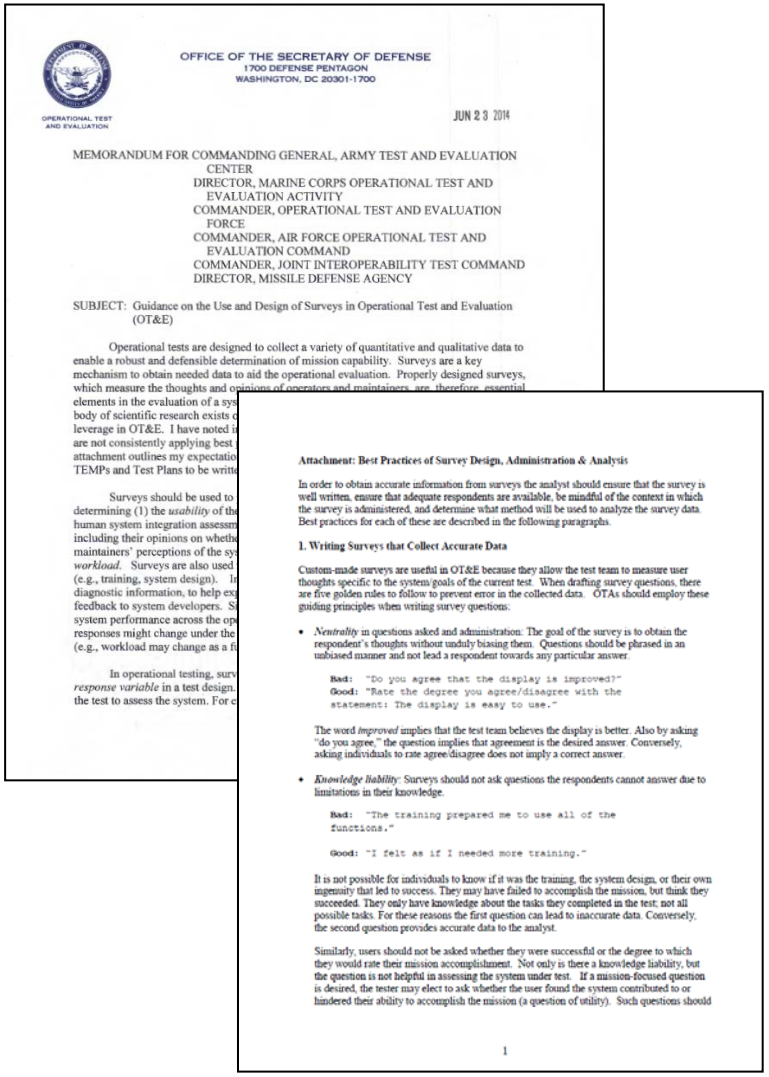
IDA

Surveys in Test & Evaluation: The Good, The Bad, & The Ugly

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

DOT&E Guidance on Surveys

- **Surveys are an important aspect of DOT&E evaluation of effectiveness and suitability**
- **Surveys are appropriate for quantitatively measuring operator and maintainer thoughts and opinions**
- **Use surveys only when appropriate**
- **Employ best practices for writing and administering surveys**
  - Memo provides a best practices guide attachment



OFFICE OF THE SECRETARY OF DEFENSE
1700 DEFENSE PENTAGON
WASHINGTON, DC 20301-1700

OPERATIONAL TEST AND EVALUATION

JUN 23 2014

MEMORANDUM FOR COMMANDING GENERAL, ARMY TEST AND EVALUATION CENTER
DIRECTOR, MARINE CORPS OPERATIONAL TEST AND EVALUATION ACTIVITY
COMMANDER, OPERATIONAL TEST AND EVALUATION FORCE
COMMANDER, AIR FORCE OPERATIONAL TEST AND EVALUATION COMMAND
COMMANDER, JOINT INTEROPERABILITY TEST COMMAND
DIRECTOR, MISSILE DEFENSE AGENCY

SUBJECT: Guidance on the Use and Design of Surveys in Operational Test and Evaluation (OT&E)

Operational tests are designed to collect a variety of quantitative and qualitative data to enable a robust and defensible determination of mission capability. Surveys are a key mechanism to obtain needed data to aid the operational evaluation. Properly designed surveys, which measure the thoughts and opinions of operators and maintainers, are, therefore, essential elements in the evaluation of a sys
body of scientific research exists
leverage in OT&E. I have noted i
are not consistently applying best
attachment outlines my expectatio
TEMPs and Test Plans to be writt

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

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

Attachment: Best Practices of Survey Design, Administration & Analysis

In order to obtain accurate information from surveys the analyst should ensure that the survey is well written, ensure that adequate respondents are available, be mindful of the context in which the survey is administered, and determine what method will be used to analyze the survey data. Best practices for each of these are described in the following paragraphs.

1. Writing Surveys that Collect Accurate Data

Custom-made surveys are useful in OT&E because they allow the test team to measure user thoughts specific to the system/goals of the current test. When drafting survey questions, there are five golden rules to follow to prevent error in the collected data. OTAs should employ these guiding principles when writing survey questions:

- *Neutrality* in questions asked and administration: The goal of the survey is to obtain the respondent's thoughts without unduly biasing them. Questions should be phrased in an unbiased manner and not lead a respondent towards any particular answer.

  Bad: "Do you agree that the display is improved?"
  Good: "Rate the degree you agree/disagree with the statement: The display is easy to use."

  The word *improved* implies that the test team believes the display is better. Also by asking "do you agree," the question implies that agreement is the desired answer. Conversely, asking individuals to rate agree/disagree does not imply a correct answer.

- *Knowledge liability*: Surveys should not ask questions the respondents cannot answer due to limitations in their knowledge.

  Bad: "The training prepared me to use all of the functions."

  Good: "I felt as if I needed more training."

  It is not possible for individuals to know if it was the training, the system design, or their own ingenuity that led to success. They may have failed to accomplish the mission, but think they succeeded. They only have knowledge about the tasks they completed in the test; not all possible tasks. For these reasons the first question can lead to inaccurate data. Conversely, the second question provides accurate data to the analyst.

  Similarly, users should not be asked whether they were successful or the degree to which they would rate their mission accomplishment. Not only is there a knowledge liability, but the question is not helpful in assessing the system under test. If a mission-focused question is desired, the tester may elect to ask whether the user found the system contributed to or hindered their ability to accomplish the mission (a question of utility). Such questions should

1

What Will Be Covered

- What can't be assessed by survey
- Some best practices for writing questions

What Won't Be Covered

- Selecting academic surveys
- Formatting surveys
- Selecting response option type
- DOE with surveys
- Analysis of survey data
- Interviewing techniques

IDA | What is a Survey?

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

What do you see?

Thoughts are relative;
Requirements are not!

~~Physical~~
System
Requirements

IDA

Examples Where Measurement by Survey Is Not Appropriate

- **"The ship has protective clothing for every crew member."**
  - Count the protective clothing & compare to number of crew.
- **"Engine exhaust levels in the mission bay do not exceed safety limits ..."**
  - Measure exhaust levels with Portable Emissions Measurement System (PEMS) & compare to safety limits
- **"Temperatures in primary work spaces were adequate."**
  - Measure by a thermometer & compare to requirements in MIL-STD-1472G
  - e.g., 5.5.2.1.4 part g: *Limited thermal tolerance zones. Where hard physical work is to be required for more than two hours, an environment not exceeding… 25 °C (77 °F) shall be provided. … shall be decreased 5.0 °C (9.0 °F) for complete chemical protective uniforms, 4.0 °C (7.0 °F) for intermediate clothing systems, and 3.0 °C (5.0 °F) for body armor.*

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

1. Collect specific data for a pre-defined purpose according to rules (i.e., not random thoughts).

2. The data collection rules determine the validity of the data & the statistical analyses possible.

Who	Role	Sources of Error
Commissioner	Defines Survey's Purpose & Uses Information from Survey	• Not Enough Information • Wrong Information
Participant/ Respondent	Gives Data	• Answers Different Question • Thinks Too Much • Doesn't Think Enough
Analyst	Translates Data Into Information	• Unable to analyze data • Data Aggregation

IDA | Question Writing Best Practices

Identifier Item Response
1. How many parts of a question are there?
1 2 3 4 5

Knowledge Liability: Respondents Have Enough Information to Answer the Question

Singularity: Only 1 Idea Per Item

User Friendly: Items Do Not Require a Lot of Thought or Interpretation (e.g., short, clear, specific)

Neutrality: Items Do Not Imply Value Judgments
Items Are Not Emotionally Charged

Independence: Responses Will Not Affect Responses to Other Questions (Branching, Redundancy, etc…)
- Especially important when aggregating responses
- Item Response Theory & Internal Reliability

Available response options match items

All possible response options

IDA

1. The system is efficiently reliable.

User Friendly

Unclear what the goal of the question is.
Therefore no recommended rewording.

IDA

2. Rate the overall ability of the system to provide situation awareness.

Not Adequate Adequate DK/NA

2b. If not adequate, rate the degree to which this deficiency degrades effectiveness.

Very Low Low Moderate High Very High

IDA

2. Rate the overall ability of the system to provide situation awareness.

Not Adequate Adequate DK/NA

2b. If not adequate, rate the degree to which this deficiency degrades effectiveness.

Very Low Low Moderate High Very High

Knowledge Liability

IDA

2. Rate the overall ability of the system to provide situation awareness.

Not Adequate Adequate DK/NA

2b. If not adequate, rate the degree to which this deficiency degrades effectiveness.

Very Low Low Moderate High Very High

Knowledge Liability

Independence

IDA

2. Rate the overall ability of the system to provide situation awareness.

Not Adequate Adequate DK/NA

2b. If not adequate, rate the degree to which this deficiency degrades effectiveness.

Very Low Low Moderate High Very High

Knowledge Liability

Independence

All Possible Responses

6. I trust the information provided by the system.

Strongly Agree Somewhat Agree Slightly Agree Slightly Agree Somewhat Disagree Strongly Disagree

IDA

3. Is the training materials complete?

IDA

3. Is the training materials complete?

Knowledge Liability

User Friendly – grammar

5. I felt as if I needed more training.

4. Rate the acceptability of system's Launch Acceptability Region (LAR) displays provided to support accurate and timely system employment.

IDA

4. Rate the acceptability of system's Launch Acceptability Region (LAR) displays provided to support accurate and timely system employment.

Knowledge Liability

4. Rate the acceptability of system's Launch Acceptability Region (LAR) displays provided to support accurate and timely system employment.

Knowledge Liability

Singularity

4. Rate the acceptability of system's Launch Acceptability Region (LAR) displays provided to support accurate and timely system employment.

Knowledge Liability

Singularity

User Friendly

(repeated and unnecessary words)

3. The Launch Acceptability Region (LAR) displays are helpful.

IDA

5. The SSO functionality increases my productive time within the clinic by reducing the amount of time I spend logging into different applications when documenting the healthcare provided.

Strongly Agree	Somewhat Agree	Slightly Agree	Slightly Agree	Somewhat Disagree	Strongly Disagree

IDA

5. The SSO functionality increases my productive time within the clinic by reducing the amount of time I spend logging into different applications when documenting the healthcare provided.

Neutrality

User Friendly – 27 words!

4. The SSO function is useful.

Strongly Agree	Somewhat Agree	Slightly Agree	Slightly Agree	Somewhat Disagree	Strongly Disagree

IDA

6. Rate the overall usefulness of the report: Accuracy

Completely unacceptable	Largely unacceptable	Somewhat unacceptable	Somewhat acceptable	Largely acceptable	Completely acceptable

IDA

6. Rate the overall usefulness of the report:

Accuracy

Completely unacceptable	Largely unacceptable	Somewhat unacceptable	Somewhat acceptable	Largely acceptable	Completely acceptable

Knowledge Liability

Response Options Don't Match Item

User Friendly

(What is being rated: report acceptability, usefulness, or accuracy?)

5. The content of the report is useful.

Strongly Agree	Somewhat Agree	Slightly Agree	Slightly Agree	Somewhat Disagree	Strongly Disagree

IDA

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

IDA

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

Independence

User Friendly

(It is the analyst's job to rate the acceptability of the system)

8. I would like to use this system to accomplish the mission.

IDA | Some Words to Avoid In Writing Surveys

Knowledge Liability Singularity User Friendly Neutrality Independence

- Accurate
- Timely
- Situation Awareness
- Effective
- Efficient
- n/a
- And
- Each
- All
- Never
- None
- Better
- Easier
- Improved
- Based on
- If
- Considering

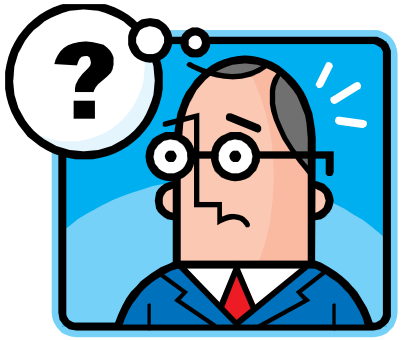




Questions?







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

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

Knowledge Liability

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

Knowledge Liability

Singularity

User Friendly – 22 words!

2. I trusted the system's engagement decisions.

IDA

7. The amount and type of training provided to the X position allowed them to employ the system.

IDA

7. The amount and type of training provided to the X position allowed them to employ the system.

Knowledge Liability

7. The amount and type of training provided to the X position allowed them to employ the system.

Knowledge Liability

Singularity

7. I felt as if I needed more training.