IDA

Surveys in Operational Test & Evaluation

Rebecca A. Grier, Ph.D.

Institute for Defense Analyses
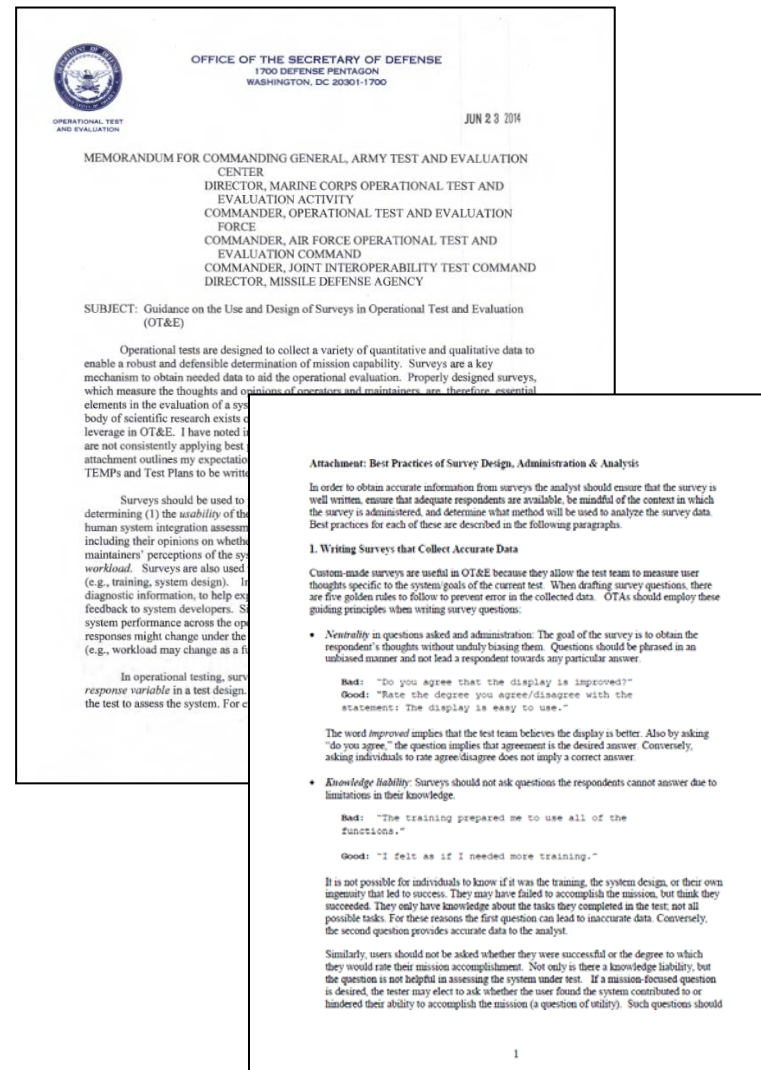
DOT&E Guidance on Surveys
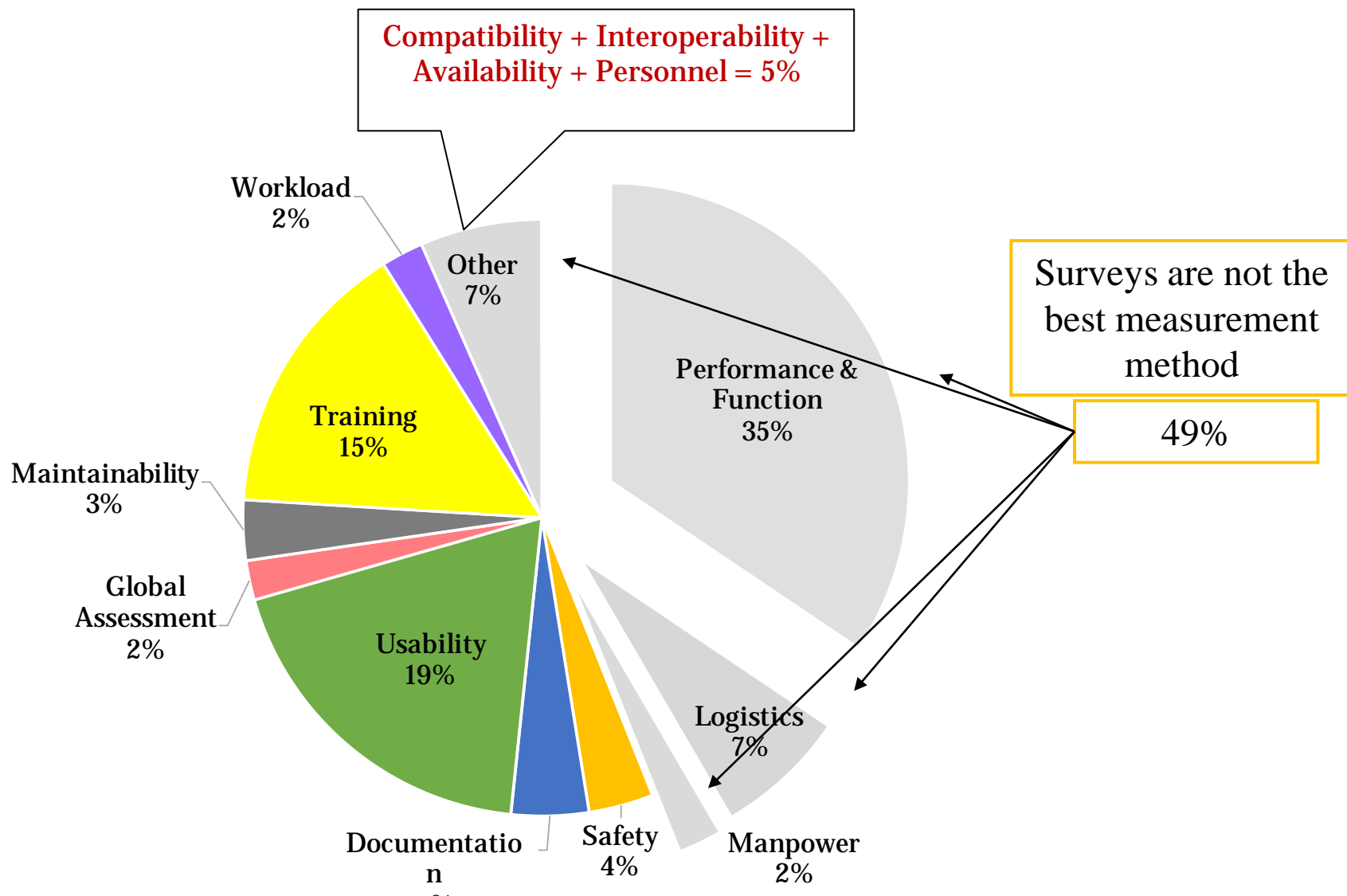## June 2014

IDA

- **Surveys are an important aspect of DOT&E evaluation**

- **Surveys should be used to (determine)**
  - the *usability* of the system
  - the operators' thoughts of the system's *utility*
  - maintainers' thoughts of the system's *maintainability*
  - the effects of system design on *workload*

- **Academically-established surveys should be used for *human factors* constructs**

- **Use surveys only when appropriate**

- **It is essential to understand the goal of why you are conducting the survey**

- **Employ best practices for writing and administering surveys**
  - Memo provides a best practices guide attachment

Review of OT&E Surveys: Percentage Appropriate Questions

Compatibility + Interoperability + Availability + Personnel = 5%

Surveys are not the best measurement method

49%

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

Surveys Measure _Thoughts_ about Performance Only

IDA

- ## 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:

3 Mile Island

Truth		
	Success	Failure
Belief Success	🙂	**!!!**
Belief Failure	☹️	😐

Bad Design = Mismatch Between Truth & Belief

Vincennes Incident

- ## Not Situation Awareness:

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

Ex: *The GPS was accurate.*

Ex: *I felt as if I needed more training*

Ex: *I would like to use this system to accomplish the mission.*

Ex: Rate the adequacy of land-based administrative support for supply support

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

- **I would like to use this system to accomplish the mission.**

- **The instructor presented the material clearly.**

- **I feel as though additional training is needed.**

- **The _(e.g., work station, cockpit)_ is well organized.**

- **I did not have the information needed to __(e.g., execute the mission, perform a specific task)__.**

- **It was difficult to _(e.g., perform a specific task)_.**

- **_(e.g., Equipment, Controls, Information, Features, Applications)_ are easily accessible.**

- **Are there any improvements that you would make to the system?**

- **Please comment on any safety concerns that you have.**

Surveys Are An Important Aspect of DOT&E

IDA

Performance Data

What: time & accuracy

Subject Matter Expert Observation

How: actions taken, moments of frustration, etc.

User Surveys

Why: usability, workload, thoughts about specific design features, etc.

- Questions known ahead to be appropriate for test

- Finite set of concise responses possible

Effectiveness & Suitability

User Interviews

Why: non-specific thoughts

- Questions in response to rare or unexpected test events

- Infinite number of possible responses

- Possible responses are long

When to Design A Survey

IDA

Appropriate

1. **There Isn't an Appropriate Academically-Established Survey**

2. **Measure Specific User/Maintainer Thoughts**
   – Utility/Ease
   – Specific features/ components
   – Specific issues with regard to CONOPS

3. **Quantify Observer Ratings**

> "A good plan is like a road map: it shows the final destination and usually the best way to get there."
> *H. Stanely Judd*

Not Appropriate

1. **Obtain Random Thoughts of Respondents**
   – Interview

2. **Measure Performance**
   – Time
   – Accuracy via Appropriate Physical Measure
   – Observers

3. **Measure Requirements**
   – Appropriate Physical Measure
   – See e.g., MIL-STD-1472G

4. **Measure Situation Awareness**
   – Numerous techniques in Human Factors Literature
   – Salmon et al (2006) for review

Some Common Self Report Workload Measures

IDA

Measure	Published	Citations	Description

Cooper Harper & Variants
  - Modified Cooper Harper (1992)
  - Bedford (1990)
	1969	2036	1 -3 Questions
			Score: (good)1-10 (bad)
			High workload: 4
			One-dimensional/Not Diagnostic
			Task Relative
			No Theory

Crew Status Survey/ Integrated Workload Scale	1993/2005	26/63	1 Question Score: (good) 0 -7/9 (bad) High Workload: ???? Uni-dimensional/Not Diagnostic Task Agnostic No Theory

NASA-TLX
  - Original/Weighted
  - RawTLX (RTLX)/ Unweighted
	1988	7020	6 or 21 Questions
			Score: (good) 0 -100 (bad)
			High workload: ?????
			Multi-dimensional/ Diagnostic
			Task Agnostic
			Resource Pool Theory

MRQ	2001/2007	217	Up to 17 Questions Score: (good) 0 -100 (bad) High workload: ?????? Multi-dimensional/Diagnostic Task Agnostic Multiple Resource Theory

IDA

- **Most Used Usability Survey**
  - 43% of usability studies
  - Sauro & Lewis (2009)

- **10 Questions**
  - 5 point alternating Likert response
  - Administered immediately after user completes tasks
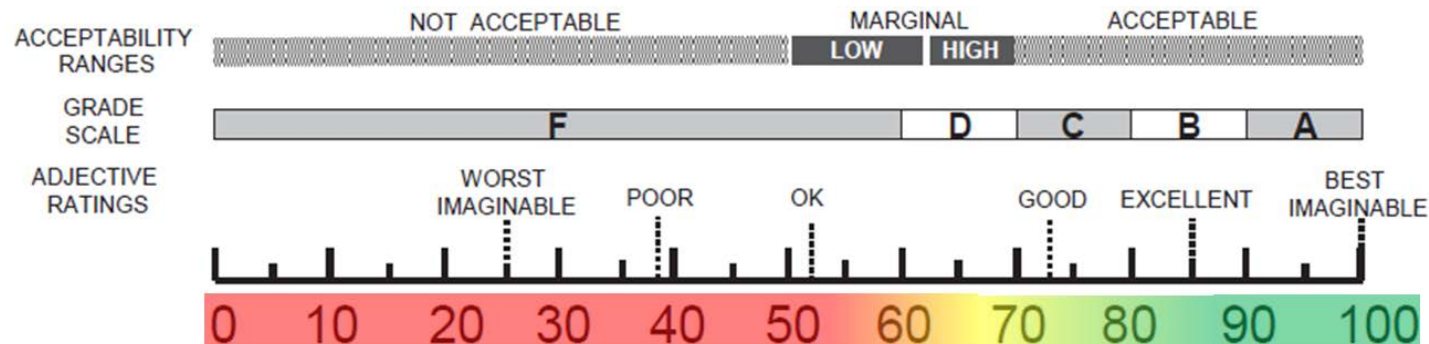
- **Score: (bad)0 – 100(good)**
  - Subtract 1 from each odd question
  - Subtract each even question from 5
  - Multiply the sum of above by 2.5
  - 2.5 [20+Q1 + Q3 + Q5 + Q7 + Q9 - Q2 - Q4 - Q6 - Q8 – Q10]

	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

Conclusions

- **HSI is an important component of Operational Test & Evaluation**

- **All measurement should be done with a goal in mind and according to best practices**

- **Academically vetted surveys tell the test team about HSI constructs**
  - **Usability:** are there likely to be critical errors in operational context?
  - **Workload:** how much effort is required to achieve performance level?

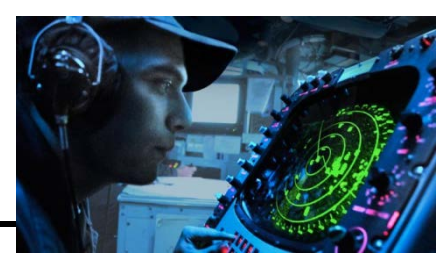- **Situation Awareness should not be measured via survey**

Further Reliability & Validity Assessments of SUS

IDA

- **Tullis & Stetson (2004)**
  - Compared SUS to other usability surveys
  - More accurate conclusions with smaller sample sizes

- **Bangor, Kortum, & Miller (2008)**
  - 2324 tests over 10 years wide range of systems
  - High internal consistency (r = 0.91)
  - Correlated to user-friendliness rating (r = 0.806)
  - Sensitive to usability differences

- **Lewis & Sauro (2009) & Borsci et al (2009)**
  - Two Interdependent Factors
    » Usability (Items 1, 2, 3, 5, 6, 7, 8, & 9)
    » Learnability (Items 4 & 10)

ACCEPTABILITY RANGES	NOT ACCEPTABLE		MARGINAL		ACCEPTABLE
			LOW	HIGH	

| GRADE SCALE | F | D | C | B | A |

ADJECTIVE RATINGS: WORST IMAGINABLE   POOR   OK   GOOD   EXCELLENT   BEST IMAGINABLE

0   10   20   30   40   50   60   70   80   90   100

Recommended Modifications to SUS

- **Learnability** (items 4 & 10)
    - Key Component of HSI
    - Key Component of Effectiveness
    - Key Component of Suitability

- **Slight Modifications to Text Suggested for Military Operators**
    - Item 1: Military missions are not frequent
    - Item 7: Clarify baseline

- **User Sophistication is a Test Design Issue**

  ISO: "The extent to which a product can be used by <u>specified users</u> to achieve <u>specified goals</u> with effectiveness, efficiency & satisfaction in a <u>specified context</u> of use."

  Effective: "<u>mission accomplishment</u> when used by <u>representative personnel</u> in the (<u>expected environment</u>) …considering organization, <u>training</u>…"

Recommended Military SUS

1. I think that I would like to use this system ~~frequently~~ *to accomplish the mission*.

2. I found the system unnecessarily complex

3. I thought the system was easy to use

4. I think that I would need the support of a technical person to be able to use this system

5. I found the various functions in this system were well integrated

6. I thought there was too much inconsistency in this system

7. I would imagine that most people *with my MOS* would learn to use this system very quickly

8. I found the system very awkward to use

9. I felt very confident using the system

10. I needed to learn a lot of things before I could get going with this system.

Case Study: DSL Self Installation

•Truck Roll
•Splitter in NID
•Bridge modem

•Multiple modems
•Splitterless
•4 manuals
•2 CDs, 2 disks

•Multiple modems
•Splitterless
•5 manuals
•2 CDs, 1 disk

•1 Ethernet modem
•Splitterless
•1 manual
•1 CD
•Color coded cables

•1 Ethernet modem
•Splitterless
•2 manuals
•1 CD
•Color coded cables

•1 Ethernet modem
•Splitterless
•2 manuals
•1 CD

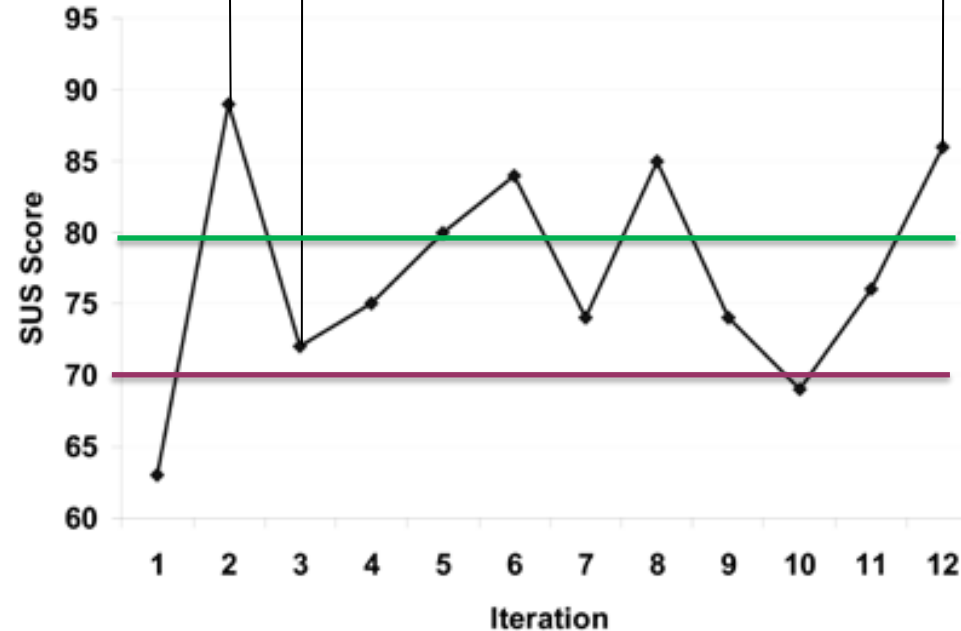•1 Ethernet modem
•Splitterless
•1 manual
•1 CD
•No Ethernet card

•New 4 port Ethernet modem
•Splitterless
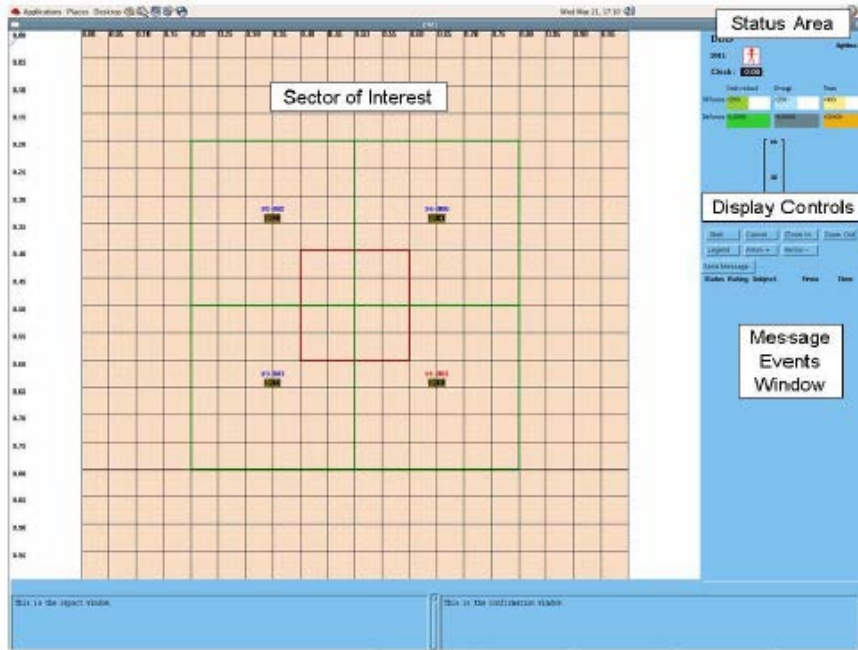•1 manual
•No CD

?

95% Success in the Lab

90% Install Ethernet Card
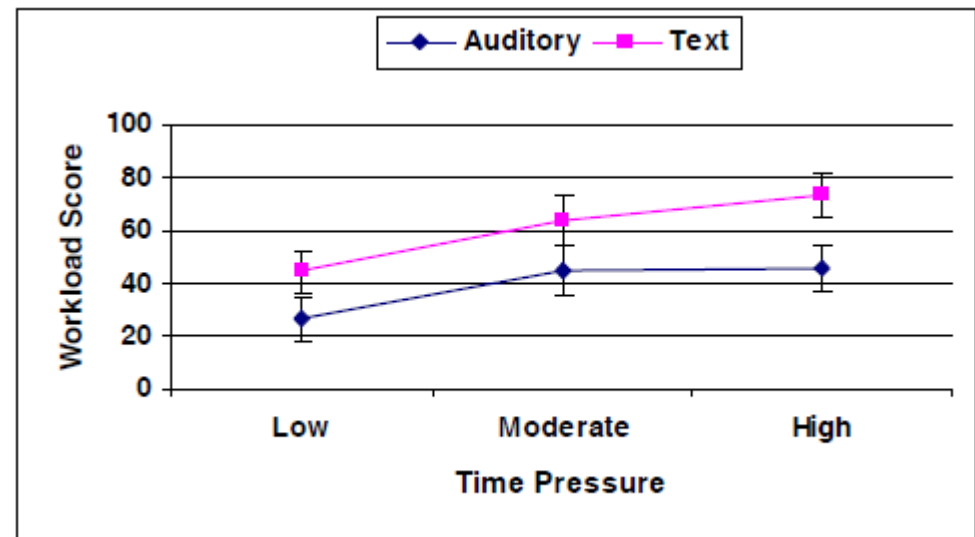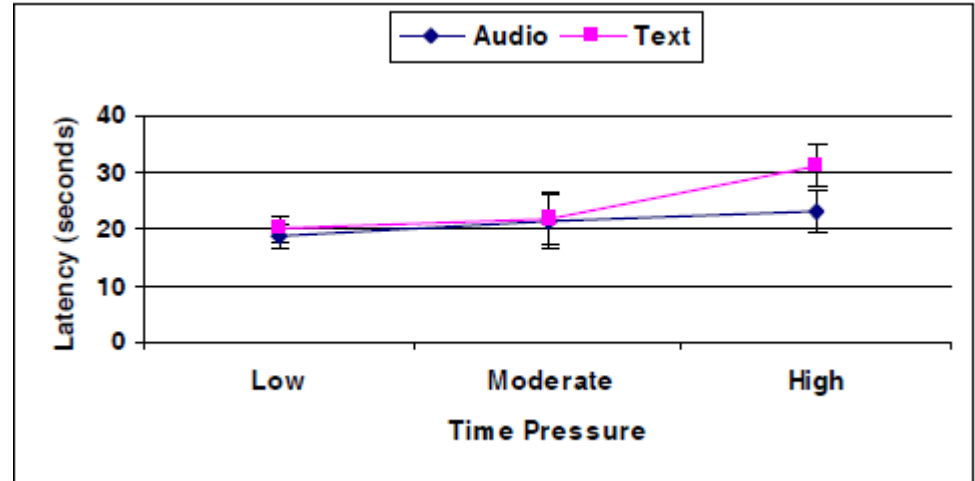
New Modems Introduced

Kortum, P., Grier, R. & Sullivan, M. (2009). DSL Self-installation: From Impossibility to Ubiquity. *Interfaces, 80*, 12-14.

Using NASA TLX to Compare Versions: Value of Multi-Modal System to C^2

Grier, R.A., Parasuraman, R., Entin, E., Bailey, N., & Stelzer, E. (2008). A test of intra- versus inter-modality interference as a function of time pressure in a warfighting simulation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting in New York City.*

IDA

Questions?