# Measures and Metrics: The Need for Consistency in HSI Measurement Terminology

LTC STONEY TRENT, PH.D. USCYBERCOM

ROBERT HOFFMAN, PH.D. INSTITUTE FOR HUMAN MACHINE COGNITION

NDIA

2016 HUMAN SYSTEMS CONFERENCE

*Valuing Human Systems R&D: Enhancing Capabilities, Reducing Cost*

February 9-10, 2016 ▶ Springfield, VA ▶ www.NDIA.org/meetings/6350

# Agenda

- Measurement Pitfalls
- Measurement Theory
- HSI Measurement Example (Cyber Protection Team Technologies)
- Summary

# *Reliance on traditional “human performance measurement”*
# →
# *Failure to measure cognitive work at the systems level*

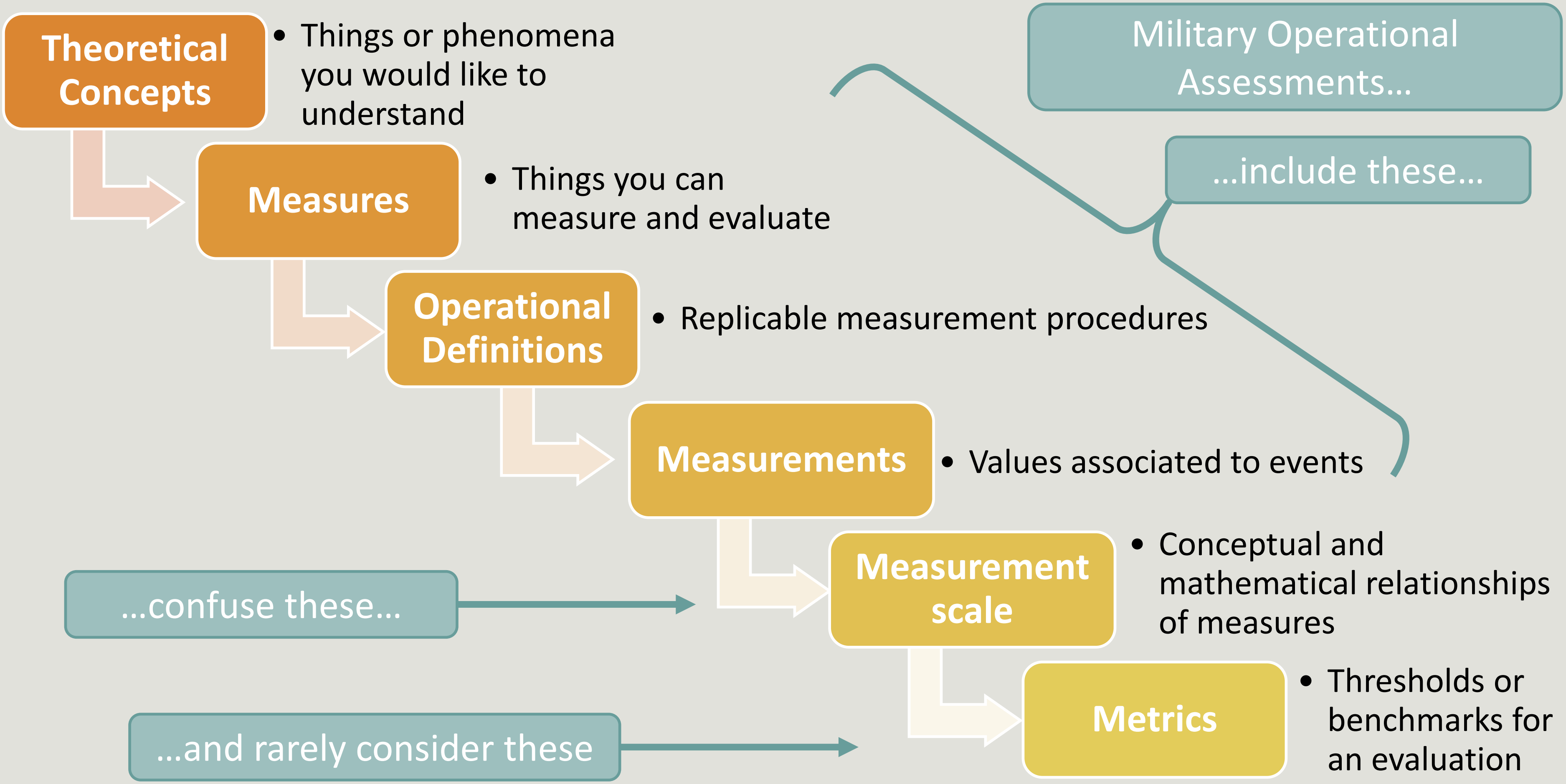
R&D Programs ask for systems that are “adaptive” or “resilient.”

OK. So how do we measure such things?

**Step 1:** Measure what can be easily measured. OK
**Step 2:** Disregard that which cannot be measured. Artificial and misleading.
**Step 3:** Presume the unmeasurable is not important. Blindness.
**Step 4:** Say the unmeasurable doesn’t exist. **Suicide.**

Daniel Yankovich, *Science*, 1977.

# *Measurement Terminology*



**Theoretical Concepts**

- Things or phenomena you would like to understand

**Measures**

- Things you can measure and evaluate

**Operational Definitions**

- Replicable measurement procedures

**Measurements**

- Values associated to events

**Measurement scale**

- Conceptual and mathematical relationships of measures

**Metrics**

- Thresholds or benchmarks for an evaluation

Military Operational Assessments…

…include these…

…confuse these…

…and rarely consider these

# *Measurement Scales*

**Qualitative (Nonparametric)**

- **Nominal –** Categories (Colors)
- **Ordinal –** Ordered Categories (Sequence)

**Quantitative (Parametric)**

- **Interval –** Meaningful distances (Time)
- **Ratio –** Absolute zero (Velocity)

Stevens (1946, 1951)

Quantitative scales can correlate to Qualitative scales

- Example: Scores of 85% correct or greater get an "A"

Parametric statistics should not be used with qualitative scales.

**Statistical significance** should not be confused with **practical significance**

# *Example:  Are you big enough to ride this roller coaster?*

**Theoretical Concepts:** Safety, Park insurability, Liability

**Measure:** Physical Stature

**Operational Definition:**  Height of child's head against a vertical ruler.

*ASSUMPTION: Height is the critical measure of stature.*

*ALL MEASURES UNDERGO INTERPRETATION*

**Measurement:**   Child stands next to a ruler

**Measurement Scale:**  Distance (inches)

At amusement parks the scale is often just a cut-out clown figure and in this case the measure **is** the metric.

*"If you are as tall as Puddles the Clown, you can ride this ride."*

**Metric:** Some minimum height.  If that height is met, the child rides theride.   If not, the child does not ride the ride

# Crucial Point: Metrics come from Policy.

Metrics do not come from the underlying science, the theory, the theoretical concepts, the measures, the measurement methodology, the measurement scales, or any of that.

Policy: Do not kill the customer or get sued.

**Metrics come from Policy. They do not magically spring from the measures or measurements.**

**Research sponsor is responsible for the policy.**

# Evaluating technologies for Cyber Protection Teams (CPTs)

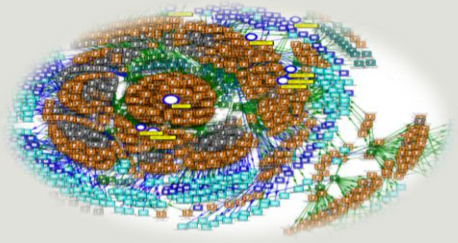
**CPT Mission:** Defend priority DoD networks and systems against priority threats

**Performance goal:** *Detect, characterize, and mitigate before any damage can be done?*

Perhaps, however consider the following:

- Decontextualization - Mitigation might reveal your capabilities to the attacker. You may not always want to do that.
- Reductive Thinking - This proposed metric is a raw performance measure.  It does not get at the “work system” level.

Conclusion:  The measurement of cognitive work system performance must involve the application of multiple measures.



# CPT Task:  Map a Cyberspace Network

**Critical Network Characteristics**

- Number/Type of devices on network
- Applications/Services/Operating Systems
- Physical/Logical Architecture
- Communication paths
- High value systems (e.g., servers, system admin devices)
- Open ports
- Roles of Devices (e.g., web server, domain controller, user workstation)

- External connections
- Directory service information (e.g., Lightweight Directory Access Protocol (LDAP))
- User privileges and roles
- Software configurations
- Router configurations
- Normal (and aberrant) traffic patterns

# Theoretical Concepts

- **Utility –** Does the tool help the team do the right things well?
- **Usability –** Does the tool work in the hands of real teams?
- **Acceptability –** Does the tool operate within the operational constraints of real teams?

# Measures and Metrics

| Theoretical Concept | Measure | Operational Definition | Metric |
|---|---|---|---|
| Utility | **Sufficiency** | Number of tasks completed with tool | 6 |
| | **Efficiency** | Time required to complete assigned tasks | <8 hours |
| | **Accuracy** | Completeness and correctness of survey data | 90% physical devices and paths enumerated |
| | **Data Integration** | Types of data used to make map | PCAP, Config files, Netflow, SNMP, ICMP |
| | **Transparency** | Ability to display what types of data were used | Yes |
| | **Map Richness** | Network attributes rendered on the map | All device types and physical routing |
| | **Exportability** | Formats possible for exporting data and products | Visio, Image, and Data |
| Usability | **User Feedback** | Ease of use/learnability | 60% positive |
| | **Map Interactivity** | Ability to explore and annotate the map | Both |
| | **Support to Job Learning** | Prompts for normative processes | Yes |
| | **Assistance Required** | User requests for help | 1/day |
| Acceptability | **Network Load** | Impact of network scans on the network | None |
| | **CPU Load** | CPU usage over time | TBD |

# Measurement Challenges and Issues

Avoiding Decontextualization by Using Multiple Measures

Avoiding Reductive Thinking and Promoting Systems-level Thinking

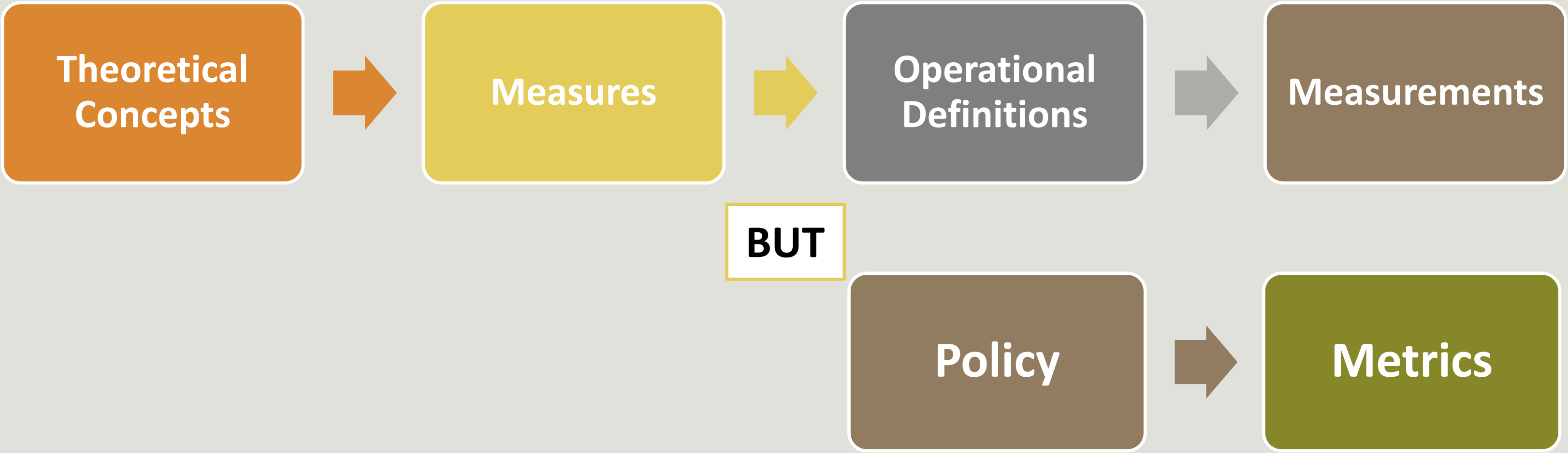
Example:  A new software support system improves performance on some task.

- Traditional Performance measurement would look only at HEAT measures: Hits, Errors, Accuracy, and Time
- This *faster-better-cheaper* techno-centric focus puts the worker in a “John Henry vs. the Steam Hammer” dilemma.
  - Worker feels like a slave to the machine.
- Does the software tool promote continued learning and expertise?
- Does it enhance worker intrinsic motivation?

# Systems-level Measurement

- Cognitive work systems must be usable, useful, understandable and observable. →Empirical evidence must accompany "deliverable."
- Measures must support:
  - Evaluation of hypotheses concerning the nature of the cognitive work (e.g., synchronous versus asynchronous communication, effects of team experience, etc.)
  - Evaluation of the software tools themselves
- Methodology:
  - Study work
  - Operationally relevant tasks and conditions
  - Representative users
  - Include developers in assessments
  - Be prepared to be surprised

# Summary



Theoretical Concepts

Measures

Operational Definitions

Measurements

BUT

Policy

Metrics

Metric - threshold for
making a valuation or decision

"Universal Metrics" do not exist, because decisions are context sensitive

See: Hoffman, R.R., Hancock, P.A., and Bradshaw, J.M. (2010, November/December). Universal Metrics? *IEEE Intelligent Systems*, pp. 93-97.

# Contacts and References

LTC Stoney Trent, Ph.D. satrent@cybercom.mil

Robert Hoffman, Ph.D. rhoffman@ihmc.us

Selected References

Chronbach, L. (1975). “Beyond the two disciplines of scientific psychology,” *American Psychologist, 30,* 116–127.

Hancock, P.A., Weaver, J.L. and Parasuraman, R. (2002). “Sans subjectivity, ergonomics is engineering,” *Ergonomics, 45,* 991–994.

Hoffman, R.R. (2010). Theory → Concepts → Measures but Policies → Metrics. In E. Patterson and J. Miller (Eds.), *Macrocognition metrics and scenarios: Design and evaluation for real-world teams* (pp. 3-10). London: Ashgate.

Hoffman, R.R., Neville, K.N. and Fowlkes, J. (2009). Using cognitive task analysis to explore issues in the procurement of intelligent decision support systems. *Cognition, Technology, and Work, 11,* 57-70.

Klein, G., Woods, D.D., Bradshaw, J.D., Hoffman, R.R. and Feltovich, P.J. (November/December 2004). Ten challenges for making automation a “team player” in joint human-agent activity. *IEEE: Intelligent Systems*, pp. 91-95.

Roth E.N. and Eggleston, R.G. (2010). Forging new evaluation paradigms: Beyond statistical generalization. In E. Patterson and J. Miller (Eds.), *Macrocognition Metrics and Scenarios*, (pp. 204-219). London: Ashgate.

Scholtz, J. (2005). Metrics for evaluation of software technology to support intelligence analysis. *Proceedings of the Human Factors and Ergonomics Society 49th Annual Meeting* (p. 918). Santa Monica, CA: Human Factors and Ergonomics Society.

Stevens, S.S. (1946). On the theory of scales of measurement. Science, 103, 677-680.

Stevens, S.S. (1951). Mathematics, measurement, and psychophysics. In S.S. Stevens (Ed.), Handbook of experimental psychology. New York: John Wiley.

Velleman, P. F., and Wilkinson, L. (1993). Nominal, ordinal, interval, and ratio typologies are misleading. *The American Statistician, 47*, 65-72.