Workload Measurement in Human Autonomy Teaming: How and Why?

Jay Shively
NASA-Ames Research Center

27 June 2016
Outline

• Me and my biases
  – NASA-TLX
  – MIDAS

• Categories of metrics
  – Subjective
  – Objective
  – Physiological
  – Computational

• Workload and Human Autonomy Teaming
  – Changes workload

• Right tool for the right question
  – Assessment
  – Prediction
  – Design
• Worked with Sandy Hart and the TLX team from 1984 until about 1990

• Managed the Man-Machine Integration Design and Analysis System (MIDAS) 1990 – 1995

• Bias?

• Yes, but not in the way you might think --- I know where the skeletons are buried !!
NASA-TLX

- NASA Task Load Index (TLX)
  - Includes six subjective dimensions of global workload
  - First validation and development studies done by Hart and Staveland (1988)
  - Studies have used NASA-TLX along with physiological measures
    - Borghini et al. (2012) conducted a study assessing workload during driving a vehicle using EEG, along with other physiological data
    - Participant given NASA-TLX questionnaire at the end of each condition for subjective workload assessment
    - Correlation between NASA-TLX scores and physiological data
  - Sometimes has strongest effect size out of other workload measures
  - Sensitive to both task type and dual tasking (Matthews, Reinerman-Jones, Barber, & Abich IV, 2015)
  - Considered to be more favorable for subjects as compared to other measures of workload (Cao et al., 2009)
**NASA Task Load Index**

Hart and Staveland’s NASA Task Load Index (TLX) method assesses workload on five 4-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.

<table>
<thead>
<tr>
<th>Name</th>
<th>Task</th>
<th>Date</th>
</tr>
</thead>
</table>

**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 Example

<table>
<thead>
<tr>
<th>TITLE</th>
<th>ENDPOINTS</th>
<th>DESCRIPTIONS</th>
</tr>
</thead>
<tbody>
<tr>
<td>MENTAL DEMAND</td>
<td>Low/High</td>
<td>How much mental and perceptual activity was required (e.g., thinking, deciding, calculating, remembering, looking, searching, etc.)? Was the task easy or demanding, simple or complex, exacting or forgiving?</td>
</tr>
<tr>
<td>PHYSICAL DEMAND</td>
<td>Low/High</td>
<td>How much physical activity was required (e.g., pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?</td>
</tr>
<tr>
<td>TEMPORAL DEMAND</td>
<td>Low/High</td>
<td>How much time pressure did you feel due to the rate or pace at which the task or task elements occurred? Was the pace slow and leisurely or rapid and frantic?</td>
</tr>
<tr>
<td>PERFORMANCE</td>
<td>Good/Poor</td>
<td>How successful do you think you were in accomplishing the goals of the task set by the experimenter? How satisfied were you with your performance in accomplishing these goals?</td>
</tr>
<tr>
<td>EFFORT</td>
<td>Low/High</td>
<td>How hard did you have to work (mentally and physically) to accomplish your level of performance?</td>
</tr>
<tr>
<td>FRUSTRATION LEVEL</td>
<td>Low/High</td>
<td>How insecure, discouraged, irritated, stressed, and annoyed versus secure, gratified, content, relaxed, and complacent did you feel during the task?</td>
</tr>
</tbody>
</table>
SWAT

- Developed by Reid & Nygren (1988)
- Develops a single, global rating scale with interval properties (Rubio et al., 2004)
- Used in a variety of task settings (Hendy, Hamilton, & Landry, 1993)
Bedford Scale

• Modification of the Cooper Harper Scale (Roscoe, 1984)
• Unidimensional
• Uses a decision tree and examines whether:
  – The task can be successfully completed
  – The level of workload experienced was tolerable
  – The level of workload was satisfactory without reduction
• Taps into operator’s spare mental capacity
• Currently very few studies have used it in controlled settings
• More often used in applied settings
• Not enough data on validity of the scale available (NATO Guidelines on Human Engineering Testing and Evaluation, 2001)
Bedford Scale

1. Was workload satisfactory without reduction?
   - Y: Workload insignificant (WL1)
   - N: Was workload satisfactory tolerable for the task?
     - Y: Insufficient spare capacity for easy attention to additional tasks (WL4)
     - N: Reduced spare capacity. Additional tasks cannot be given the desired amount of attention (WL5)
   - Y: Enough spare capacity for all desirable additional tasks (WL3)
   - N: Little spare capacity. Level of effort allows little attention to additional tasks (WL6)

2. Was it possible to complete the task?
   - Y: Very little spare capacity, but maintenance of effort in the primary task not in question (WL7)
   - N: Very high workload with almost no spare capacity. Difficulty in maintaining level of effort (WL8)
   - Extremely high workload. No spare capacity. Serious doubts as to ability to maintain level of effort (WL9)
   - Task abandoned: Pilot unable to apply sufficient effort (WL10)
Modified Cooper Harper Scale

- Used most often in aviation
- Unidimensional
- Uses decision tree rating scale, with a score of 1 indicating “best” and a score of 10 indicating “worst”
- Relatively sensitive to changes in workload (Wierwille & Connor, 1983) and various types of workload
- Data is collected after the trial, participants poor at recalling past mental events (Wierwille & Casali, 1986)
- Limited to manual control tasks
Modified Cooper-Harper Scale

- **Operator Demand Level**
  - **Very Easy, Highly Desirable**: Operator Mental Effort is Minimal and Desired Performance is Easily Attainable
  - **Easy, Desirable**: Operator Mental Effort is Low and Desired Performance is Attainable
  - **Fair, Mild Difficulty**: Acceptable Operator Mental Effort is Required to Attain Adequate System Performance
  - **Minor But Annoying Difficulty**: Moderately High Operator Mental Effort is Required to Attain Adequate System Performance
  - **Moderately Objectionable Difficulty**: High Operator Mental Effort is Required to Attain Adequate System Performance
  - **Very Objectionable But Tolerable Difficulty**: Maximum Operator Mental Effort is Required to Attain Adequate System Performance
  - **Major Difficulty**: Maximum Operator Mental Effort is Required to Bring Errors to Moderate Level
  - **Major Difficulty**: Maximum Operator Mental Effort is Required to Avoid Large or Numerous Errors
  - **Major Difficulty**: Intense Operator Mental Effort is Required to Accomplish Task, But Frequently Numerous Errors Perceived
  - **Impossible**: Instructed Task Cannot Be Accomplished Reliably

- **Rating**
  - 1
  - 2
  - 3
  - 4
  - 5
  - 6
  - 7
  - 8
  - 9
  - 10

- **Operator Decision**
  - Yes
  - No

- **Flowchart**
  - **Is Mental Workload Level Acceptable?**
    - Yes
    - No
  - **Are Errors Small and Inconsequential?**
    - Yes
    - No
  - **Even Though Errors May Be Large or Frequent, Can Instructed Task Be Accomplished Most of the Time?**
    - Yes
    - No
Cooper-Garrod Estate Vineyards
Subjective

Pros:

• Pretty good idea of operator’s experience of workload (Crabtree, Bateman, & Acton, 1984)
• Cheap and easy (Stanton, Salmon & Walker, 2007)
• Years of use = validation (?)
• Gold standard ???
• Diagnostic when combined with objective measures (Crabtree, Bateman, & Acton, 1984)

Cons:

• Phenomenon has to be available for introspection (see SA) (Yeh & Wickens, 1984)
• Retrospective, i.e., not real-time
• Memorial (prone to memory failure ?) (Muckler & Seven, 1992)
• Not continuous, reflect average or peak
• Subjective – NOT objective data
Almost any of these (or a scale 1 – 100) gives you a pretty good idea of overall workload experienced by the operator.

NOT really diagnostic – I know of no system design ever modified because of too high “physical workload”

Individual differences “weightings” reduce variance – but mathematically have to !!

No one knows what the “own performance” scale means – maybe Sandy

Have effectively become the gold standard against which other metrics – such as physio or computational models are judged
Issues of Subjective Measures

• **Processing characteristics are lost** (Yeh & Wickens, 1984)
  – Limited in scope
  – Only provides scalar measures

• **Dissociations between subjective and objective measures** (Yeh & Wickens, 1988)

• **Difficult to compare results across scales** (Gopher & Braune, 1984)
  – Lack of formal theory for workload
  – Subjective measurement scales are influenced by how experimenters select scalar dimensions for rating
Objective

Metrics:

- Embedded Secondary tasks
- Naturally occurring secondary tasks
- WL probes
Workload Probes

• **Situation Present Assessment Method (SPAM)**
  – On-line probe method that can measure workload, in addition to SA (Stanton, Salmon, & Walker, 2007)

  – **Readiness latency**: Time from onset of “ready” prompt for query to an individual’s response to the prompt acts as an indicator for workload
    • Objective (Stanton, Salmon, & Walker, 2007)
    • Sometimes accompanied by an auditory warning signal (Pierce, 2012)

  – Queries can also ask operator to report current mental workload (Silva et al., 2013)
    • Scale
    • Subjective rating

  – Not intrusive to operator performance and workload (Silva et al., 2013; Keeler et al., 2015)
Embedded Secondary Tasks

• Operator performs a primary task in addition to a secondary task

• A wide variety of secondary tasks have been used in studies (Ogden, Levine, & Eisner, 1979)

• RT on secondary tasks often shows greatest sensitivity to workload changes

• Time estimation task is also sensitive, but can be intrusive (Wierwille, Rahimi, & Casali, 1985)

• Change detection also successful and less intrusive (Teo, Reinerman-Jones, & Szalma, 2015)
Objective

Pros:
• Objective data: RT, error

Cons:
• Specific to each implementation
• Low data rate
• Can be difficult to implement
• Momentary measure (not continuous)
• Can’t implement in some situations (real cockpit- can’t add secondary tasks)
Physiological

Metrics:

- Heart-rate (variability)
- Eye gaze
- GSR
- Eye-blink
- fMRI
- EEG
- fNIRS
Physiological Measures

• Eye Gaze
  – Gold, Körber, Lechner, and Bengler (2016)
    • Study to determine how traffic density and verbal tasks affect take-over performance in highly automated driving
      – I.e. how much time does the driver need in order to regain control of an automated vehicle if a situation necessitates this? Does traffic density and a verbal task have an affect on performance of taking back control?
    • Used eye gaze behavior as a measure of workload
      – Lower horizontal gaze distribution (HGD) = More workload (Wang, Reimer, Dobres, & Mehler, 2014)
Physiological Measures

• **Heart Rate Variability**
  
  – Decrease in heart rate variability may indicate an increase in mental workload (Mulder, 1980)
  
  – Strang, Best, and Funke (2014)
    
    • Studied mental workload of participants in a simulated training exercise involving realistic, large-scale air-combat scenarios.
    
    • Examined the ability of heart rate to predict mental workload.
    
    • Some data to support that heart rate may be able to predict mental workload, but this relationship is inconsistent.
Physiological Measures

• Eye Blink
  – Sometimes measured through electrooculography (EOG) (Veltman & Gaillard, 1996)
  – Length or frequency of blink
  – Not always sensitive to changes in workload (Wierwille & Connor, 1983)
  – Might need to be combined with other eye tracking techniques to be more reliable (Orden, Limbert, & Makeig, 2001)
Physiological Measures

• Functional Magnetic Resonance Imaging (fMRI)
  – Monitoring cerebral blood flow velocity (CBFV)
  – As CBFV increases in the prefrontal cortex, mental workload increases (Parasuraman & Caggiano, 2005)
  – Highly constrained environment
  – Limits what kind of activities can be analyzed (Warm, Parasuraman, & Matthews, 2008)
Physiological Measures

• Electroencephalogram (EEG)
  – Electrodes are placed on the scalp over various brain areas:
    • Fz, F3, F4, Cz, C3, C4, Pz, P3, P4

  – Different types of brain waves
    • Alpha (7-14 Hz)
    • Beta (14-30 Hz)
    • Theta (4 to 7 Hz)
    • Delta (up to 4 Hz)

  – As mental workload increases, alpha waves are replaced by beta waves, and frontal theta waves are increased (Borghini, et al., 2012)
Physiological Measures

- **Galvanic Skin Response (GSR)**
  - Measurement of resistance of skin tissue to electrical current
  - Measured through palms, inside / outside of wrist, arch of foot, forehead, or fingers
  - Susceptible to individual differences in response (Wierwille, 1979)
  - Found to be associated with cognitive workload (Shi et al., 2007)
    - Mean GSR increases as cognitive load increases
Physiological Measures

• **Functional Near-Infrared Spectroscopy (fNIRS)**
  – Relatively new measure
  – Monitors elevation of rSO$_2$
  – Higher rSO$_2$ levels = more cognitively demanding (Matthews, Renierman-Jones, Barber, & Abich IV, 2015)
  – Not always sensitive enough to changes, but does correlate well with other physiological measures of workload (e.g. HR) (Teo, Reinerman-Jones, & Szalma, 2015)
Physio

Pros:
• Continuous
• Potentially unobtrusive
• Face Validity – looks scientific (Levin et al., 2006)
• Supplements subjective measures (Wierwille & Eggemeier, 1993)

Cons:
• Not a “pure” workload signal
• Individual differences (Wierwille, 1979)
• Sensitive to external events/sources
• Poor correlation to subjective metrics (gold standard)
Computational Models

- MIDAS
- IMPRINT
- OMAR
- ACT-r*

MIDAS

- Man-machine Integration Design and Analysis System (MIDAS)
- NASA Ames Research Center human performance model (HPM) software tool
- Predict human-system performance and model human error
- “What-if” analyses (Gore, 2011)
- 3-D rapid prototyping, dynamic simulation, and human performance modeling with the aim to reduce design cycle time
- Links a virtual human to a computational cognitive structure that represents human capabilities and limitations (NASA, 2016)
- Currently v5 being used
• Cognitive component
  – Perceptual mechanism
  – Memory
  – Decision Maker
  – Response selection architecture
• Outputs include:
  – Dynamic visual representations
  – Timelines
  – Task lists
  – Cognitive loads along resource channels
  – Actual / perceived S.A.
  – Human error vulnerability
  – Human performance quality
• Air MIDAS
  – Aviation specific version of MIDAS
  – Models ATCo
  – Switches between control strategies depending on number of aircraft under control and the complexity of maneuvers the aircraft have to perform
• Improved Performance Research Integration Tool (IMPRINT)
• Developed by the U.S. Army Research Lab, Human Research & Engineering Directorate
• Software is available for free for
  – U.S. government agencies
  – U.S. private industry with U.S. government contract
  – U.S. colleges and universities working in HSI
• Designed to support system performance through lifecycle
• Can help to
  – Set realistic system requirements
  – Identify Soldier-driven constraints on system design
  – Evaluate the capability of available manpower and personnel to effectively operate and maintain a system under environmental stressors (U.S. Army, 2016)
IMPRINT Modules

**Warfighter**
Estimate the type of individuals who will be available to operate and maintain the system

**Missions**
Estimate the effect of operator performance on system performance, including time, accuracy, or mental workload

**Equipment**
Estimate maintenance man-hours required to attain acceptable system availability

**Forces**
Estimate the manpower needed to complete the routine and unplanned work performed by a force unit
• Provides a simulation environment that allows for modeling human operators, where they work, and the entities of the larger world that are reflected in their workplaces
• A “production rule-based executive process” regulates scheduling of competing tasks
• Emphasis on developing multiple-task behaviors from “functional centers” that are operating at the same time without an executive or central control (Deutsch, 1998)
Unique Characteristics of OMAR

- Stimuli directly affect procedural memory
- “Function-specific procedures” that represent specific brain areas coordinate the completion of tasks
- Resulting behaviors may be considered “intelligent”
- Task completion is mediated on a pair wise basis and not through a central executive. (Deutsch, 1998)
Computational

Pros:
• Learn a lot by formalizing description of your system
• Objective (sort of – input, assumptions, etc.)
• What if questions can be asked
• Can mod and re-run
• Consistent
• Can be done with notional system

Cons:
• Benefit might largely be in the process (more of my bias)
• Need a detailed task analysis/system design
• !@$#% input > !@#$ output
• Automation does not necessarily reduce workload, **just changes it**.

• Automation changes an operator’s role from manually **controlling** a system to **monitoring** the automated system (Parasuraman & Riley, 1997)

• **Examples**
  – **Wiener (1989)**
    • Pilot responses were divided when asked whether workload was decreased in a more automated cockpit
  – **Warm, Dember, & Hancock (1996)**
    • Monitoring tasks can lead to underarousal and increased mental workload
  – **Wiener & Curry (1980)**
    • Although automation may reduce *manual* workload, it may increase *overall* workload as a result of increased mental workload.
Relationship Between WL and Automation

- Workload
- Unpredictability
- Competency

Increased Human Mgt (Adaptable)

Increased Automation Mgt (Adaptive)

A Playbook® Approach to Delegation

- A means of Delegation
- Plays contain an implicit goal
- Plays define a “template” of plan/behavior alternatives—a “space” of delegated planning authority
  - “pre-compiled” with convenient label
  - Supervisor can further constrain/stipulate as desired—by reference to play structure
  - Monitoring and information reporting facilitated by shared intent structure
  - Dynamic, real time revision and tuning = “calling signals”
- Subordinates responsible for best-effort attempts within play constraints
Playbook & HAT

• Single Operator control of multiple, heterogeneous UAS
  (Simulations and flight tests)
  – Top ten pre-defined Plays – from operators
    • Convey support
    • Troops in contact
    • Recon an area
  – Increased Performance
  – Decreased Workload

• Human Autonomy **Teaming**
  – Reduced Crew Compliment in Commercial Aviation
  – One step further – not just delegation, but discussion, negotiation, joint problem solving
  – Automation (and interface) adapts by (largely) pilot-directed context
Why measure workload?

System Life-cycle

• Design
• Evaluation (R & D)
• Evaluation (Operational)
• Embedded (adaptive automation)

• WC Fielde:  Workload Consultant for Field Evaluation
Design

Environment:
• System doesn’t exist
• SME’s may be tangential
• Non-real time

Decisions:
• Roles and responsibilities
• Information flow/ displays
• Crew size

Metric:
Computational Models
Evaluation (R&D)

Environment:
• Prototype system
• Focus on other variables
• Real-time

Questions:
• Workload too high/low
• Effect of variables on WL

Metrics:
• Subjective
• Objective/secondary
• Phsyio

* Choice depends on ability to insert/identify secondary tasks
Environment:
• System
• Real users
• Real-time

Questions:
• Workload too high/low

Metrics:
• Subjective
• Physio (if non-intrusive)
Environment:
• System (WL eval is part of the system)
• Real users
• Real-time

Questions:
• Workload too high/low

Metrics:
• Subjective
• Performance
• Physio (if non-intrusive)
• Pros and Cons of all approaches
• Driven by the QUESTION
• Strongly advise using a battery of measures to converge on “workload”

• Adaptable vs. Adaptive Automation...
References


