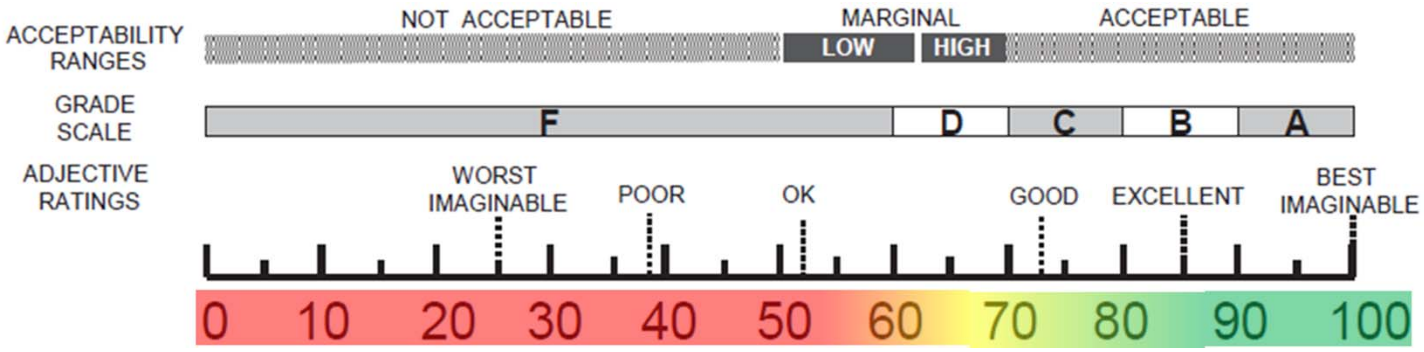


7. I would imagine that most people would learn to use this system very quickly	□	□	□	□	□
	1	2	3	4	5

in my current position

SUS Comparison Standard

Average usability score of 70.



Bangor, A., Kortum, P.T., & Miller, J.T. (2009) Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. Journal of Usability Studies, 4, 114-123.

Empirical Surveys

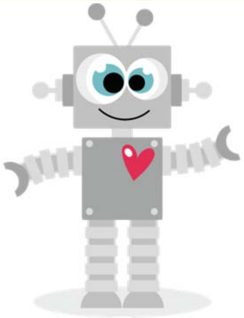
Strive to use empirical surveys.

- The survey's reliability and validity are known.
- Effect sizes and variances are available to aid in power analyses.
- Average scores can be used as a standard for comparison.

Currently, empirical surveys available for **workload**, **usability**, **trust**, **fatigue**, and **stress**.

NASA-TLX

SUS



Your Turn!

Activity: Part 2

Selecting Empirically-Vetted Surveys

Selecting Empirically-Vetted Surveys

Can the concepts you identified be measured by empirically-vetted surveys?

- Do the concepts relate to **workload** or **usability**?
  - If so, there is likely an empirically-vetted survey to fit test needs.
- If appropriate, select an instrument from the list below to measure your concept.

Workload

- NASA Task Load Index (NASA-TLX)
- Multiple Resource Questionnaire (MRQ)
- Crew Status Survey (CSS)
- Modified Cooper-Harper

Usability

- System Usability Scale (SUS)

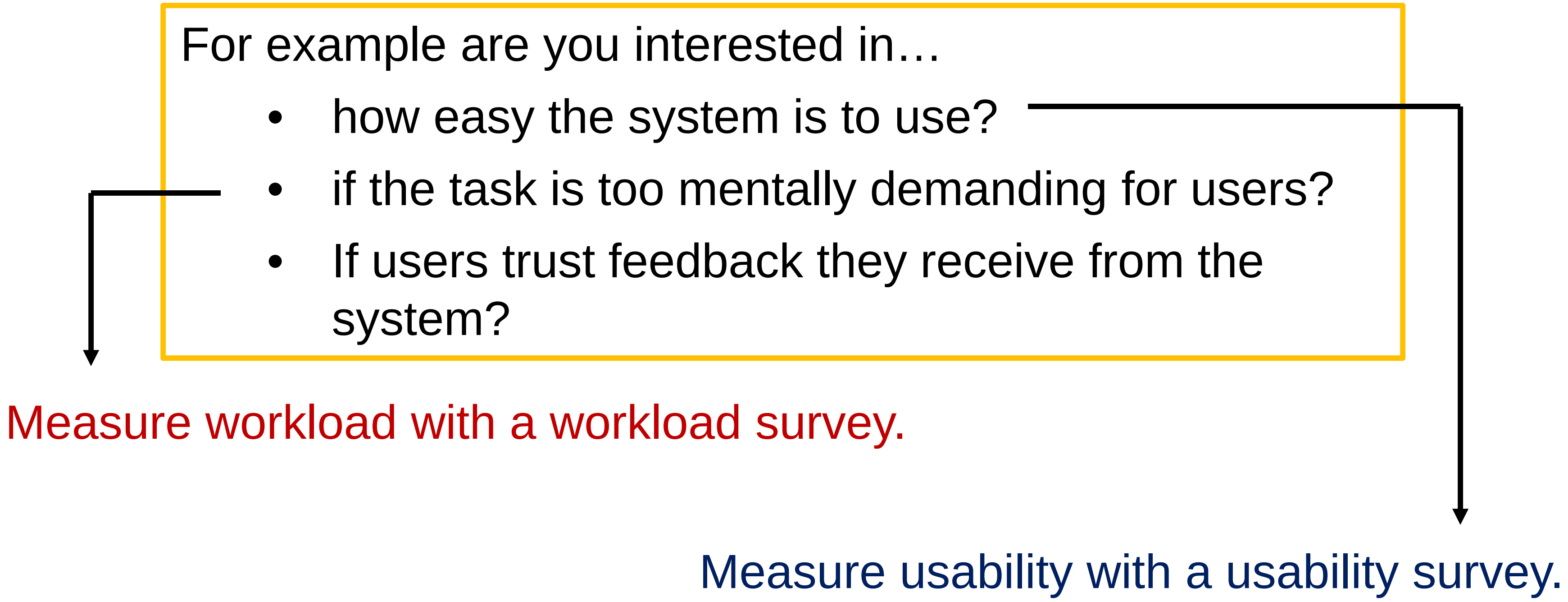
Empirically-vetted surveys are constantly being developed. Do your research before concluding that there isn't one available.

Case Study:

Selecting Empirically Vetted Surveys

Selecting a Measurement Method

- **Define the concept**
  - What do you want to measure?



For example are you interested in…

- how easy the system is to use?
- if the task is too mentally demanding for users?
- If users trust feedback they receive from the system?

Measure workload with a workload survey.

Measure usability with a usability survey.

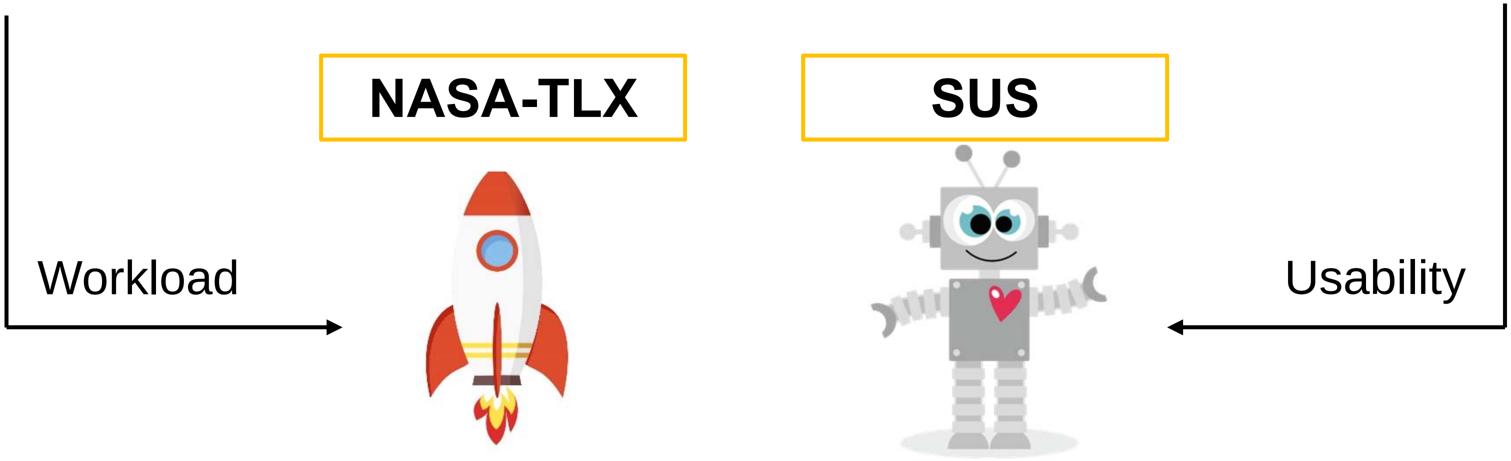
Selecting a Measurement Method

- **Define the concept**
  - What do you want to measure?
- **Define the purpose of the measure**
  - Are you trying to identify a problem with the system? (diagnostic)
  - Will you use it to predict an outcome? (factor)
  - Is it the outcome of interest? (response)
- **How will the data be analyzed?**
  - Different response types support calculation of different statistics.
  - How small is the effect you want to detect?
    - » Some surveys are sensitive to larger/smaller effects.

Do **NOT** select a measurement method until you can clearly describe what is being measured.

Selecting a Measurement Method

- Select the most rigorous method for the concept you want measure and the expected analysis



Workload

NASA-TLX

SUS

Usability

Measure different aspects of each concept

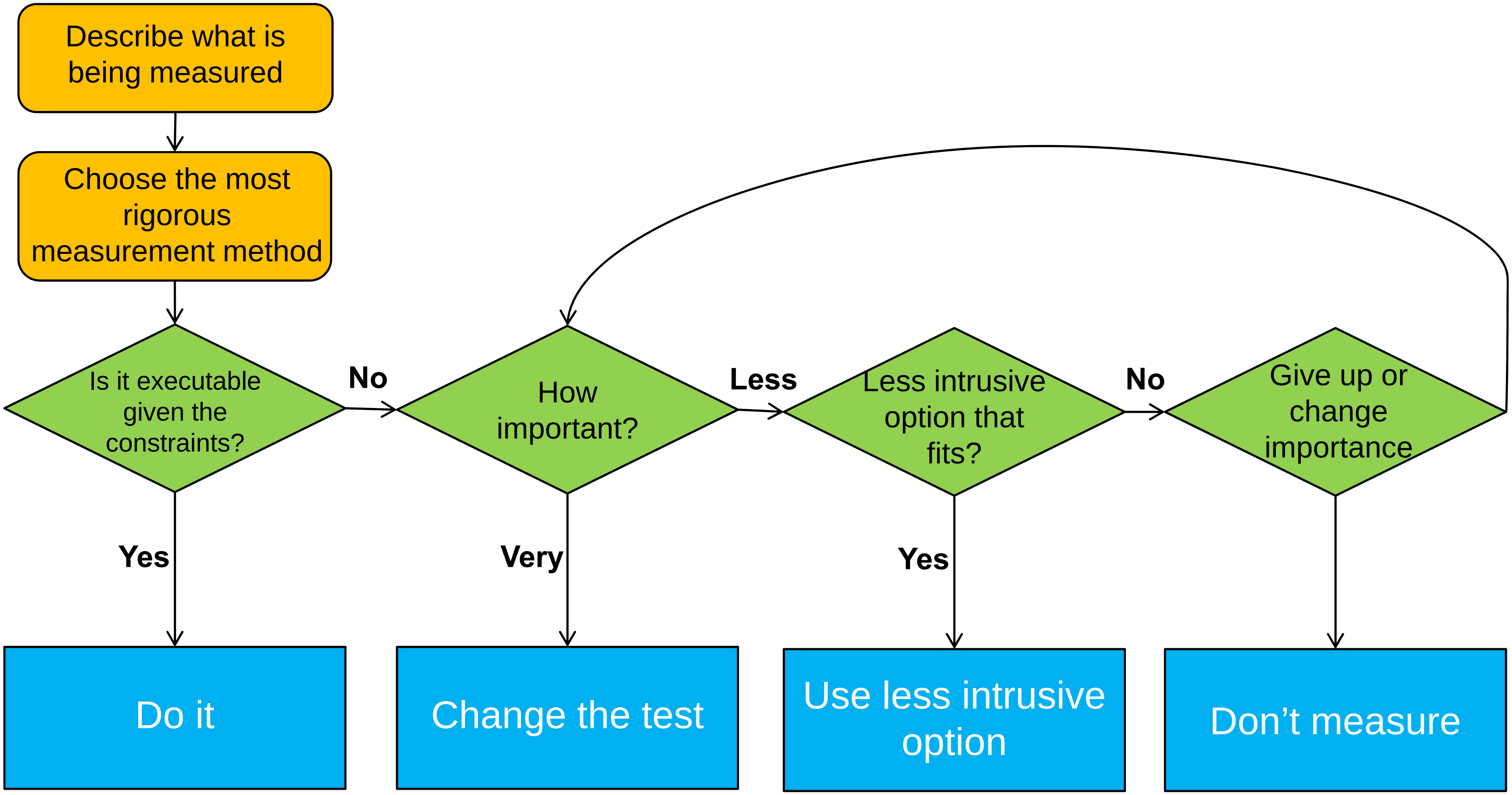
Selecting a Measurement Method

- Select the most rigorous method for the concept you want measure and the expected analysis.

- Identify constraints of the test, method, and environment.
  - Time available (NASA-TLX takes 1 to 3 minutes)
  - Physical limitations (Is it safe for the pilot to take a 3 minute survey?)
  - How often the survey will be given (Once? More than once?)

- How important is the concept being measured?
  - If the concept is a response variable or factor, use the most rigorous method. Alter test design to overcome limitations.
  - If the concept is a secondary or minor part of the test, use a less rigorous method in the face of constraints.

Decision Flowchart


Describe what is being measured
Choose the most rigorous measurement method
Is it executable given the constraints?
No
How important?
Less
Less intrusive option that fits?
No
Give up or change importance
Yes
Very
Yes
Do it
Change the test
Use less intrusive option
Don’t measure

KC-46A Workload Example

- **New Aerial Refueling Operator Station**
  - Aerial Refueling Operator (ARO) views aircraft being refueled through 3-D video screens rather than a window
  - Want to understand ARO workload in this environment
- **Choosing a method**
  - Describe what is being measured

What: Workload during specific tasks in a multi-hour mission

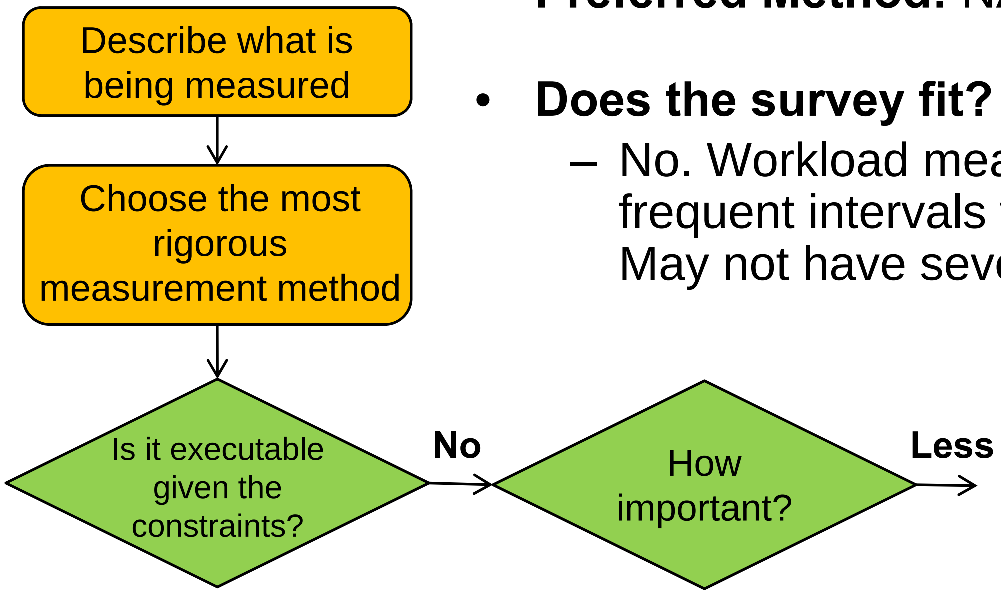
Why: To support a workload Measurement of Effectiveness (MOE)

How: Compare factors – operational conditions (e.g., day/night), different receiver aircraft being refueled.

  - Most rigorous method: NASA-TLX

KC-46A Workload Example

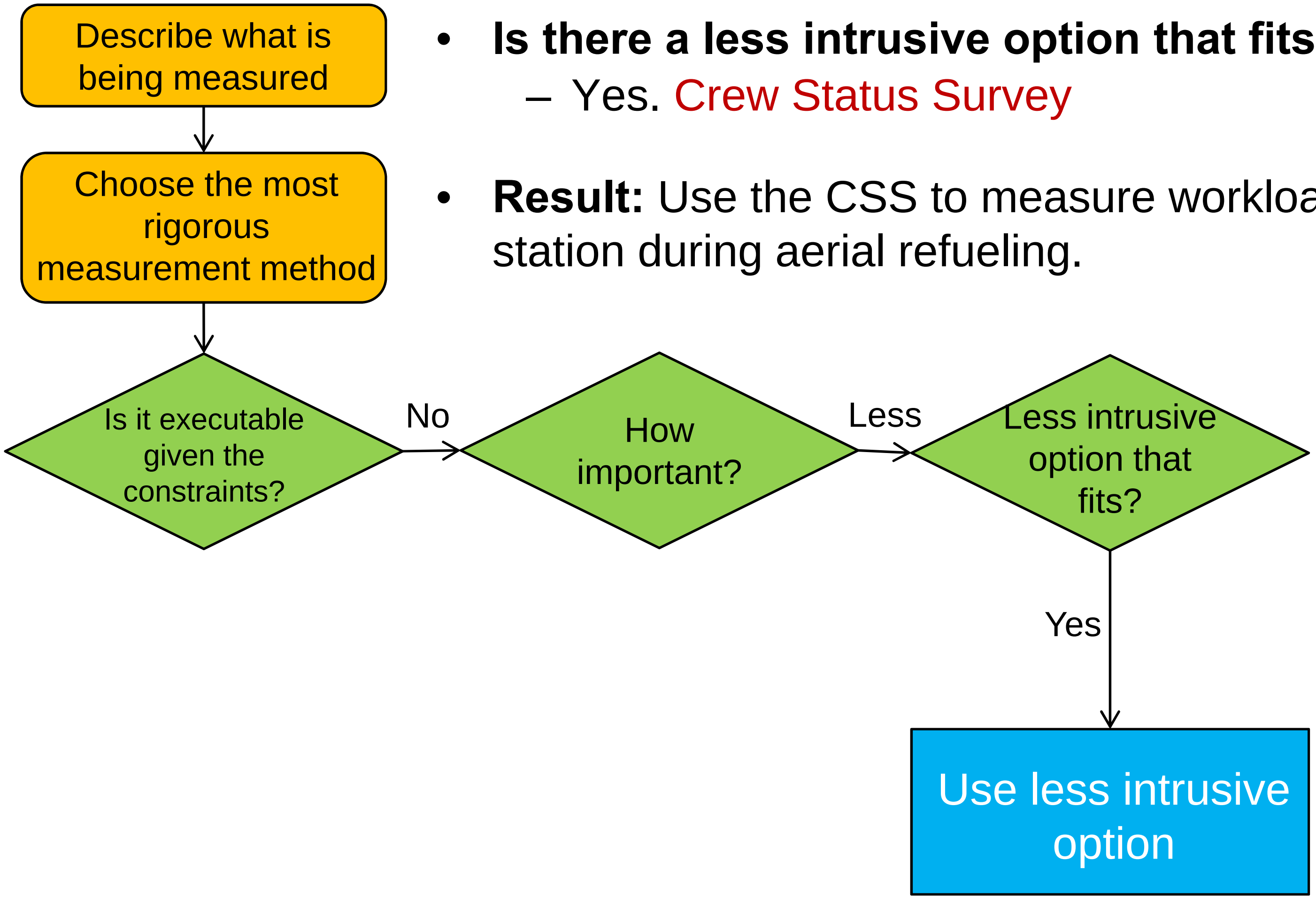
- **Measure:** Workload
- **Preferred Method:** NASA-TLX
- **Does the survey fit?**
  - No. Workload measurements will be taken at frequent intervals while receivers are waiting. May not have several minutes between tasks.


Describe what is being measured
Choose the most rigorous measurement method
Is it executable given the constraints?
No
How important?
Less

- **Is the measurement important enough to change the test?**
  - No. Workload is important, but not a primary response variable

KC-46A Workload Example

- **Is there a less intrusive option that fits?**
  - – Yes. Crew Status Survey
- **Result:** Use the CSS to measure workload at the ARO station during aerial refueling.


Describe what is being measured
Choose the most rigorous measurement method
Is it executable given the constraints?
No
How important?
Less
Less intrusive option that fits?
Yes
Use less intrusive option

KC-46A Workload Analysis

- **What can we say with data from CSS?**
  - Change in scores across test indicate changes in workload.
  - Can identify high vs. low workload scenarios.
  - Results will be analyzed with respect to Performance
    - Does user experience conflict with reality – for instance, low workload with low performance
    - Support performance results with human responses
  - Comments analyzed for problem identification
  - Can't make general comparisons– no current research supports known workload benchmarks in CSS results.

AH-64E Apache Workload Example

- **Lot 4 AH-64E Apache Attack Helicopter FOT&E**
  - Systems were upgraded, to include Link 16, upgraded sensors, and new video transfer capability
    - Expected outcome: improved Joint operations and mission effectiveness
    - Experiment: time to find first target during a mission
    - Workload was measured in conjunction with this primary metric

- **Choosing a method**
  - Describe what is being measured

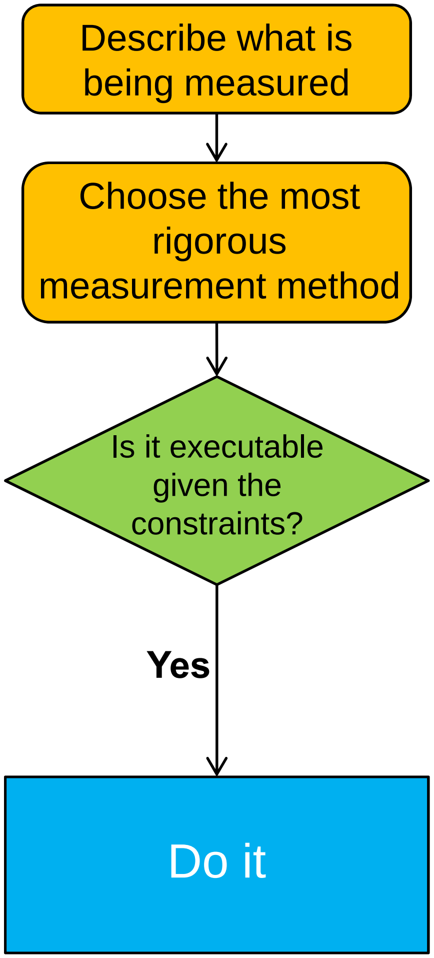
What: Workload over the entire mission

Why: To support a primary response variable

How: Compare workload in different missions

  - Most rigorous method: NASA-TLX

Apache Workload Example


Describe what is being measured
Choose the most rigorous measurement method
Is it executable given the constraints?
Yes
Do it

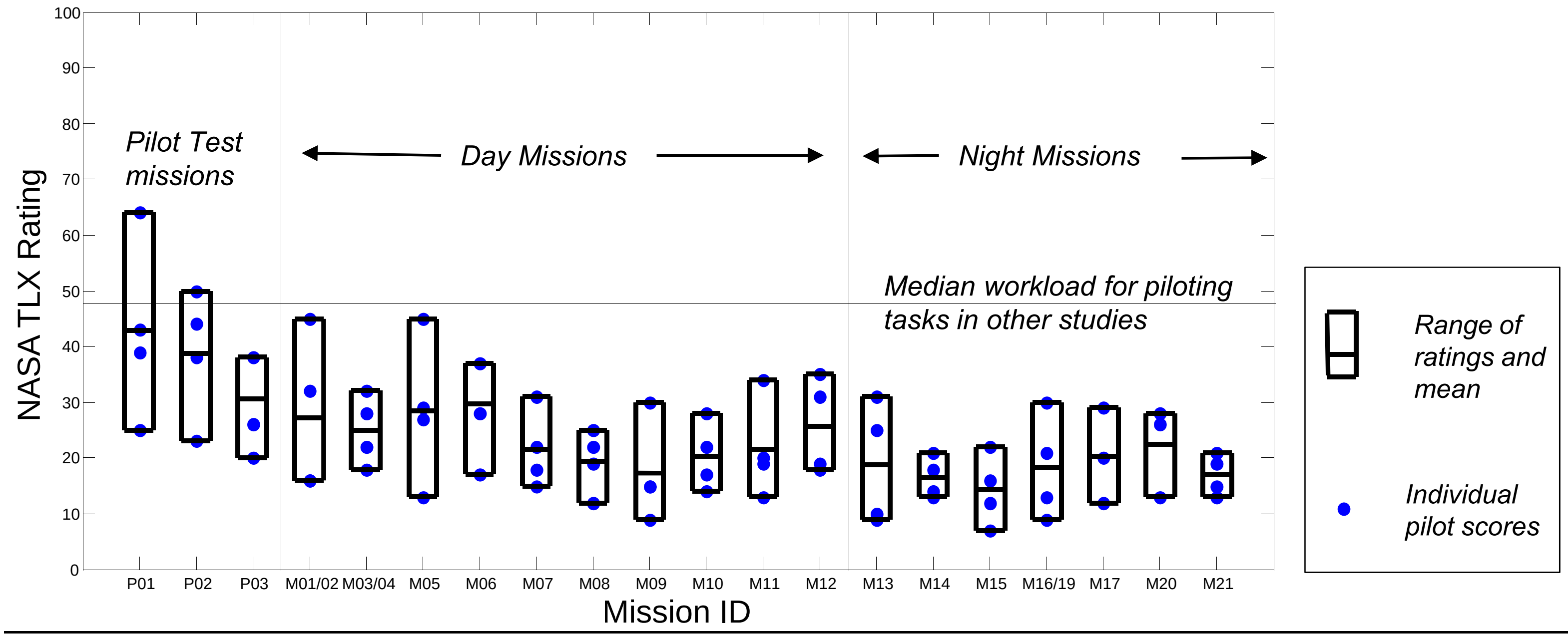
- **Measure:** Workload
- **Preferred Method:** NASA-TLX
- **Does it fit?**
  - Yes! 3 minutes of time available after mission, before debrief
- **Do it!**

Apache Workload Analysis

- **NASA-TLX survey administered after each mission**
- **Four Factors chosen for primary metric (time to find first target)**
  - Link 16 Targeting Data (yes or no), Battlefield Density (high or low), Light Level (day or night), Pilot Seat Location (front or back)
- **Analysis shows several significant correlations**
  - High Density resulted in higher workload with Link16 (*p* = 0.02)
  - Front seat pilot had higher workload with Link 16 (*p* = 0.10)
  - Night missions were significantly lower workload than day, but all day missions were accomplished first, then night missions. Unclear if results were due to time (experience) or to light level

*** 80 % confidence, 10% significance**

Terms	p-value
Link 16 Targeting Data	0.22
Battlefield Density	0.76
Light Level	**0.001**
Pilot Seat Location	0.16
Targeting Data*Battlefield Density	**0.02**
Targeting Info*Light Level	0.73
Targeting Data*Pilot Location	**0.10**
Battlefield Density*Light Level	0.64
Battlefield Density*Pilot Location	0.39
Light Level*Pilot Location	0.33

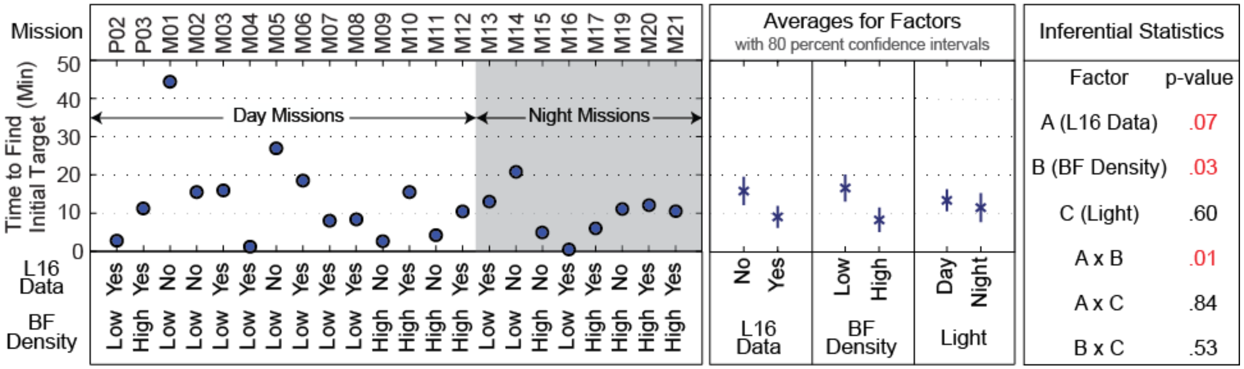


Apache Workload vs. Performance

- **Workload differences were found – what do they mean about the mission?**

- **Primary metric – time to find first target**
  - Key finding – Link 16 improved time for low density battlefield (*p* = 0.01)
  - When battlefield density was high – many targets were present – time to find first target was shorter (*p* = .03) whether or not Link 16 was available

- **What does this mean?**
  - Higher effectiveness with Link 16 and low density– no increase in workload
    - » Clear benefit!
  - Higher workload and similar effectiveness with Link 16 and dense battlefield
    - » Correlation, not causation, but potential information for developing TTPs or further testing



Inferential Statistics

Factor	p-value
A (L16 Data)	.07
B (BF Density)	.03
C (Light)	.60
A x B	.01
A x C	.84
B x C	.53

KC-46A Usability Example

- **KC-46A –Air Refueling Operator Station**
  - Refueling Boom controls and system interface significantly changed from previous designs
  - Expected outcome: improved capability (video feed, IR)

- **Choosing a method**
  - Describe what is being measured

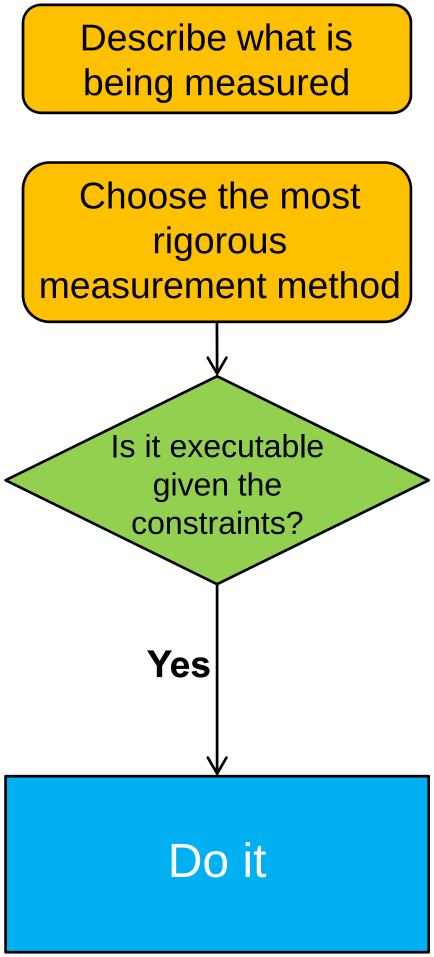
What: Usability of Air Refueling Operator Station

Why: To support “User rating” MOEs

How: General comparison to usability benchmarks, identify problems

- Most rigorous method: SUS (+ open-ended responses for problem ID)

KC-46A Usability Example


Describe what is being measured
Choose the most rigorous measurement method
Is it executable given the constraints?
Yes
Do it

- **What is being measured?**
  - – Usability
- **First choice: SUS + open-ended comment, several times throughout test**
  - – Shows effect of experience
  - – Comparative ability
  - – Problem ID via open-ended comment
- **Does it fit?**
  - – Yes! 3 minutes are available at periodic times throughout test period
- **Do it!**

KC-46A Usability Analysis

- **How will SUS scores be used?**
  - Scores will be compared to established standards.
  - Change in scores indicates changes in usability.
  - Change in usability scores by demographic characteristics.
  - Results will be compared with Performance
    - » Can identify conflicts in perception and help interpret performance results
  - Comments analyzed for problem identification

What if there is no empirically-vetted survey available?

When to Design a Survey

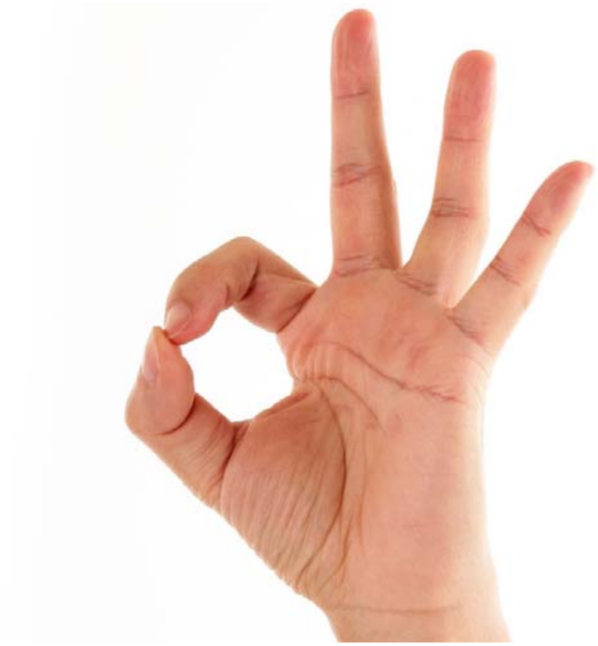
Appropriate

1. **There is not an appropriate empirical survey.**

2. **You are assessing system-specific issues.**

3. **You are measuring thoughts, feelings, and opinions about:**
   - System features and components.
   - Issues related to CONOPS.
   - System experience.

4. **You want to quantify observer ratings.**
   - SME surveys



When NOT to Design a Survey

Inappropriate

1. Performance
   (For instance, accuracy and timeliness)

2. Situation Awareness
   (Situational Awareness Global Assessment Technique)

Use more direct measures of performance and requirements
(e.g., physical indicators, SME or test-team observation)

DOT&E Guidance

http://www.dote.osd.mil/pub/policies/2014/6-23-14_Guidance_on_Use_and_Design_of_Surveys_in_OTE(7985).pdf

Observation vs. User Rating

- **Should you survey users?**
  - Limit survey burden on users by identifying alternative observers.
  - Distinguish **existence** of features/components/processes from **experiences** with them.

Design issues observable by SME or test-team outside of test.

Example

- “The icons were visible against the desktop background.”
- “The LINE function could be used to draw a box.”
- “The system enabled identification of threats.”

Internal phenomena (thoughts, feelings, and opinions) often best measured by surveying users.

Example

- “It was easy to find the information I needed in the maintenance guide.”

Translating Concepts to Questions

Conceptualization → **Define Measurement**

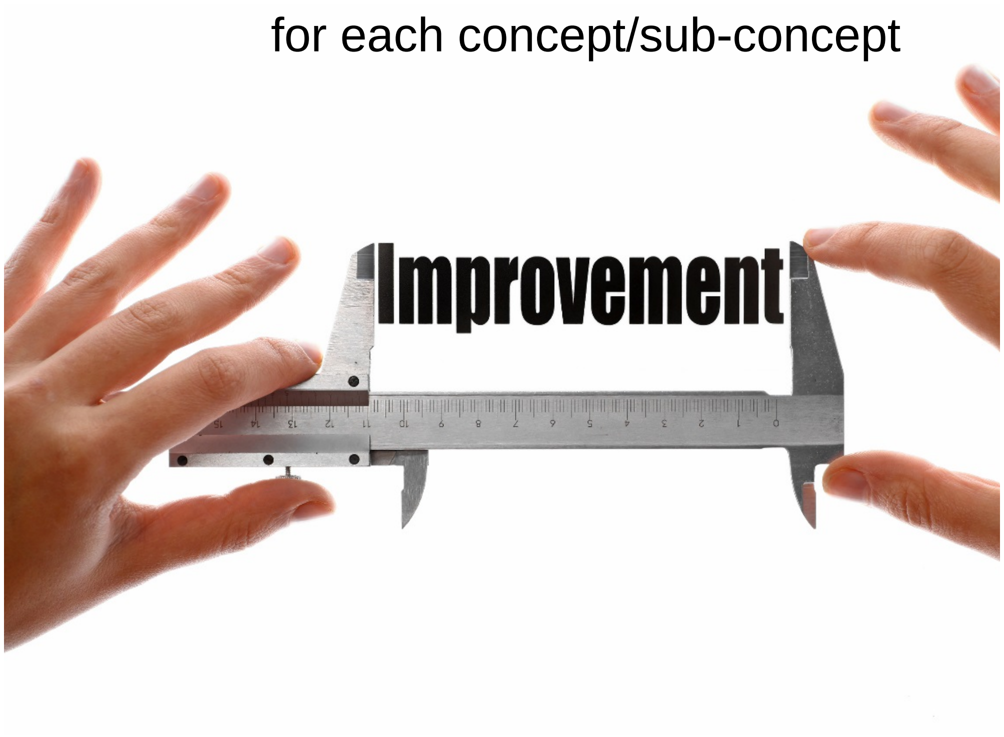
Name Concept

Define Concept

Specificity

Sub-concepts

Choose a measurement method for each concept/sub-concept



IDA

Addressing Concepts with Questions

Surveys are comprised of several parts.

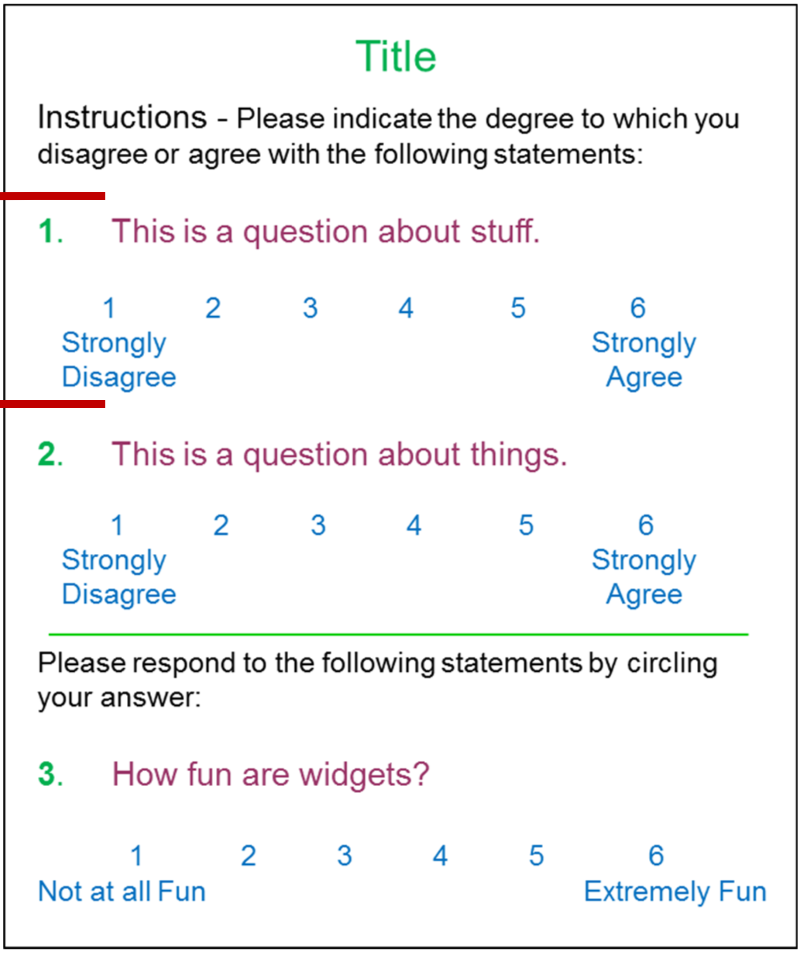
Survey: a collection of questions

Question: item and response option

Item: words a respondent addresses

Response Options: how the respondent provides an answer

Identifier/Formatting: symbols and layout to assist in organizing the survey



Title

Instructions - Please indicate the degree to which you disagree or agree with the following statements:

1. This is a question about stuff.

1 Strongly Disagree	2	3	4	5	6 Strongly Agree

2. This is a question about things.

1 Strongly Disagree	2	3	4	5	6 Strongly Agree

Please respond to the following statements by circling your answer:

3. How fun are widgets?

1 Not at all Fun	2	3	4	5	6 Extremely Fun

Translating Concepts to Questions

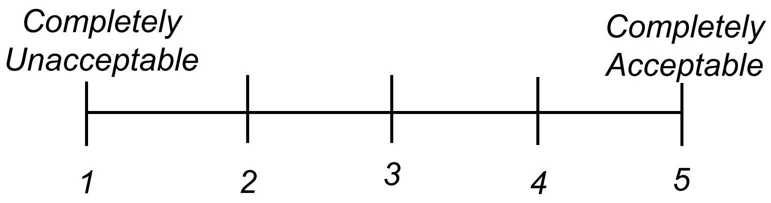
- **When appropriate for user rating, measures of effectiveness/suitability (MOE/MOS) may need “unpacking”**
  - May contain multiple concepts.
  - Will likely benefit from rewording.
- **What do MOEs/MOSs mean for user experience?**
  - Focus on the task users will complete.
  - Ask questions about what user should notice and remember.
- **Write question from user’s perspective, not tester’s.**
  - How testers think about the system may not be how users think about it.
  - Do respondents readily think in these terms?

Questions should address specific, well-defined tasks or attributes.

Example 1

Criteria: System must employ acceptable color display.

Rate the colors:



***Problem:** What attribute is being measured?*

– What should the user consider when determining acceptability?

- **If possible, learn about underlying issue:** Certain colors on the display monitor were not clear in direct sunlight.
  - Ask about user experience
- **Alternative wording based on intent of measure:**

“The contents on the screen were easy to read.”

Write Direct Items

- **Can be helpful to consider two components of an item.**
  - **Object** = The primary target of the rating.
    - » E.g., system, interface, task, perception
  - **Attribute** = The characteristic of the object being rated.
    - » E.g., difficulty/ease, accessibility, disagree/agree
- **Identify ONE object and ONE attribute per item.**
  - Be clear whether object is **overall** system or **specific** component.
  - Can be appropriate to ask at both levels of specificity, but remember working with limited resources.

Example 2

Criteria: Rating of maintenance panel accessibility.

Item: The system adequately supports maintenance panel accessibility.

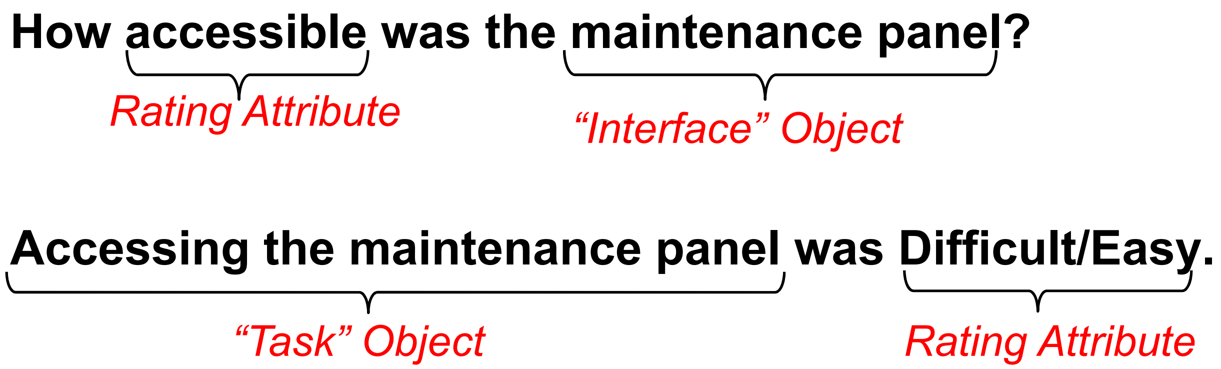
Object: System? Maintenance panel? Support? Accessibility?

Attribute: Adequacy? Accessibility?

Problem: Indirect, imprecise/vague

Define Concepts

~~The system adequately supports~~ maintenance panel accessibility.



How accessible was the maintenance panel?

Rating Attribute

“Interface” Object

Accessing the maintenance panel was Difficult/Easy.

“Task” Object

Rating Attribute

General Item Writing Tips

Goal: Write a simple, direct item for each object-attribute pair

- **What do we want to know?**
  - Ensure items clearly address goals
- **Who do we want to ask?**
  - Put yourself in the mind of the operator
  - Use a conversational tone
- **What will we do with the answers?**
  - Ask questions in a way that produces data needed for analysis

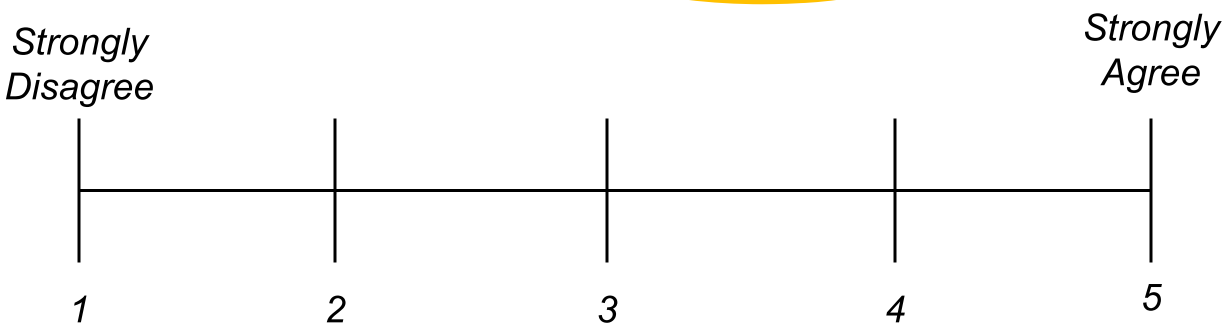
Item wording strongly influences the quality of responses

Golden Rules of Writing Items

Golden Rule	Definition
Singularity:	1 idea per question
User Friendly:	Items require little thought or interpretation
Neutrality:	Items do not imply value judgments
Knowledge Liability:	Respondents have sufficient info to answer question
Independence:	Earlier responses will not affect later responses

Singularity: 1 Idea Per Question

The display was bright and easy to see.

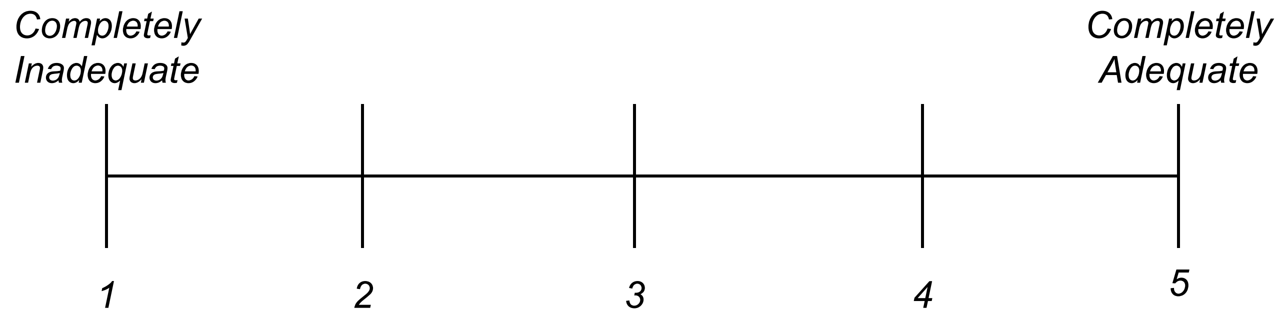


Avoid double barreled questions!

Respondents may not be clear which attribute to assess.

User Friendly: Requires Little Thought or Interpretation

Rate the adequacy of air-search radar & combat system to correctly decide to engage/not engage each track per Combat System Engagement Doctrine.



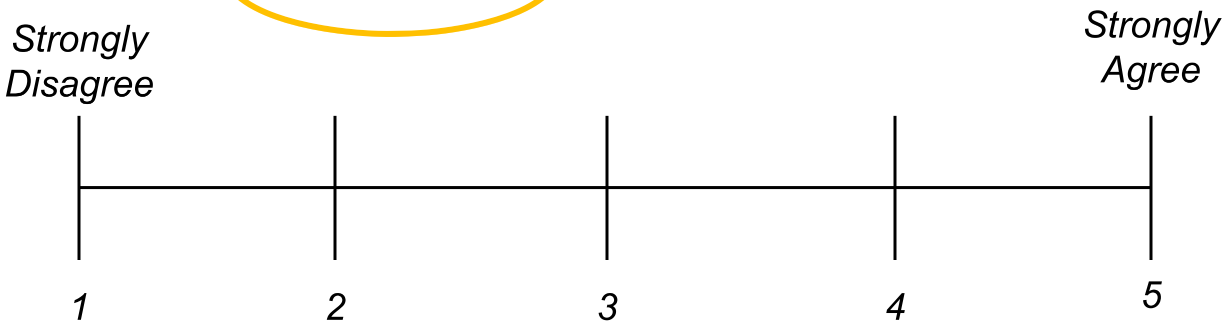
Be concise, clear, and specific!

↓

"I trusted the systems engagement decisions."

Neutrality: Items Do Not Imply Value Judgments

The software upgrade was necessary to eliminate annoying interruptions.



Strongly Disagree

Strongly Agree

1 *2* *3* *4* *5*

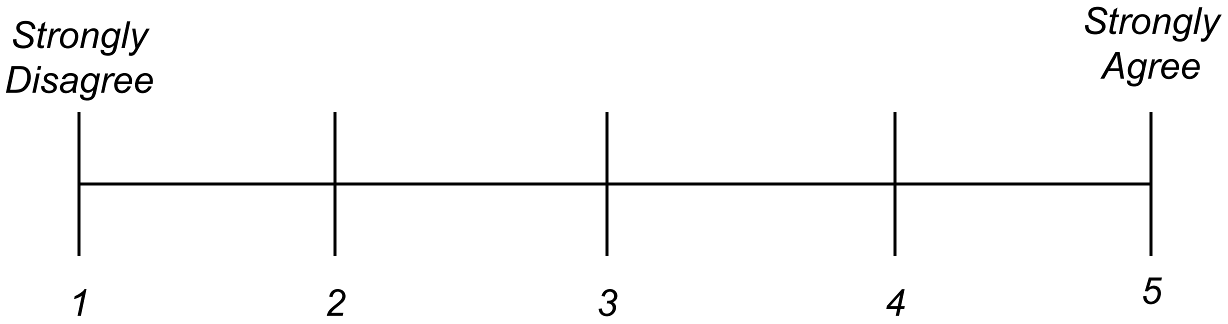
Violating neutrality can bias responding

Who decided they were annoying?

Was the upgrade necessary?

Knowledge Liability: Sufficient Info to Answer Questions

The targeting system performed accurately by tracking all targets.



Users are not in a position to judge accuracy because they may not have ground truth.

How could the user know if all targets were tracked?

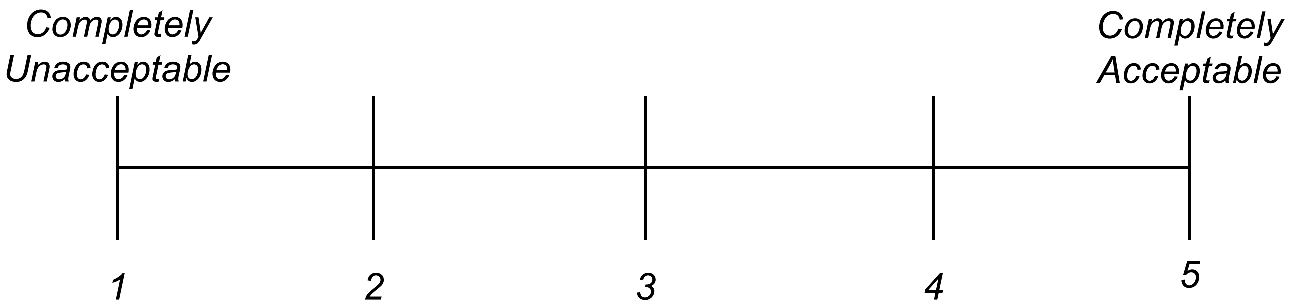
A Note on Accuracy

		Reality	
		Accurate	Inaccurate
Belief	Accurate	😍	☹
	Inaccurate	☹	😟

Mismatch

Independence: Earlier Responses Will Not Affect Later Ones

Based on your responses above, rate the acceptability of the system.



Can increase respondent burden

Rating may be independent of previous answers.

Response Types

Closed Response: restricts responses to a limited number of options.

Dichotomous

☒ No

□ Yes

Multiple Choice

□ Blue

☒ Green

□ Red

□ Orange

Rank

1 Killer Robots

2 Aliens

4 Zombies

3 Vampires

Scale

Strongly Disagree … *Strongly Agree*

1 *2* *3* [✓ *4*] *5*

Response Types

Open Response: unrestricted response options.

Fill-In The Blank

Age ________

Open-ended Response

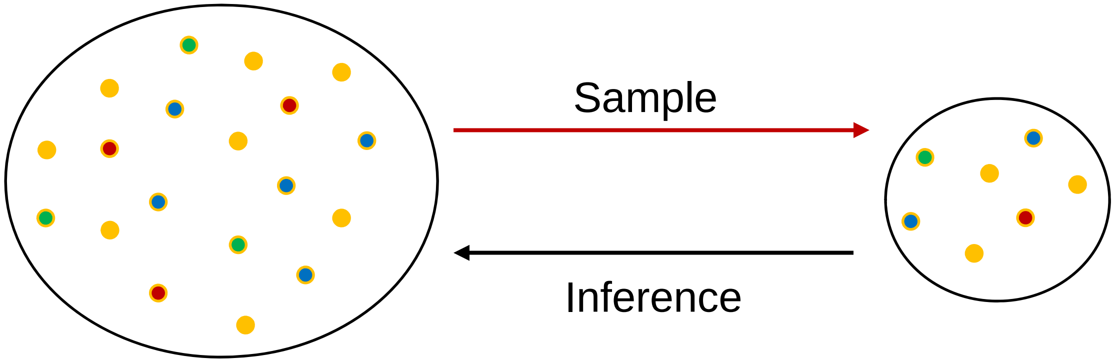
I can write whatever I want in this box! FREEDOM!

Selecting Response Options

Apply appropriate, precise measurement units

Consider the type of analysis required

- **Qualitative Analysis:** report characteristics of the sample.
- **Inferential Statistics:** make inferences about population characteristics from sample.



Open-Ended Responses

Useful when numerous or unanticipated responses are possible.

Open-ended Response

> I can write whatever I want in this box! FREEDOM!

- Qualitative analyses are possible.
- Cannot make inferences to the population.
- Can reduce respondent motivation.

Levels of Measurement

Inferential statistics require a quantitative measurement

- Measurement is more precise at higher levels.
- Higher levels of measurement require smaller sample size.

- **Nominal Level:** numbers simply serve as categories.
  - No ordering of cases is implied.

Dichotomous

□ Dead

□ Alive

Multiple Choice

□ Blue

□ Green

□ Red

□ Orange

Levels of Measurement

- **Ordinal Level:** numbers can be rank-ordered.
  - Cases can be ordered.
  - Distances between numbers are not meaningful.

Rank

_____ Killer Robots

_____ Aliens

_____ Zombies

_____ Vampires

Multiple Choice

□ No High school

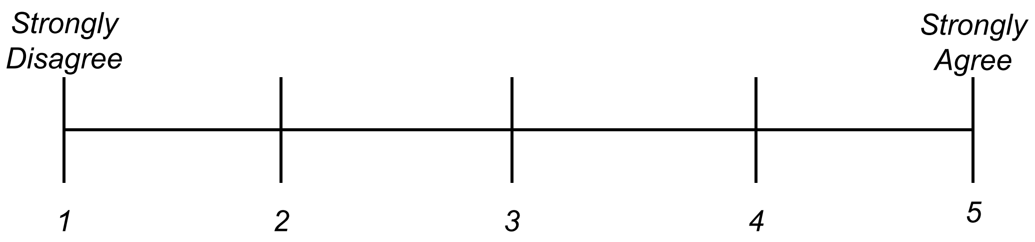
□ High school

□ Bachelors Degree

□ Graduate Degree

Levels of Measurement

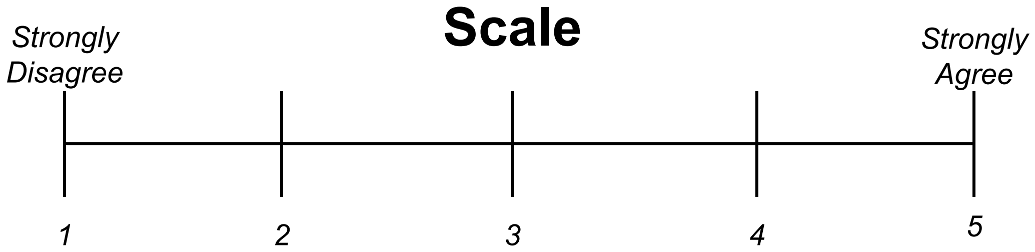
- **Interval/Ratio Level:** difference between numbers is constant.
  - – Can calculate averages!



Provides greatest statistical flexibility.

Best Practices of Scale Response

Scale responses best approximate interval level data



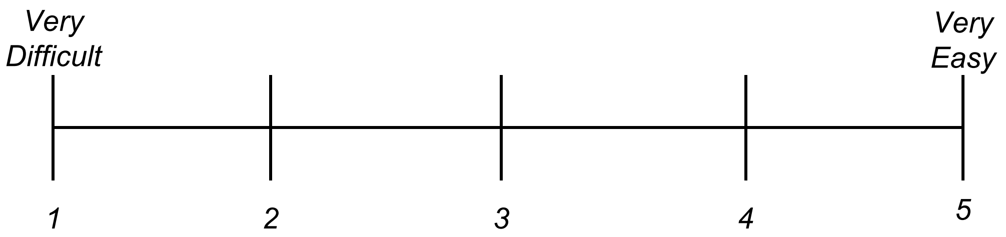
- 5-7 response options are recommended.
- Balanced, bipolar scales.
- Avoid neutral response option unless justifiable.
- If “NA” is included, do not position on scale.

DOT&E Guidance

http://www.dote.osd.mil/pub/policies/2015/4-2-15_Discussion_onIncludingNeutralResponses_onSurveyQuestions(9096).pdf

Scale Anchors

- **Anchors provide rating dimension.**



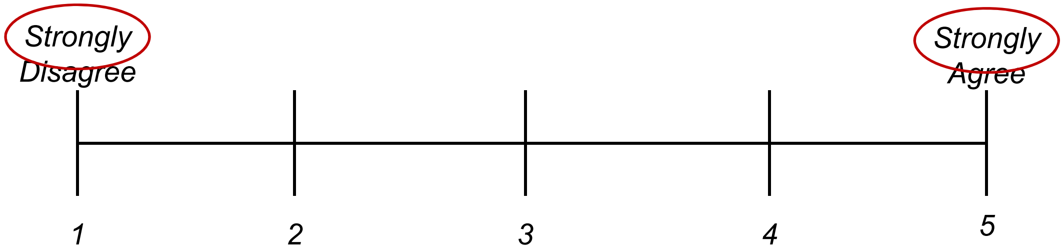
Ensure anchors match item!

How acceptable was the display? ≠ Inadequate/Adequate

U.S. Army Research Institute (1989) Questionnaire Construction Manual Annex.

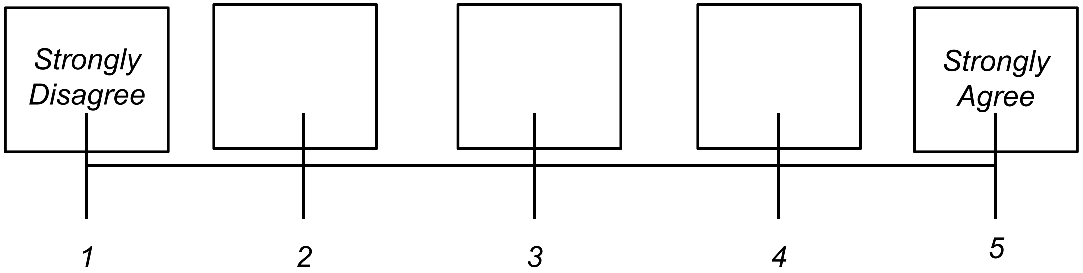
Scale Anchors

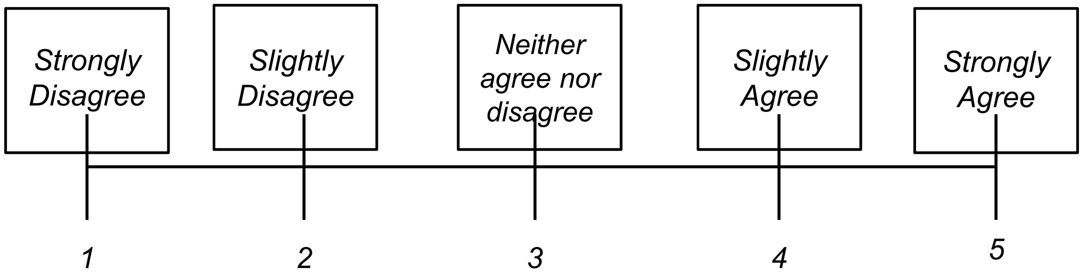
- **Modifiers/Qualifiers must be carefully selected.**
  - Modifiers indicate degree or magnitude
  - Ensure modifiers on both sides of bi-polar scale are equivalent
  - Make mutually exclusive categories
    - » E.g., Difference between adequate and completely adequate?



Scale Anchors

- **Anchor/Modifier positions can vary.**
  - End points only
  - Every other point
  - Every point (Note: Very difficult to do correctly)





Your Turn!

Activity: Part 3

Designing Surveys

Measurement Approach and Question Writing

Write question items and response options for the concepts that require user ratings.

1. **Recall the cell phone concepts you generated.**
2. **Which concepts are appropriate for user ratings?**

User Experience	Design Issue	Performance
Operator/ Maintainer Survey	Count/Checklist	Physical Measurement
	SME, Test Team, or Limited User Observation	SME Rating

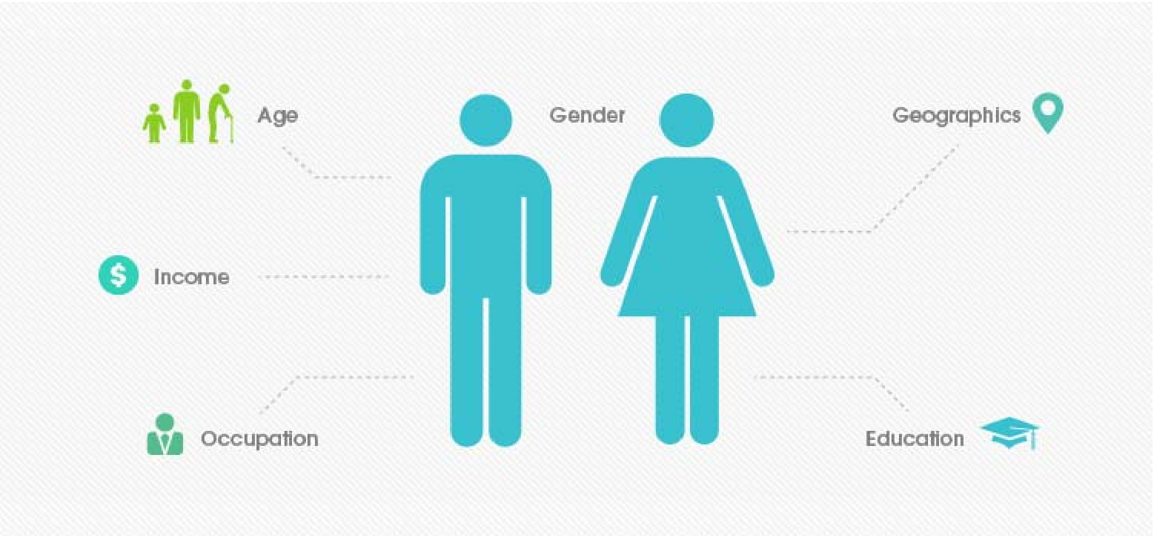
3. **Identify Object-Attribute pairs to address in each question.**
4. **Experiment with the wording. Which version gets closest to the intent of the question?**

Demographics

Demographic Sheets

Demographics should be administered once and tied to all surveys.

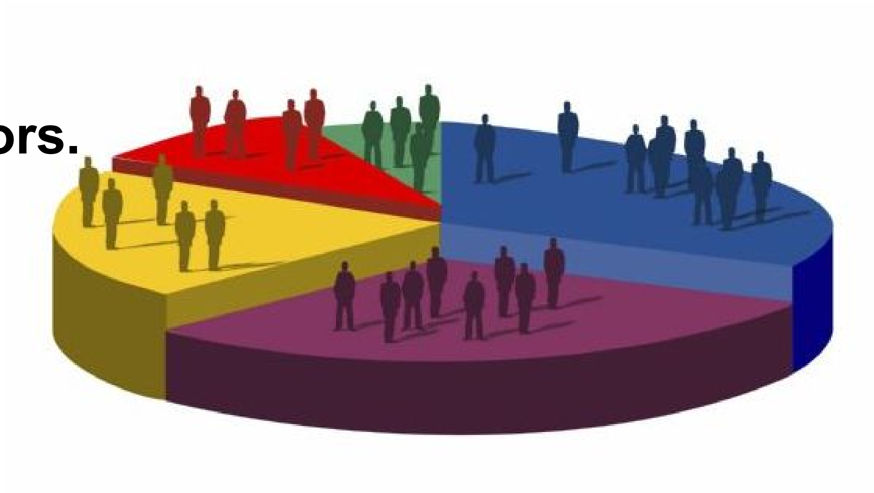
- **Information describing the respondent**
  - Personally Identifying Information (PII)
  - **Meaningful** background data
    - Characteristics that could influence interactions with system
    - Characteristics that could influence responses



Importance of Demographics

Demographics help us understand respondents

- **Characterize users**
  - Experience with legacy/new system
  - Training received
  - MOS/Role
- **Is the sample representative of the user population?**
  - Sampling bias
- **Can be used as predictive factors.**
  - Experience impacts performance



How are Demographics Measured?

Demographics are collected on data sheets.

- Factual questions, not ratings/opinions
- Typically fill-in-the blank and multiple choice
- Use examples

Date (mm/dd/yy):__/__/____ Name:___________________

Age:_______ Sex: Male Female

Grade (ex. E-5):________

Military Occupation Specialty (ex. 39B):_________________

Demographic Considerations

- **Don’t ask for too much**
  - Be brief! Ask only relevant information.
- **Try to provide standard response options when possible**
  - Anticipate potential responses
  - Especially important in joint scenarios
    - Ranks, MOSs, etc., differ across samples
  - Different “kinds” of answers to the same question result in incomparable data
- **Maintain confidentiality and protect privacy**
  - E.g., keep PII in separate record, not directly tied to responses
    - Use a participant number!
- **Put easy/NOT sensitive questions first**

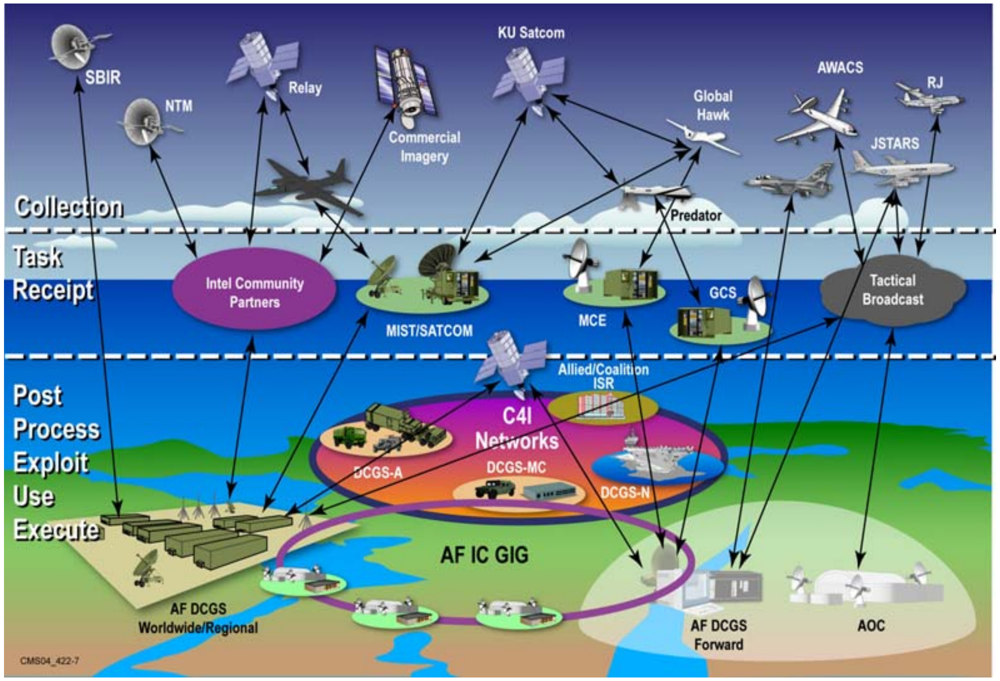
Case Study:
Air Force Distributed Common Ground System (AF DCGS)

Overview

- Introduction to the Air Force Distributed Common Ground System (AF DCGS)
- Usability testing for AF DCGS
- Analysis of findings
- Lessons learned

AF DCGS System Description

- An intelligence enterprise system
  - Hardware housed in 5 core sites and 16 distributed sites
  - Network connects them to each other and to other intelligence networks, sensors, and mission command systems
- Analysts manage, process, exploit, and disseminate information from various sources
  - Geospatial intelligence
  - Signals intelligence



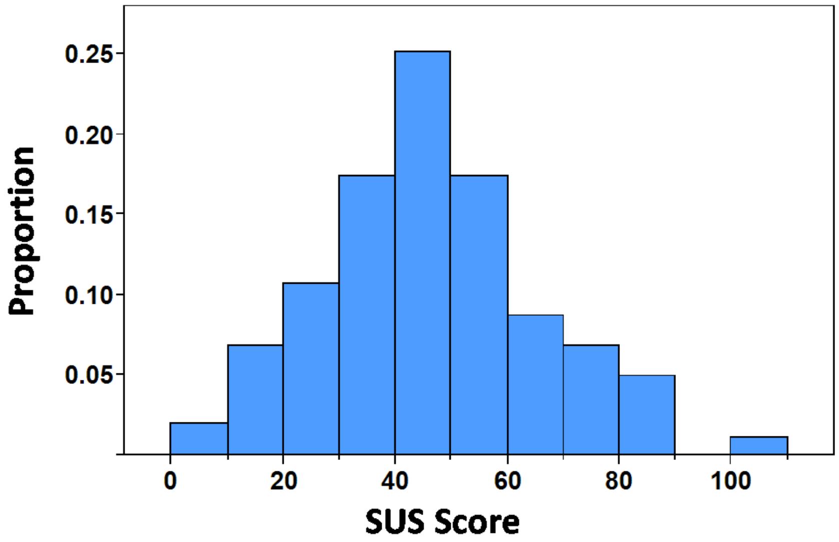
- Testing Bulk Release 10B Upgrade
  - Hardware Capability: Replace older servers that have reached end of service life
  - Software Capability: Two new web applications designed to increase operator workflow and enhance ability create/modify sensor tasks

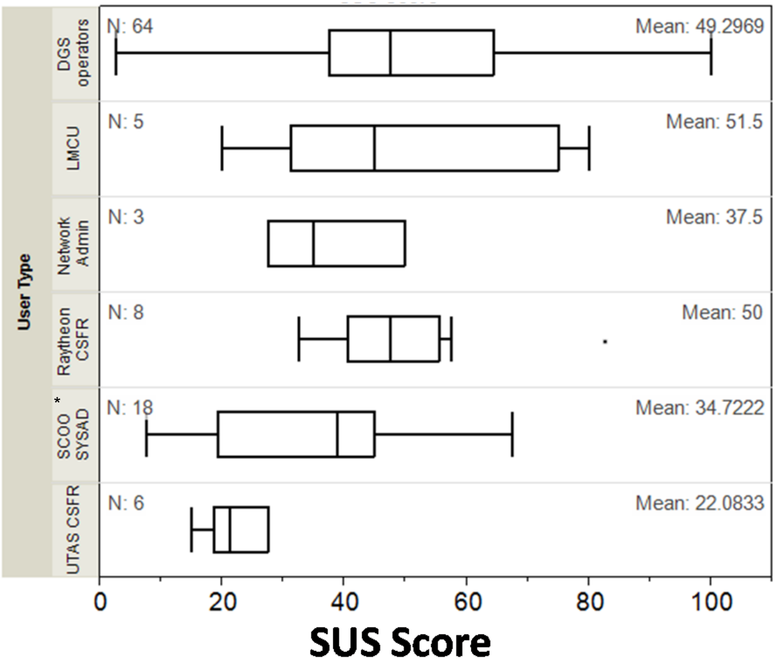
Acronyms this slide: Distributed Common Ground System (DCGS); Geospatial intelligence (GEOINT) External Tasking Service (ETS)

AF DCGS Survey Motivation and Test Plan

- Survey goal: Usability assessment of AF DCGS with upgrades
- Sample: DCGS operators, system administrators / network maintainers, and original equipment manufacturers' field representatives
- Surveys administered after every mission
    - » SUS
    - » Open-ended responses
- Proposed analyses
    - – Quantitative analysis of SUS
    - – Analysis of open-ended feedback
- Missing components of survey test plan:
    - – Unique ID linking surveys to open ended responses
    - – Unique ID linking surveys to missions

Quantitative Analysis: Bulk Release 10B is Difficult to Use





- 104 test participants
- Average System Usability Scale (SUS) score was ~45 (80% CI [42.6, 47.5])
- Significantly lower than the minimum score of 70 for a system to be considered acceptable

- Operators, system administrators, maintainers, and original equipment manufacturers all scored the usability as low

Acronyms this slide: System Usability Scale (SUS); Confidence Interval (CI); Department of Defense (DoD); Distributed Ground Station (DGS); Lockheed Martin (LMCU); Contractor Support Field Representative (CSFR); System Administrator (SYSAD); United Technologies Aerospace Systems (UTAS)

Analysis of Free Response Comments

- “Most of the issues that arise with the BR-10B system are due to a lack of TTPs when working with the system. Once these issues are understood there is little to no mission impact. However, there are no apparent benefits when working with 10B over 10.1”
- “BR-10B system is great in theory, but poorly implemented. Program still not fully functional. Complete lack of training, for a system that changes the entire way of issuing targets.”
- “I do not remember going to a training class for BR-10B. Other than that, BR-10B with 10.1 TTPs functions the same as 10.1.”
- “The system is much better implemented when using 10.1 TTP's for research. Workflow has potential to be more effective than 10.1, however it has fundamental problems.”
- “Although the MOC does not use BR-10B, when issues on 10B cause us to be unable to exploit HA imagery, the mission in general gets backed up. From what I gather, most of our analysts would rather not use 10B.”

Conclusions and Lessons Learned

- AF DCGS upgrade has poor usability, likely due to:
  - Poor software design
  - Insufficient training
  - Documentation on the system

- Additional data and better data collection techniques could have produced more / improved analyses

- Include an anonymous identifier on all surveys and free response sheets would allow the test team to:
  - Match scores with comments
  - Assess if usability varied across demographic variables (e.g., experience)
  - Make quantitative comparisons by mission (e.g. was the software more usable on some missions that others?)

IDA

Formatting Surveys

Survey format can greatly impact your response rate.

- **Begin with an introduction**
  - Title
  - Survey Topic
  - Voluntary & Confidential
  - Sponsor & Contact
- **Create a professional look**
  - Standard, readable font
  - Not too "busy"
- **Logically order and group questions**
  - Consider topic and response option type
  - Response option matrices can ease burden
  - Use general to specific ordering
- **Keep it short**

Evaluation Checklist

Goal: Identify ambiguous questions and those that don't reflect the user experience.

□ **Question relevance:** items that don't address test issues/measures should be excluded.

□ **Question wording:** items that violate golden rules or that don't match the descriptor set should be revised.

□ **Questionnaire format:** ensure instructions are clear, the survey is not too long, and organization is easy to follow.

Your Turn!

Activity: Part 4

Survey Evaluation

Survey Evaluation

Switch surveys with another group

- **Use the evaluation checklist to critique your partners' work.**

□ **Question relevance:** items that don't address test issues/measures should be excluded.

□ **Question wording:** items that violate golden rules or that don't match the descriptor set should be revised.

□ ~~**Questionnaire format:** ensure instructions are clear, the survey is not too long, and organization is easy to follow.~~

Outline

- **Measurement Basics**
- **Selecting a measurement method**
- **Survey Basics**
- **Types of surveys & how to construct them**
  - Empirical Surveys
  - Custom-Made Surveys
  - Demographics Surveys
- **Survey Administration & Data Collection**
- **Data Analysis**

Survey Administration

Key Administration Concerns

- **Administration techniques ultimately impact data quality**
- **Respondents are sensitive to:**
  - Test schedule & constraints
  - Operational context
  - Motivation and ability
- **Survey administrators should consider:**
  - Timing & Frequency
  - Survey environment
  - Administration method
  - Introduction & Instructions

Notes on Motivation

- **Quality and quantity of data is limited by respondent motivation and ability**
  - Adequate knowledge
  - Demand of mission & resulting fatigue
- **Survey demand must match motivation**
  - Consider energy and desire
  - Surveys add to fatigue
  - Asking more from unmotivated respondents yields lower quality data

Administration Timing

- **Driven by test schedule, constraints, and goal**
- **Common administration times**
  - Posttest/Exit survey
    - » Overall assessments
    - » Typically longer
  - Natural break points (End of Task/Mission/Day)
    - » May be more specific questions
    - » Captures change due to different conditions and tasks
    - » Typically shorter
  - Event-driven
    - » Delivered in response to spontaneous events
      - E.g., Task interruptions
    - » Shortest

Survey Frequency

- **Frequency depends on question**
  - Do you need overall assessments?
  - Do you intend to measure change across conditions?
  - Do you need data on unique incidents?
    - E.g., Software bug reports

- **Repeated measurement**
  - Asking the same question about the same issue multiple times
  - Must track/match respondent forms
  - If administered too frequently or if too long surveys can result in survey fatigue.
    - Respondent should know there is a reason for asking again

- **Not equivalent to replication**
  - Dependent data points

Other Timing Considerations

- **Time since task**
  - Trade off between memory and fatigue
  - Impressions fade as time passes
- **Time between tasks**
  - Respondent availability may be limited by upcoming tasks
  - Motivation is reduced just before leaving shift

Survey Environment

- **Similar to test environment…**
  - – Preserves memories and impressions of the system/task
- **…But, as free from distraction as possible**
- **Minimize interaction among respondents as well as with test team**
  - – Reduce potential bias

"Your ideal survey environment!"®

Administration Method

Carefully consider logistics of delivery and collection

- **Paper and pencil**
  - Most common
  - Fairly flexible and convenient
  - Cheap and widely available
  - Data requires manual entry

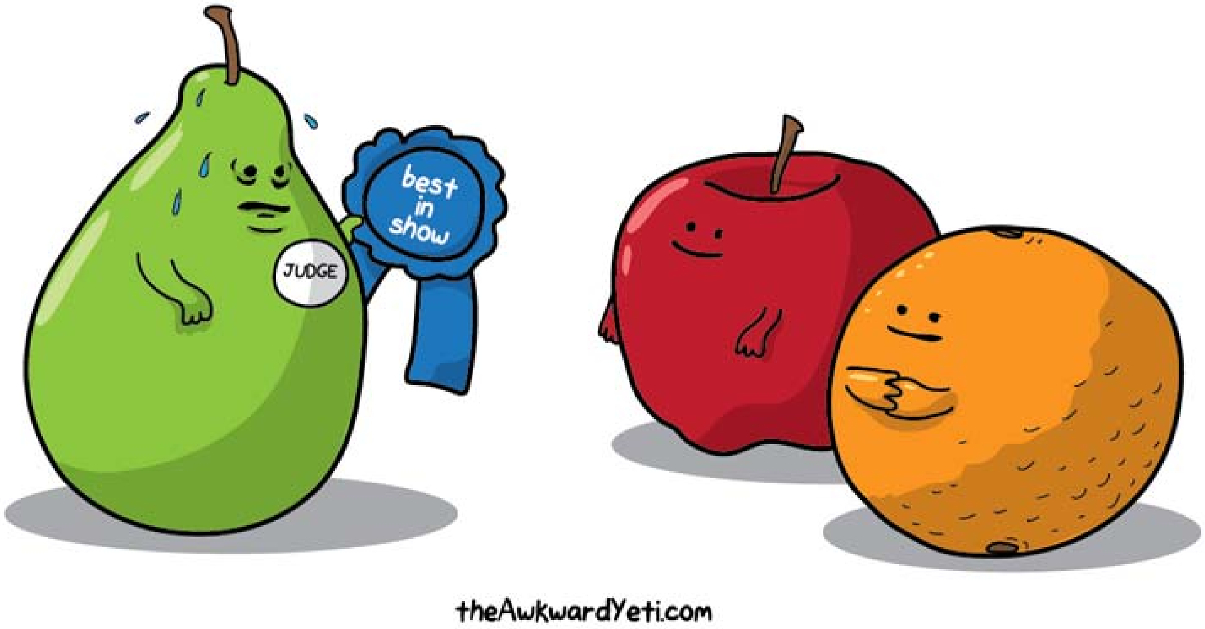
- **Electronic**
  - Question branching
  - Automatic database generation
  - Network security & data storage constraints
  - Variation of software capabilities
  - Consider availability of devices
- **Verbal**
  - Allows brief responses during operation
  - Beneficial for follow-up questions
  - Test team must record each individual response
  - Least confidential

Set the Tone with Introductions

- **Motivation can be influenced with survey introductions and tone**
- **Use introductions to encourage investment**
  - Communicate purpose and context
  - Relevance to participants (e.g., how responses will be used)
  - Responses are important to fully characterize the system
  - Responses are confidential
- **Include directions and examples to help frame responses**
- **Acknowledge time spent and thank them for their input**

Administration Consistency

- **Administration techniques can impact survey results**
- **Maintain consistency to improve data quality**
  - Minimize noise and potential confounds
  - Especially important for comparisons



Outline

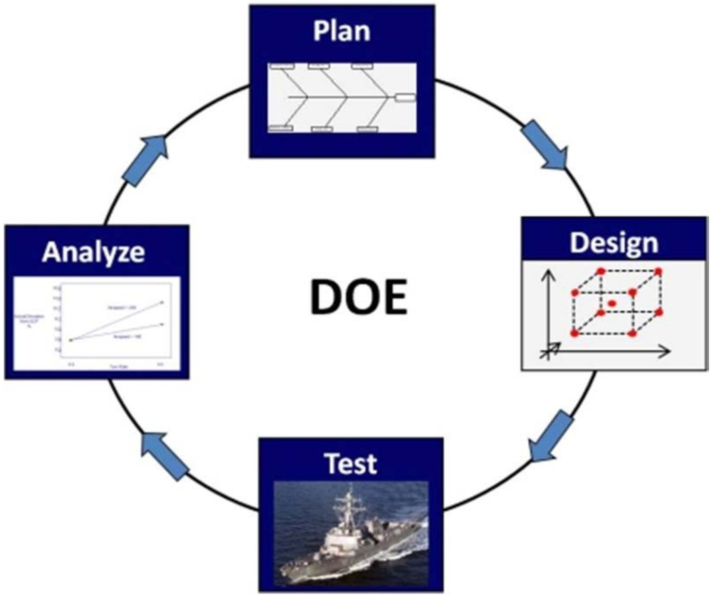
- **Measurement Basics**
- **Selecting a measurement method**
- **Survey Basics**
- **Types of surveys & how to construct them**
  - Empirical Surveys
  - Custom-Made Surveys
  - Demographics Surveys
- **Survey Administration & Data Collection**
- **Data Analysis**

Data Analysis

IDA

Surveys in the DOE Context

1. Define the objective of the experiment
2. Select appropriate response variables
3. Choose factors, levels
4. Choose experimental design
5. Perform the test
6. **Statistically analyze the data**
7. **Draw conclusions**



Successful analysis is contingent on adequate test planning and implementation

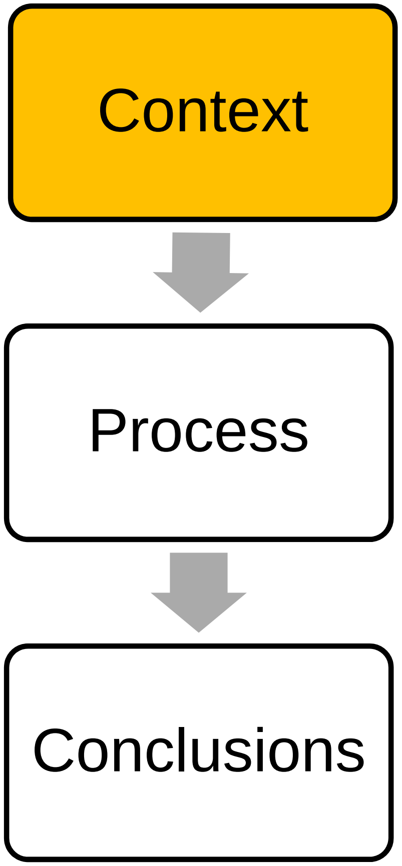
The Survey Context

- **Survey analysis is bounded by:**
  - Data collection processes
  - Experimental design
  - Questions of interest
- **Questions to consider:**
  - How were the data collected?
  - Who were the data collected from?
  - Are there other variables that may affect the interpretation of observed relationships?

The Survey Context

- **Be consist by controlling factors or recording variables**
  - Administrators with similar training
  - Environment (may not be controllable, but **is recordable**)
  - Administer surveys at the same time
  - Use the same instructions
- **Ensure participants represent population**
  - Range of skills & abilities representative of operational users
  - **NOT** just best individuals
- **Reduce experimenter bias**
  - Little interaction between respondents & test team
  - Administrator presents as if no opinion on survey

Survey Analysis Road Map



Context

Understand the data type, scale, and survey context

Process

Error check, describe, and visualize survey data

Conclusions

Statistical analysis and recommendations

Survey Scenario

Example

The Army is interested in assessing the **effectiveness and suitability** of an updated **mine detection software suite**. As a part of the test, operators completed a **survey on the usability** of the system and their **willingness to take the system to war**. Further, **performance data** were collected on the number of targets processed. Each operator was surveyed once during either a **day or night mission** (randomly assigned).

Data Context

- **SUS Score:** SUS scores range from 0-100 with larger scores meaning better usability
- **Mission Type:** Day vs. night mission
- **I would take this system to war:** Likert-scale response- 6 levels ranging from strongly disagree to strongly agree
- **Performance:** Number of targets successfully processed

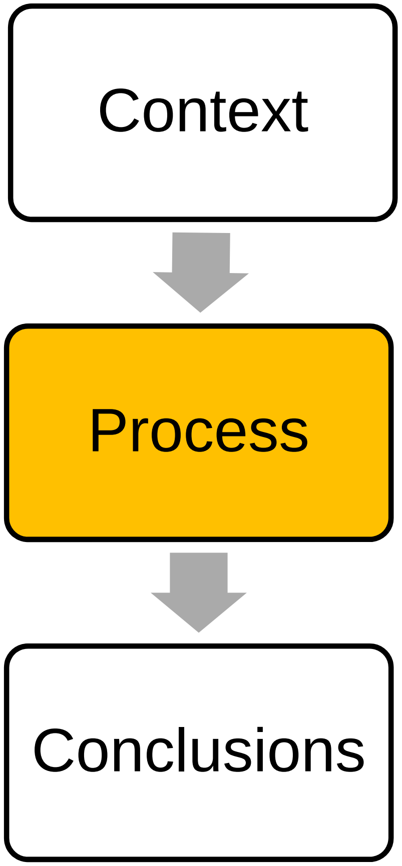
Data Context

Operator ID	Mission Type	SUS Score	I would take this system to war	Number of Targets Processed
1	Day	70	1	10
2	Night	55	5	5
3	Night	41	2	7
4	Day	81	4	15
5	Night	40	5	11
6	Day	65	3	12
7	Night	55	2	4
8	Day	65	1	8
9	Day	95	5	13
10	Night	51	1	6
11	Night	70	4	9
12	Night	65	3	6
13	Day	73	4	11
14	Day	75	3	14

Rows = Respondents

Columns = Survey Items, Factors, Demographics

Survey Analysis Road Map



Context

Understand the data type, scale, and survey context

Process

Error check, describe, and visualize survey data

Conclusions

Statistical analysis and recommendations

Processing Data and Descriptive Statistics

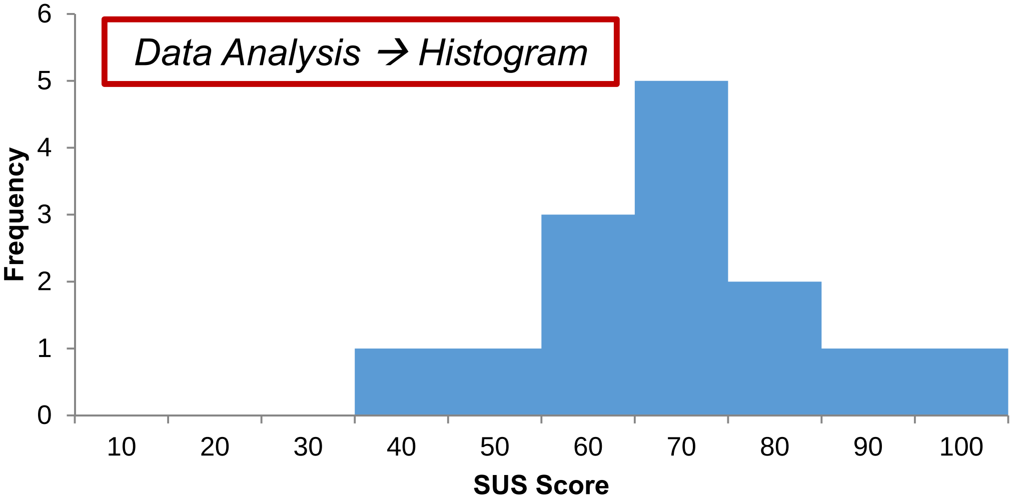
Variable Type	Description
Nominal	Categories (e.g., Mission type)
Ordinal	Ranks & Ordered Categories (e.g., 1st, 2nd place)
Interval	Numerical data with equal intervals (e.g., SUS Scores)

Descriptive Statistic	Question of Interest	Data Type
Mode	What is the Most Common Response?	Nominal
Median	What is the 50th Percentile Response?	Ordinal
Mean	What is the Average Response?	Interval
Std. Deviation	How Variable is the Data?	Interval

Processing Data

- **Goal Driven:**
  - Univariate Analysis: Characterize operators' SUS scores
- **Visualize and describe**
  - Do the largest and smallest values make sense?
  - Is the count consistent with sample size?
  - What is the shape, center and spread of the distribution?

Histogram: SUS Scores



Data Analysis → Histogram

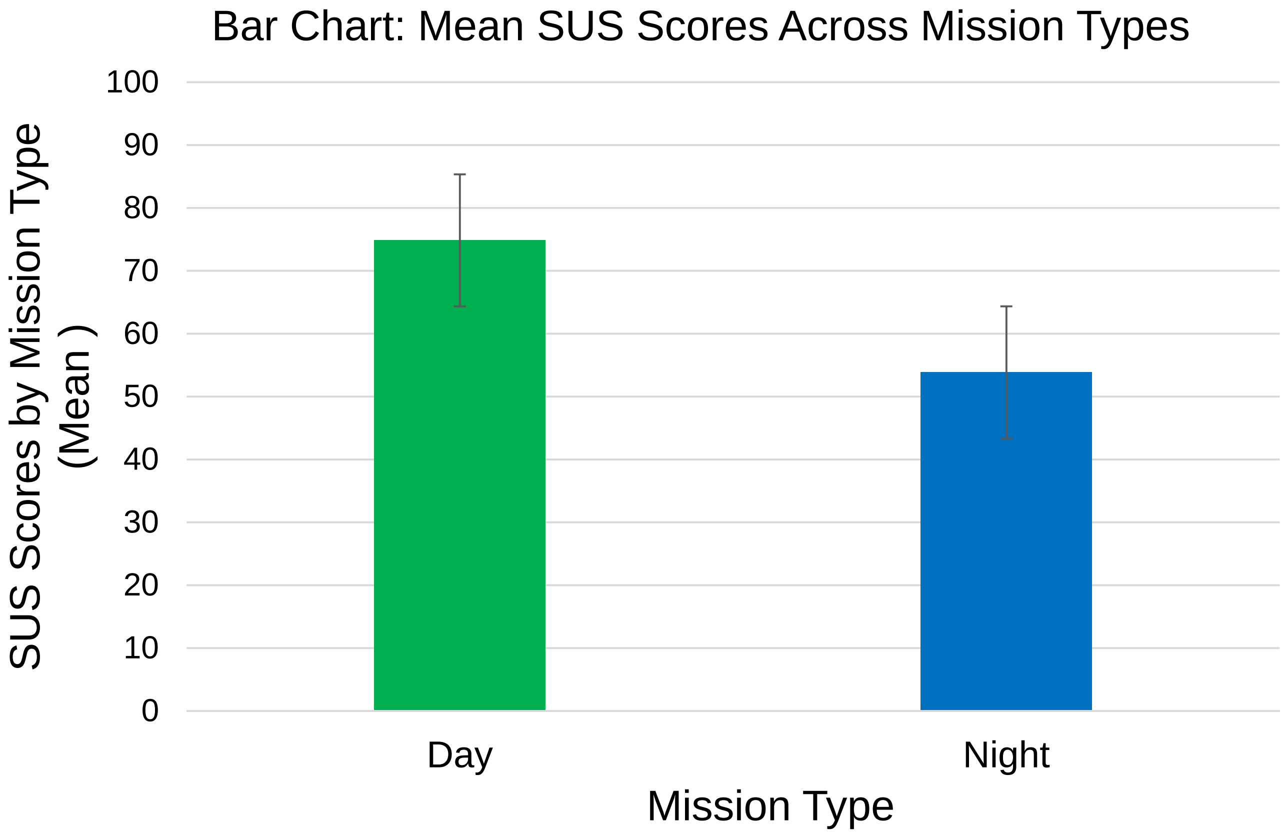
Descriptive Statistics	
Mean	64.36
Standard Error	4.03
Median	65.00
Mode	65.00
Standard Deviation	15.09
Sample Variance	227.79
Kurtosis	0.11
Skewness	0.11
Range	55.00
Minimum	40.00
Maximum	95.00
Sum	901.00
Count	14.00
Largest(1)	95.00
Smallest(1)	40.00
Confidence Level(95.0%)	8.71

Data Analysis → Descriptive Statistics

Processing Data

- **Bivariate Analysis: Characterize operators' SUS scores by mission type.**

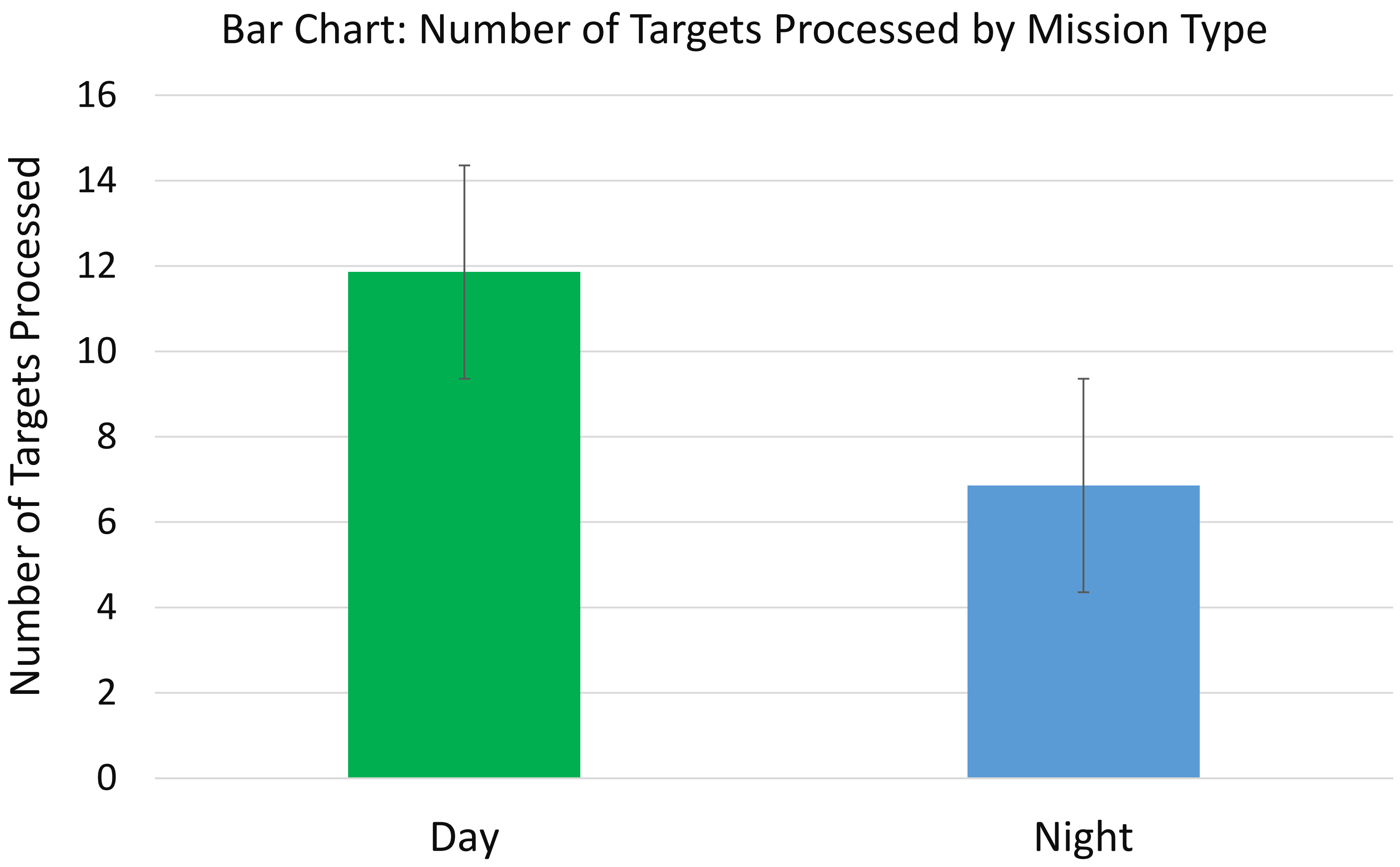
Mission	Mean SUS
Day	74.86
Night	53.86



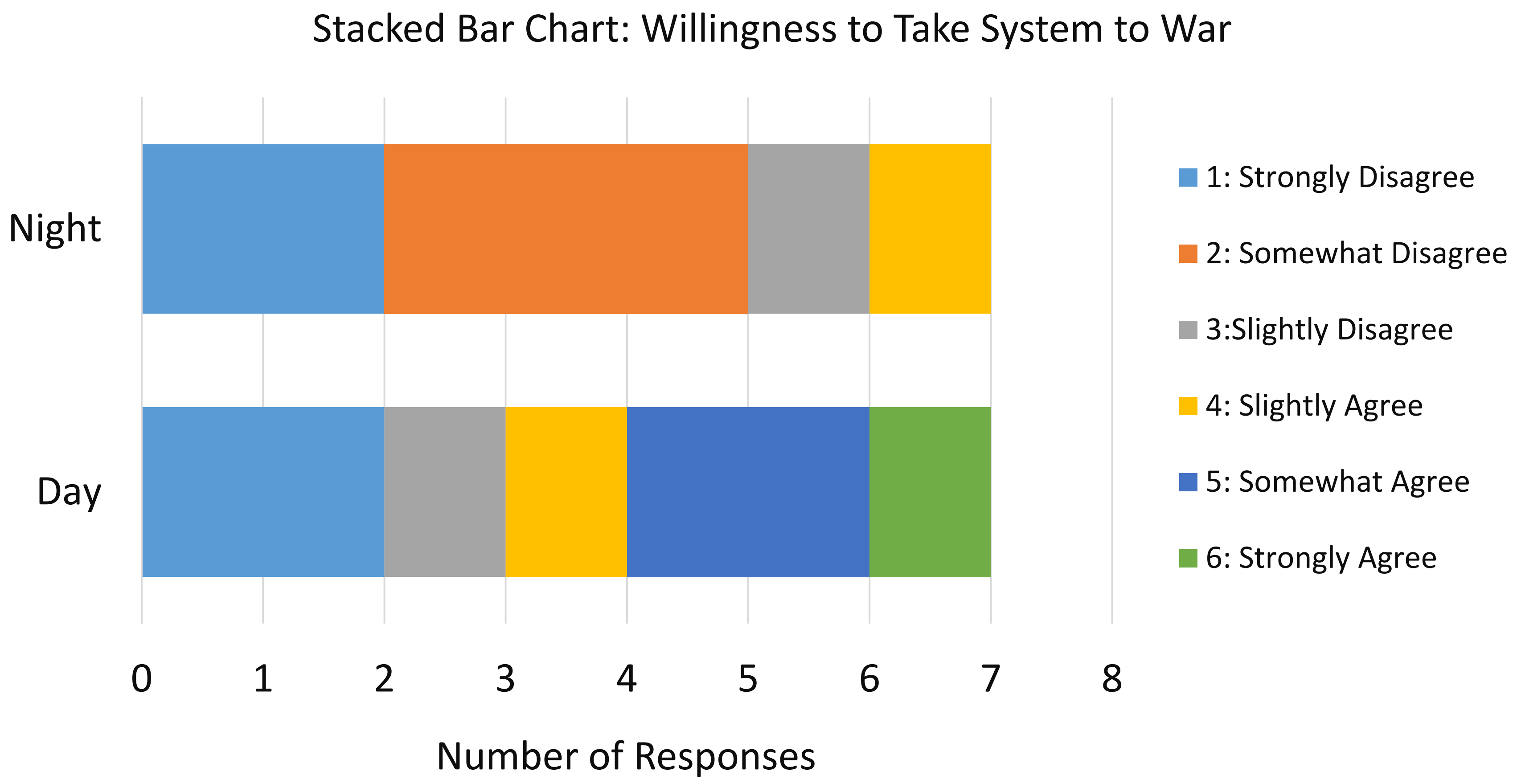
Note. Standard error bars presented

Insert → Pivot Chart → Click variable boxes → Change from sum to average

Processing Data

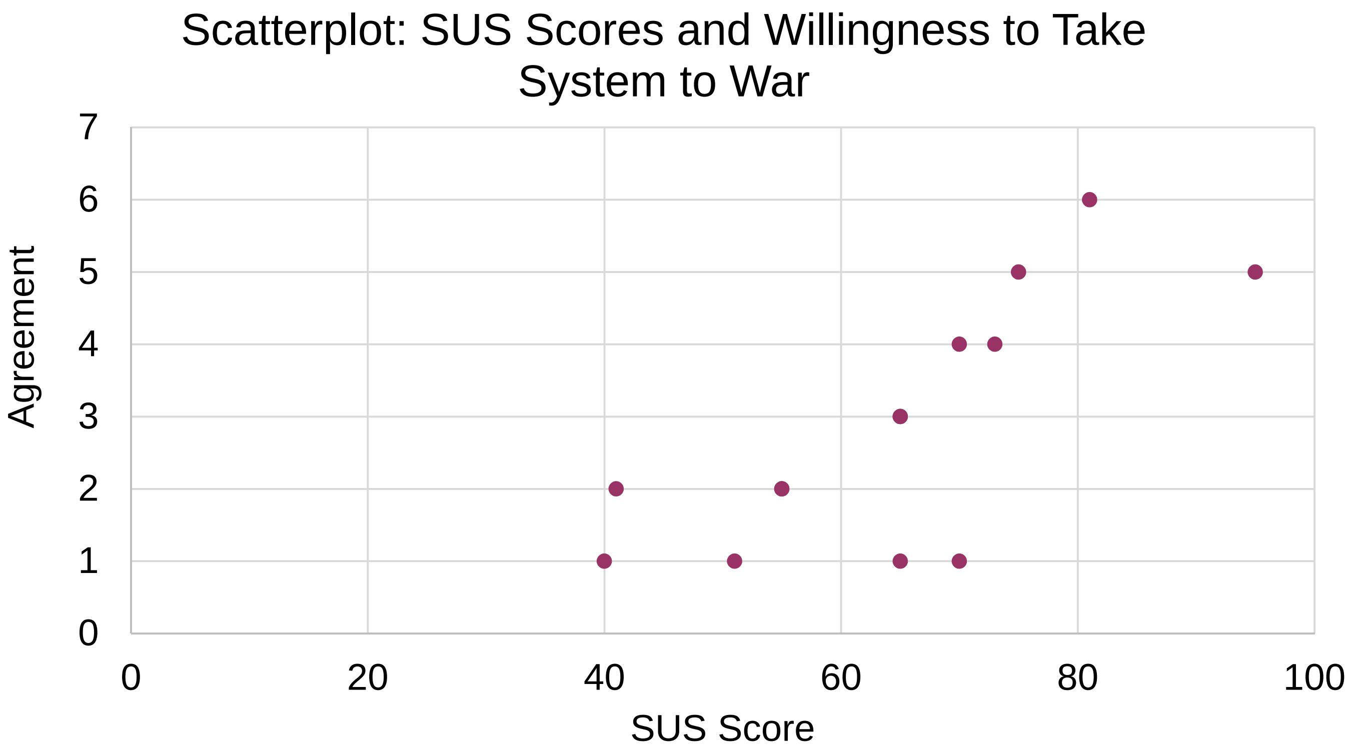


Processing Data



Example: Processing Data, cont. 2

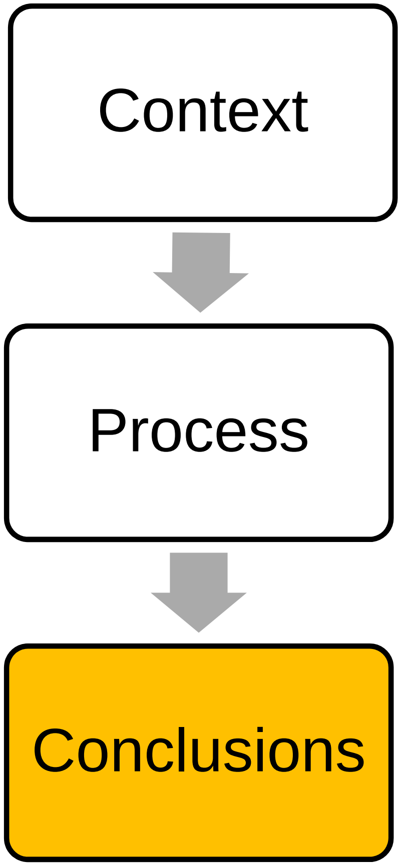
Bivariate Analysis: Characterize the association between SUS scores and willingness to take system to war.



$r = 0.75$

Insert → Scatterplot

Survey Analysis Road Map



Context

Understand the data type, scale, and survey context

Process

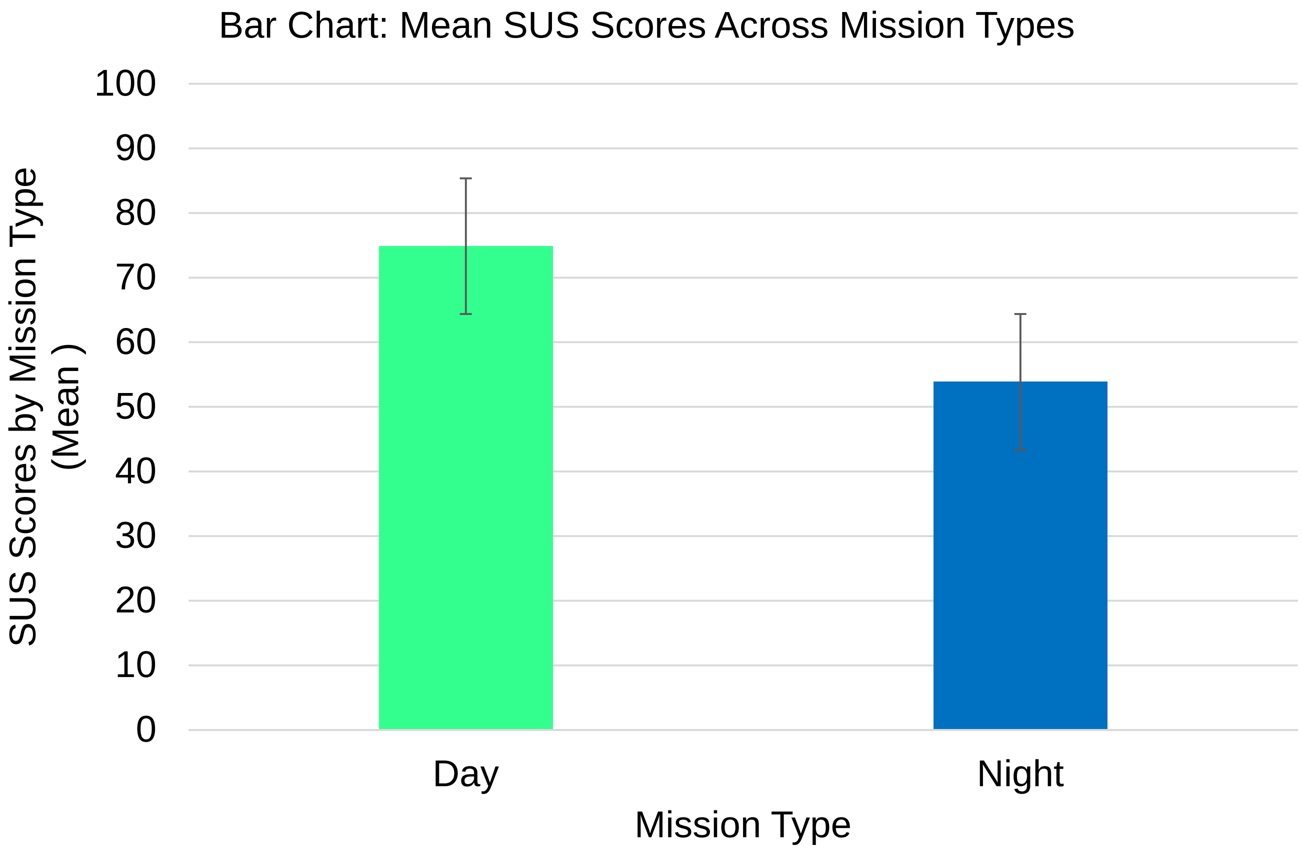
Error check, describe, and visualize survey data

Conclusions

Statistical analysis and recommendations

Drawing Inferences

Mission	Mean SUS
Day	74.86
Night	53.86



Note. Standard error bars presented

Insert → Pivot Chart → Click variable boxes → Change from sum to average

Drawing Conclusions Through Inferential Statistics

- **Statistical analysis of survey data doesn't differ from other data types**

- **Goal: Determine if SUS scores significantly differ for day vs. night missions in the population of operators**
  - Test selection: Bound by questions of interest and data
    - Independent sample *t* test appropriate for comparing two groups.
    - Each inferential test has assumptions that should be checked before reporting statistics
    - Inferential statistics are not always appropriate or to be desired.
  - The sample dictates the population to which you can generalize.
    - If you did not survey novice operators, you cannot generalize your findings to this population

Drawing Inferences

- **Normality: Visually inspect histograms**
  - Roughly normally distributed. SUS scores for day missions are somewhat skewed.

- ***F* Test for equality of variances**
  - $\text{Variance}_{\text{day}}$ = 110.81, $\text{Variance}_{\text{night}}$ = 125.48
  - *F* test is not statistically significant (*F* = 0.88, *p* =.44)

- **Compute statistic**
  - Excel: T.test function.
  - *p* = .004

- **Interpret**

Mission Type	Mean SUS Score
Day	74. 86
Night	53. 86

Inferential Statistics

Inferential Statistic	Question of Interest	Variable Type
Chi Square	**Is Frequency Distribution Different from Expected?** Example: Are the operator types distributed similarly across mission types (Day Vs Night)	Nominal
Sign Test/ **One sample *t* test**	**Is RV Different from a Threshold/Standard?** Example: Are average SUS scores significantly greater than 70?	Ordinal/ Interval
Correlation	**Are 2 Factors Correlated?** Example: Are SUS scores correlated with performance metrics?	Interval
Regression	**Can a RV Be Predicted from Other Factors?** Example: Does amount of training predict performance outcomes?	Interval
***t* test**	**Are there statistically significant differences in the means between two groups?** Example: Do mean SUS scores vary across day and night missions.	Interval
ANOVA	**Is RV Different Under Different Levels of Factor(s)?** Example: Do mean SUS scores vary three or more different operator types?	Interval

RV: Response Variable

Evaluation of Findings

- **Operators reported poorer usability for night missions (unacceptably low)**
- **Performance was worse for night missions**
- **Evaluate different features of the test environment across mission type**
  - Features of the test scenario:
    - Were test conditions different for day and night missions?
  - Features of the system:
    - Did operators report difficulty using the computer software at night?
  - Features of the team:
    - Were different teams assigned to night missions (less experienced teams)?
    - Did teams assigned to day and night missions receive different training?

Special Considerations

- **Missing Data**
  - There are complex methods for dealing with missing data
- **Confidence Intervals (CIs)**
  - Supplementing measures of central tendency with CIs provides an indication of the uncertainty of our estimates.
- **Effect Sizes**
  - Consider the practical and statistical effects of your findings

Conclusion

- **Survey data, like any data resulting from DOE, can be subject to rigorous analysis**
  - Analysis begins by establishing the data context
  - Data can then be processed and visually explored
  - Inference can be drawn from data, if appropriate.
- **Possible conclusions depend on the design and nature of data**
- **Become empowered with Excel's Data Analysis ToolPak**

Survey Question Contact Information

- **Heather Wojton – hwojton@ida.org**
- **Jonathan Snavely – jsnavely@ida.org**
- **Chad Bieber – cbieber@ida.org**
- **Justin Mary – jmary@ida.org**