Introduction to Surveys in Operational Test & Evaluation

Rebecca A. Grier

Justin Mary

IDA

Course Agenda

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

Goals

- What is a survey?
- How is a survey different from a data sheet? Interview? Focus Group?
- What role do surveys have in DOE?
- Are there different kinds of surveys?
- How are surveys incorporated in TEMPs and Test Plans

Human Factors

- **There's a Human in the System!**
  - Operational effectiveness of military systems depends on the users and maintainers
  - To assess the complete system, the human component of the system must be measured
- **Some system metrics are affected by human performance & thought**
  - Mission effectiveness, accuracy, maintainability, time to complete mission, time to employ
  - Some do not – weight, speed, range
- **It’s not the individual, it’s the population**
  - Measurements have to be made individually, but with the goal of telling a story about the whole population
  - Confidentiality increases likelihood of honest opinions
  - Measuring how the system supports the user, not the user’s performance

Survey Purpose & Goal

A ***systematic collection & analysis of data*** relating to the ***thoughts*** of a population.

1. Collect specific data for a pre-defined purpose.

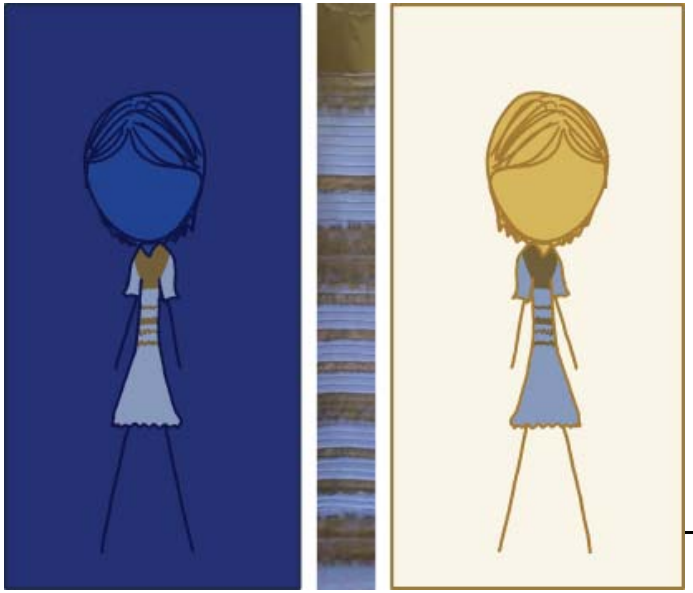
2. The survey design determines the validity of the data and the statistical analyses possible.

Who	Role	Sources of error
OTA/ DOT&E	Defines survey's purpose & uses information from survey	• Not enough information • Wrong information
System operator /maintainer	Gives data	• Answers different question • Thinks too much • Doesn't think enough
Analyst	Translates data into information	• Unable to analyze data • Aggregates data incorrectly

IDA

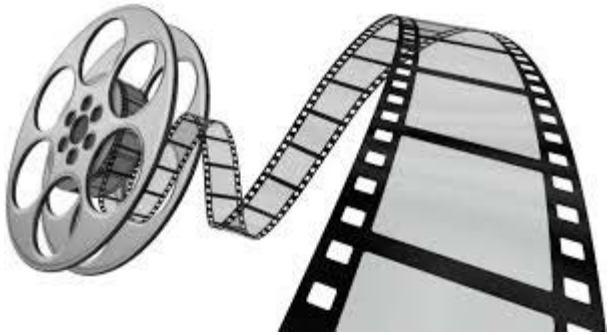
Surveys Measure Thoughts, Which Are Context Specific!

What color is the dress?



What do you see?

Video on color illusions



Which is bigger?

What Is The Difference Between Surveys, Data Sheets, Interviews, & Focus Groups?

Data Sheets

Record distinct observables

- *Dichotomous (component present: Y/N)*
- *Clear categories (weather)*
- *Easily countable (amount of gear)*
- *Start/stop time*

Surveys

Measure thoughts for statistical analysis

- *Static*
- *Utility, usability, workload*
- *Paper, electronic, or verbal administration*
- *Appropriate for planned events*
- *Concise set of known short responses*
- *Free response questions are exception*

Interviews

Collect non-specific diagnostic thoughts

- *Not static; Respond to unplanned events*
- *Countable; not analyzable by statistics*
- *Problem identification*
- *Questions with large number of potential responses*
- *Long responses anticipated*

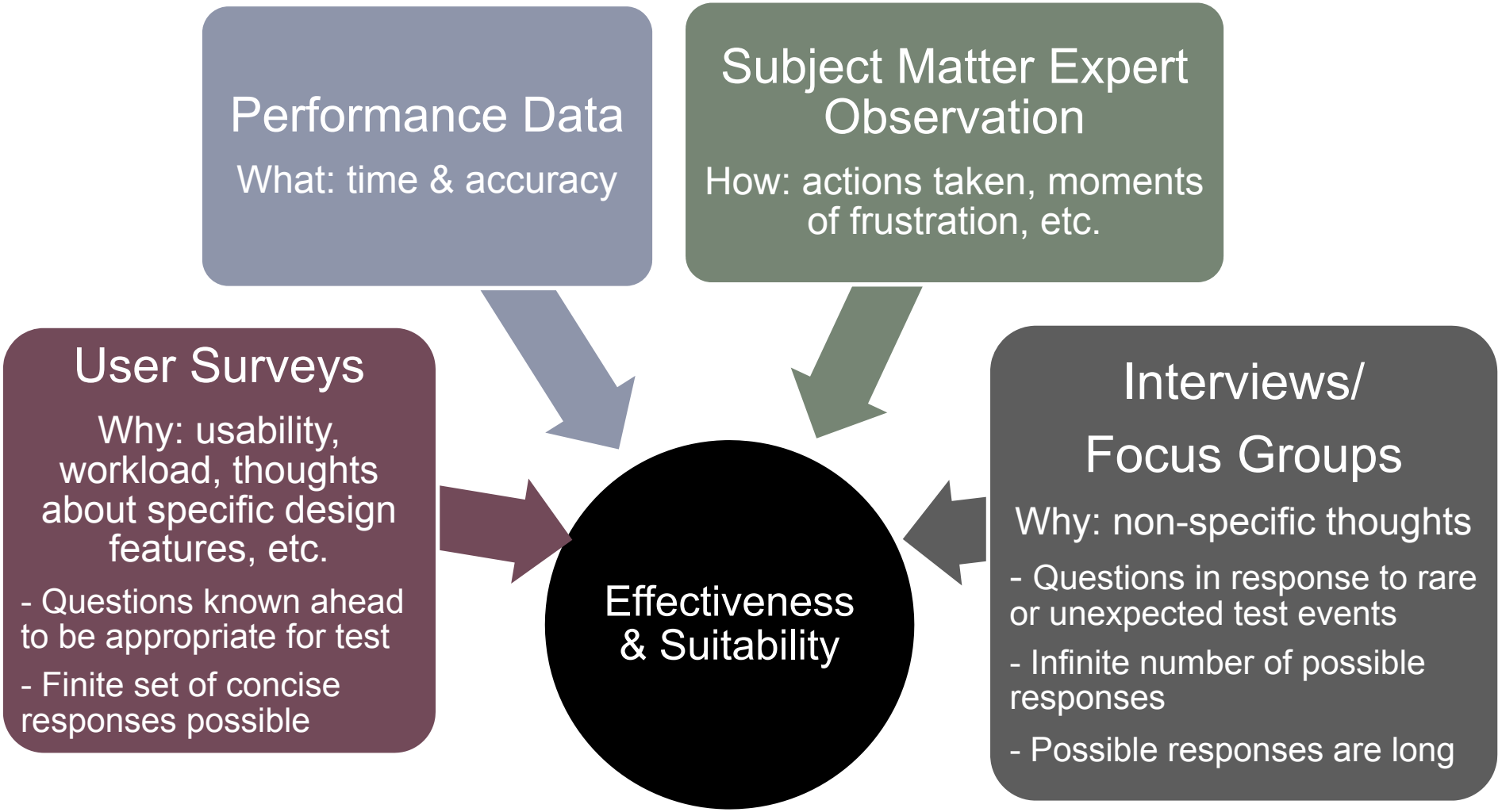
Focus Groups

Collect non-specific diagnostic thoughts

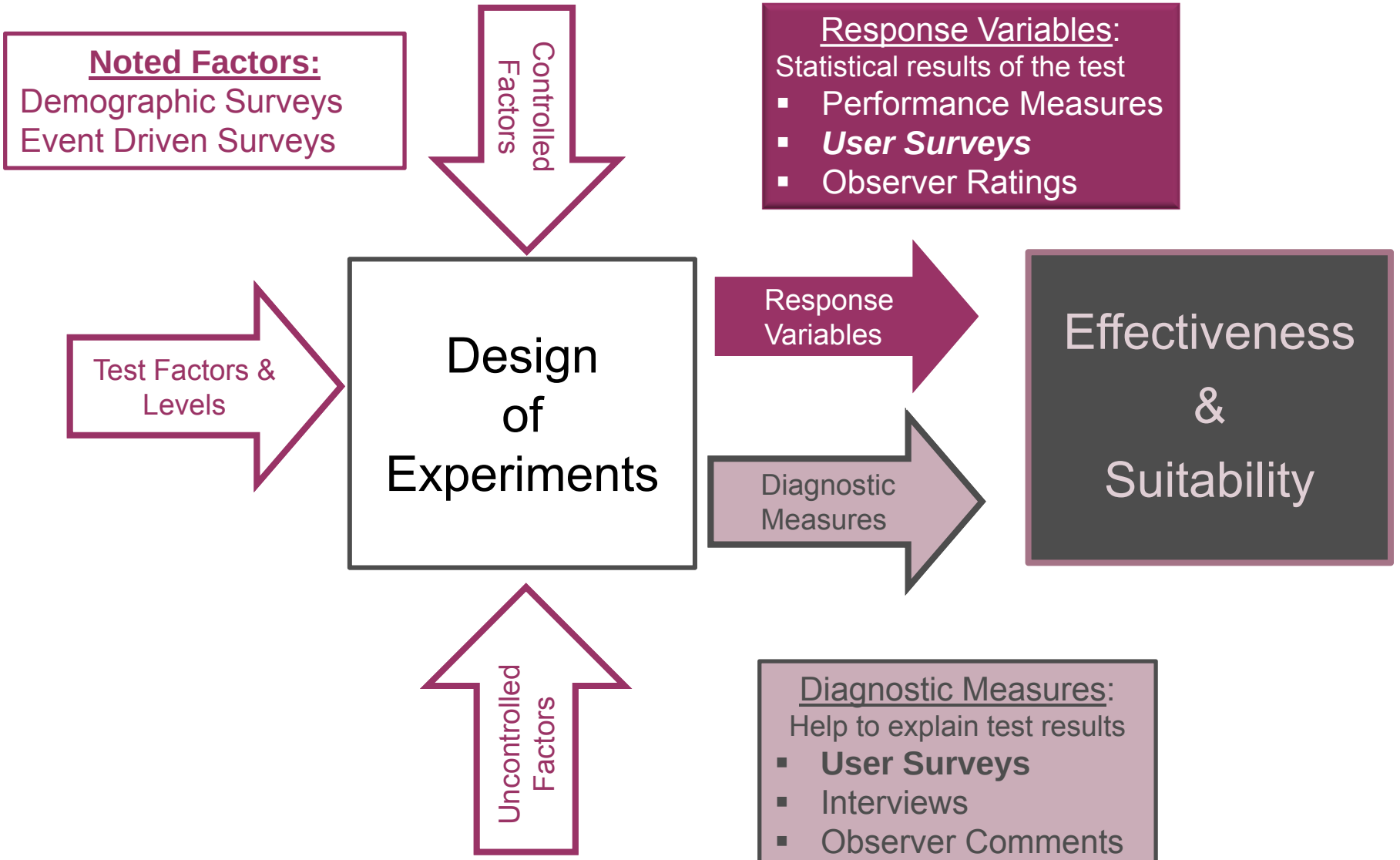
- *Similar to interviews, but strong belief group discussion adds value to assessment*
- *Group dynamics change responses – can* provide *creative solutions or biased responses*
- *Useful anecdotes and quotes as context for survey/performance data*

Different review standards depending on data collected

All Data Types Are Needed



User Surveys Have Many Roles In DOE



Surveys: Response Variable v. Diagnostic Measure

- **Surveys can be used as response variables or diagnostic measures**

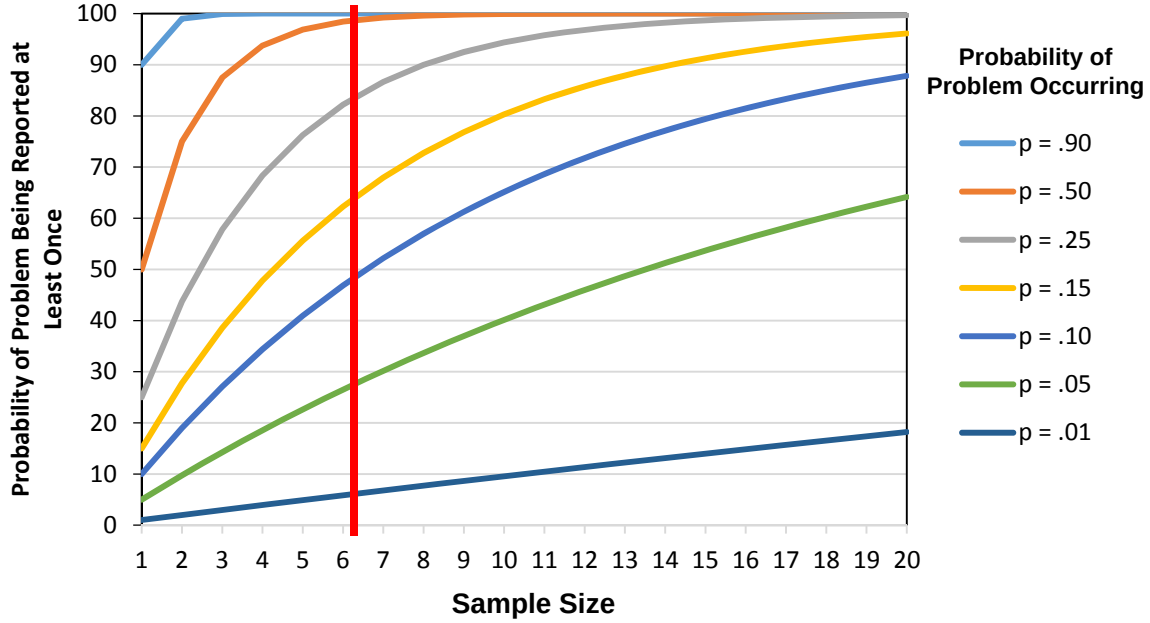
- **Questions to ask when evaluating if your survey question is a response variable:**
  - Will it be compared to a threshold?
    - » Is usability above a 70 on the SUS
  - Will it be used to make comparisons between different factors (e.g. day versus night or by mission type)?
    - » Is workload higher during night?
  - Do you need to make risk based statements about the question results in the report?
    - » X% of users said they would not like to use this system to accomplish the mission

- **Question may be diagnostic if:**
  - It will not be compared to a threshold
  - It will not be quantitatively rolled into a MOE
  - It will not be used to make comparisons across conditions
  - It will be used to provide feedback to the developer/test team regarding why response variable outcomes were observed

Sample Size

- **Most surveys in OT&E are diagnostic and are not response variables**
  - No requirement to create a DOE for these surveys
  - Sample size is a minimum of 6 participants per condition
    - » Refer to the figure below, probability of a problem being reported at least once based on likelihood of problem existing in overall population
    - » Sample size 6 provides high probability of detecting of problems with p>0.50 in population

- **If survey is a response variable, a power analysis and DOE is needed to determine sample size**

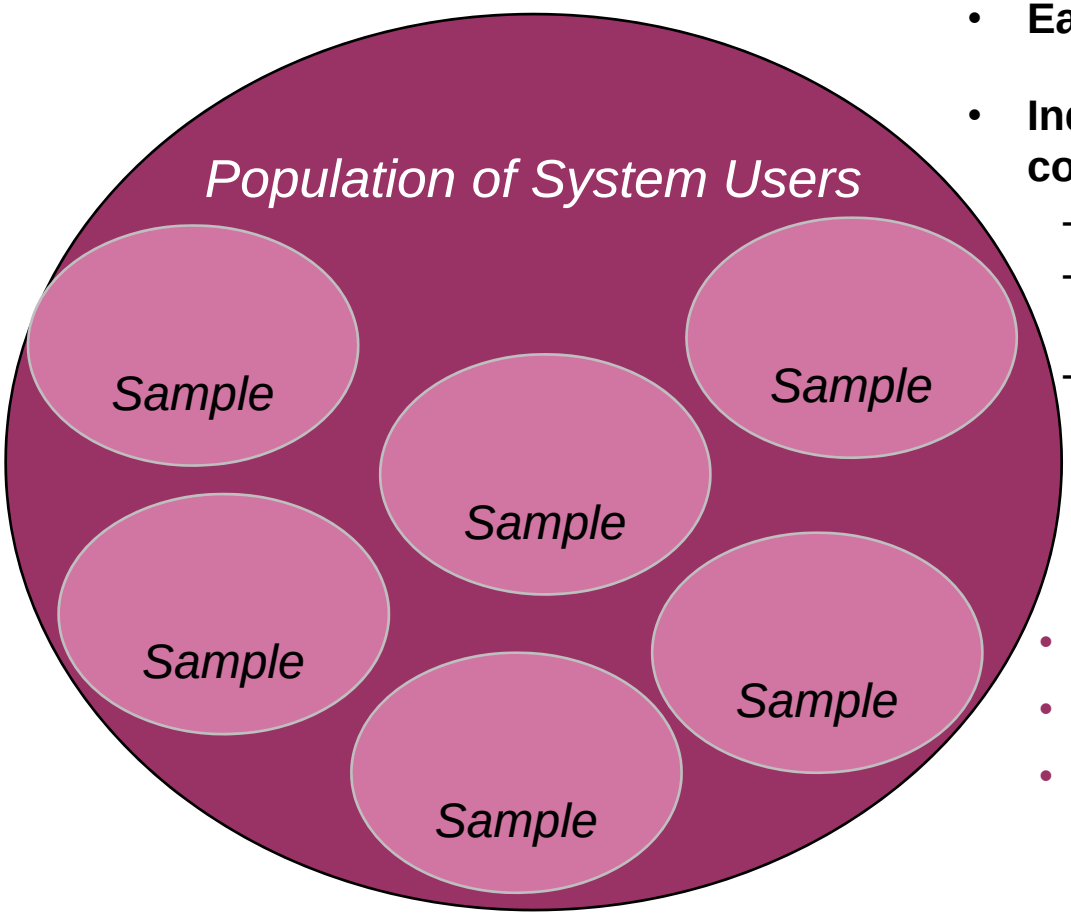


Condition *the aspect of the test the survey is intended to assess (e.g., system, task, demographics, time of day, weather,…)*

Rule of Thumb
Minimum 6 Participants per Condition

Importance of Sampling to Surveys

Sample: Members of a population participating in the test.



- **Each Individual in a population is unique**
- **Individual differences are important considerations**
  - Samples must represent entire population
  - Each participant must contribute equally to analysis!
  - Demographics may serve as factors

Sample Size Considerations

- **Effect size for practical significance**
- **Size/heterogeneity of population**
- **Experimental designs and statistical power/confidence are not required unless survey is primary response, which is rare**

When less than 6 respondents, treat as anecdotes

When to Administer Surveys

Post Test

After all activities completed

- Thoughts/feelings that will not change based on test factors or time
- Examples:
  - Overall satisfaction/ preference
  - User interface component satisfaction

Natural Break Points

Daily/ Post task

- Thoughts/ feelings that will change with time/ test factors:
  - Workload
  - Usability
  - Task specific questions
- Examples:
  - Ease of maintaining DIFFERENT components
  - Ease of operating during DIFFERENT mission types

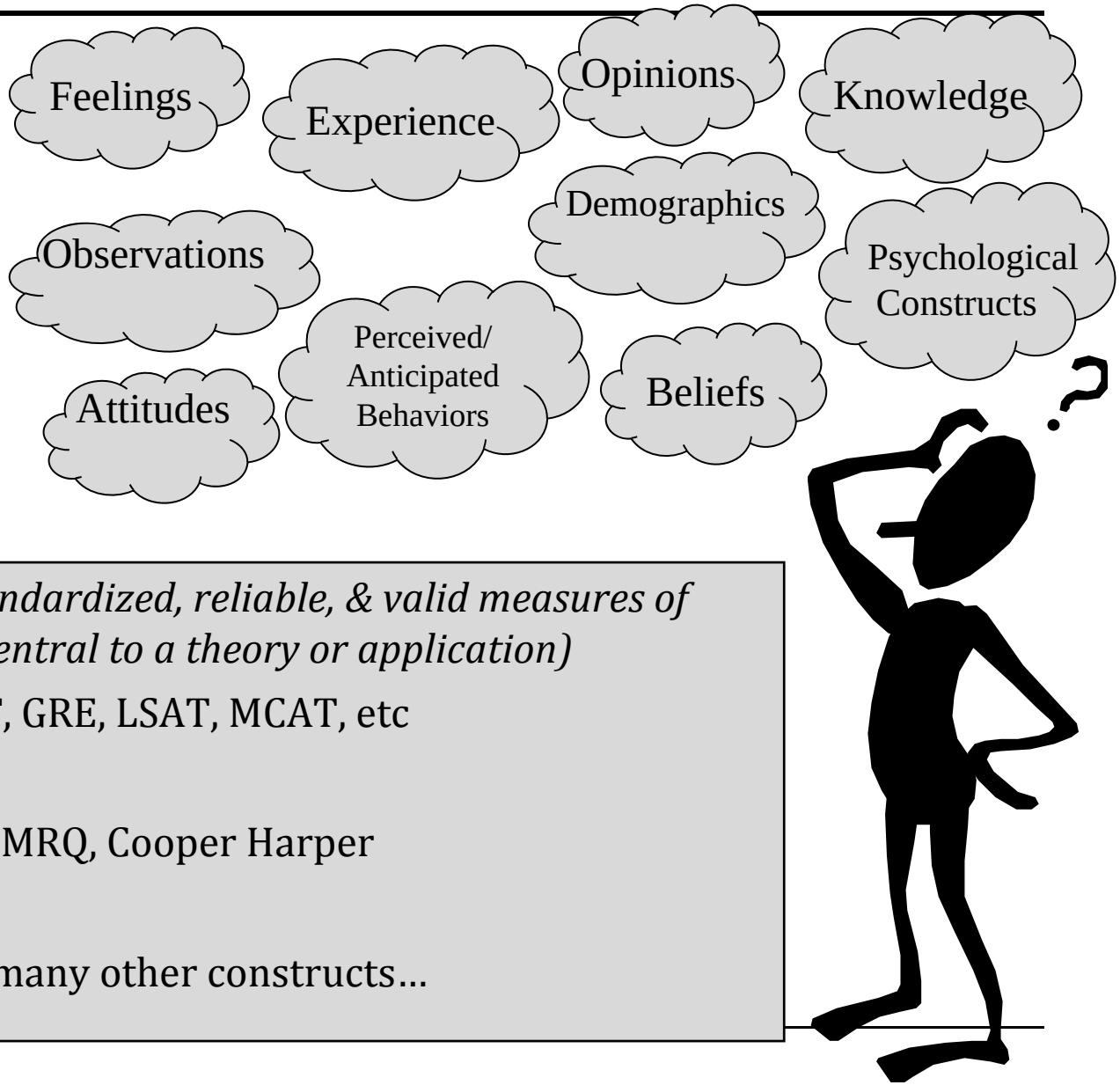
Event Driven

In response to unique occurrence

- Thoughts/ feelings about critical events (expected or unexpected)
  - Safety
  - Bugs
  - Uncommon tasks

Custom-Made & Empirically Vetted

<u>Custom-made surveys:</u> Developed for a specific purpose



<u>Empirically vetted surveys</u>: standardized, reliable, & valid measures of constructs (i.e., attributes central to a theory or application)

- Aptitude: ASVAB, ACT, SAT, GRE, LSAT, MCAT, etc
- Intelligence: WAIS
- Workload: NASA TLX, CSS, MRQ, Cooper Harper
- Usability: SUS
- Trust, Fatigue, Stress, and many other constructs…

What Is an Empirically Vetted Survey?

- **A good survey is one that is valid and reliable**
  - Reliability – are the measurements consistent?
  - Validity – actually measures what it is intended to measure
    - » Number of dimensions – different aspects of the thing being measured
    - » Sensitivity – effect size that can be reliably be detected

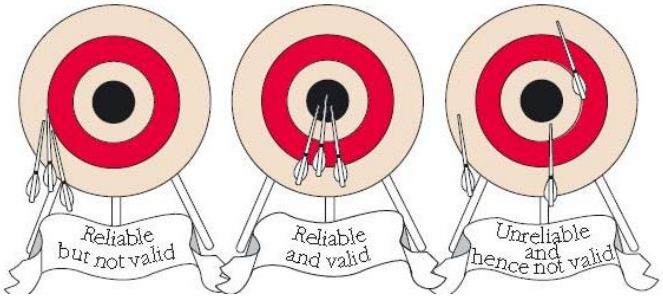


Figure 5.1 *Reliability and validity. (Source: Open University, 1979,* Classification and Measurement, *DE304, Block 5, The Open University, Milton Keynes, p. 68)*

- **An empirically vetted survey has been tested for these characteristics**
  - Example: NASA Task Load Index (NASA-TLX) [Hart 1988]
    - » Began with solid theoretical foundation
    - » Conducted tests to assess reliability & validity
      - Different workload conditions
      - Same participants & conditions at different times
      - Correlations to other valid workload measures
  - Later studies confirmed results [Nygren 1991, Matthews 2014]
    - » In an analysis across 5 Army programs, NASA-TLX was shown to have consistently higher validity than Modified Cooper-Harper, Subjective Workload Assessment Technique (SWAT), or Overall Workload (OW) rating scales [Hill 1992]

Benefits of Empirically Vetted Surveys

- **Known validity, reliability,**
- **General comparative ability (i.e., standardized)**
  - Some surveys are used widely enough to make general comparisons
  - NASA-TLX, SUS [Grier 2014, Bangor 2008 and 2009]
- **Specific comparative ability - all empirical surveys**
  - Results in one test can be compared to results in another test
  - Requires that task, environment, demographics, other factors be similar
- **Can analyze parametrically**
  - NASA-TLX, CSS, SUS, and Likert-type questions have all been shown to behave like interval measurements
- **When used in an well designed test, empirically vetted surveys can be used to discriminate between factors**
- **When used in conjunction with performance measures, all surveys can be used to diagnose performance results**

When Is a Custom-Made Survey Appropriate?

Appropriate

1. **There isn't an appropriate empirical survey**
2. **Measure specific user/maintainer thoughts**
   - Specific features/ components
   - Specific issues with regard to CONOPS
3. **Quantify observer ratings**
   - Special case

Not Appropriate

1. **Non-specific information sought from respondents**
   - Interview
2. **Measure performance**
   - Time and accuracy via appropriate physical measure
3. **Measure requirements**
   - Appropriate physical measure
4. **Measure situation awareness**
   - Numerous techniques in human factors literature
   - Salmon et al (2006) for review

Make Everything As Simple As Possible, But Not Simpler
– Albert Einstein

Custom-Made Surveys

- **Useful for measuring known or suspected problem areas**
  - Requires careful wording, formatting, and consideration of overall structure of survey
  - Take time to create, recommend test runs
  - Range from Likert questions for statistical analysis…
    - » Overall, the system was easy to maintain. Strongly Agree – Strongly Disagree
  - To simple open ended questions for diagnostic data
    - » What problems did you encounter today?

- **Choose questions based on desired analysis technique**
  - More questions don't necessarily find more problems
    - » Adds to respondent burden, reduces validity of data
  - Likert questions useful for statistical analysis
    - » Identify a statistically significant difference between subsystem component A and component B
  - Open-ended questions aid problem discovery, finding unknown-unknowns
    - » For example, to identify if a subsystem has a problem

- **No outside comparative ability**
  - No standardized scale to compare against – can't say "good" or "fair"
  - Can compare systems within a test – system A is <better/worse> than system B

Custom-made surveys require a lot of thought & preparation

Surveys in TEMPs &Test Plans

- **TEMP should indicate which COIs will be assessed using surveys, focus groups, or interviews**

- **Test Plans should contain the following information**
  - Specific survey
    - » Empirically Vetted
    - » Custom-Made
  - When will the survey be administered
    - » At break points, end of test, event driven?
  - The goal of the survey
    - » Why is it important to collect these data at these points in the test?
  - How the survey will be administered
    - » Verbally, Electronically, Paper
  - Who will complete the survey
    - » Which users/maintainers in the test?
  - How will data be vetted, stored, & analyzed

Summary

- **What Is a survey?**
  - Measure of thoughts of a population
- **How is a survey different from a data sheet? Interview? Focus group?**
  - **Data sheet:** quantitative & static measure of observables
  - **Survey:** quantitative & static measure of thoughts
  - **Interview:** qualitative & flexible collection of thoughts
  - **Focus group:** group interview
- **What role do surveys have in DOE?**
  - Noted factors
  - Response variables
  - Diagnostic measures
- **Are there different kinds of surveys?**
  - Demographic
  - Event driven /Break point / End of test
  - Empirical /Custom-made
- **How are surveys incorporated in TEMPs and Test Plans**

IDA

Up Next

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

Backup

Surveys Measure *Thoughts* about Performance Only

- **Not Time:**

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

  - Albert Einstein

- **Not Accuracy:**

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

Bad Design = Mismatch Between Truth & Belief

3 Mile Island

Vincennes Incident

- **Not Situation Awareness:**

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