Selecting Empirically Vetted Surveys

Dean Thomas, Project Leader

Chad Bieber

Rebecca Grier

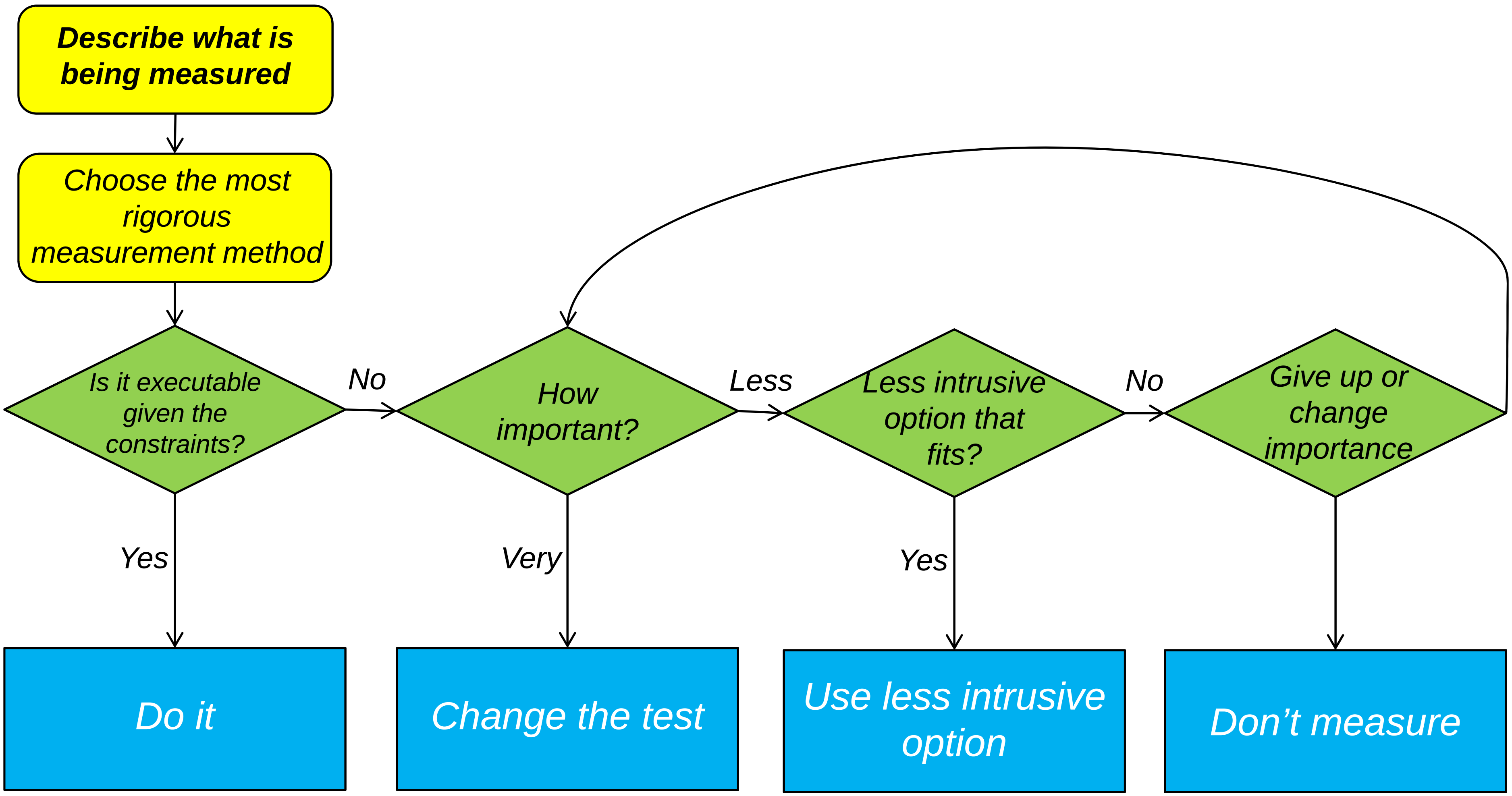
Justin Mary

Outline

- **How to choose a measurement method**
  - Clearly describe what is being measured
  - Choose the most rigorous measurement method that provides the data required to answer the question being asked
  - Identify the constraints – of the test, of the method, of the environment
  - If the chosen method does not fit the constraints, adjust the test or the method until they do

- **Examples**
  - KC-46 Workload
  - Apache Workload
  - RQ-7BV2 Workload
  - KC-46 Usability
  - KC-46 Diagnostic

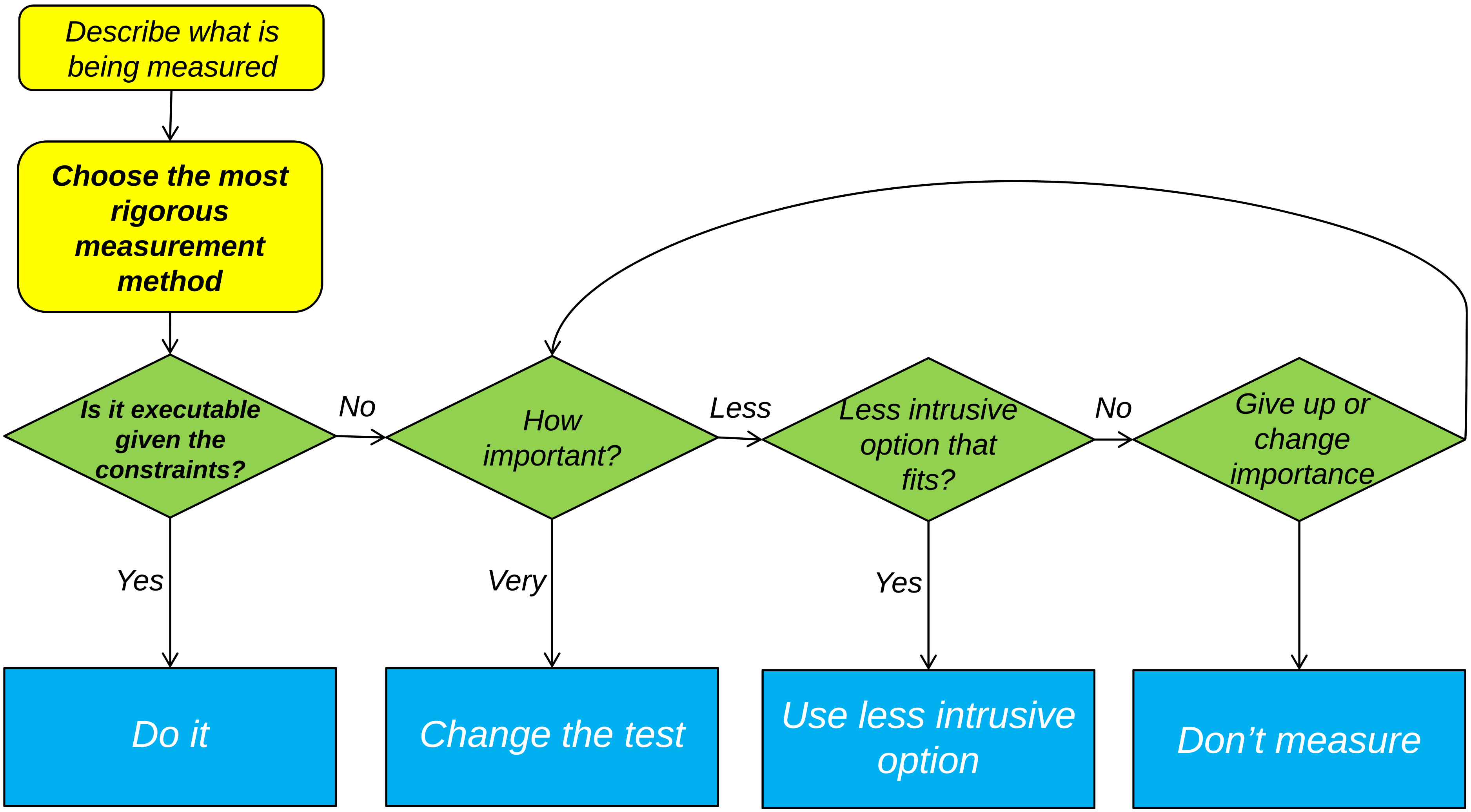
Decision Flowchart



Describe what is being measured

- **Which human measurement?**
  - Workload measurements have to be made with workload surveys
- **What is the purpose of the measure?**
  - Collecting demographics, supporting diagnostic analysis of a performance metric, or a primary response variable
  - Comparing factors - more power with continuous (or continuous-like) data
- **How will data be analyzed?**
  - Different statistics address different questions, and different response types support calculation of different statistics
  - What size difference between factors or vs. a threshold is meaningful?
    - Some surveys can detect larger/smaller differences (sensitivity).
  - Will data from multiple questions be aggregated into a single score?
    - Empirical surveys use aggregated data
    - Aggregating responses increases power
    - Un-answered questions are greater concern when aggregating questions

Decision Flowchart



Choose a measurement method, identify constraints

- **Use the question being asked and expected analysis to choose the most rigorous measurement method**
  - Are widely varying systems or Tactics, Techniques, and Procedures (TTPs) being tested to see which one reduces operator workload the most?
    - » Choose NASA-TLX – most sensitive, measures different dimensions of workload (e.g., mental, physical, temporal)
  - Is there a need to show clear improvement in a new training system before implementing across entire command?
    - » Measure training at Results level – quantify mission outcome improvement

- **Identify the Constraints – of the test, of the method, of the environment**
  - No one-size-fits-all list
  - Includes
    - » test (cost, range availability)
    - » environmental (Weather radar test needs weather)
    - » method (NASA-TLX takes 1 to 3 minutes, used shortly after task)
    - » physical (single-seat aircraft have no room for an observer)
    - » contract constraints
    - » number of times survey will be given
    - » many others

Fitting the measurement method in the test

- **How important is the thing being measured?**
  - If a primary response variable or major aspect of the system is being measured, then other parts of test design can change to fit requirements of most rigorous measurement method.
  - If a secondary metric or minor part of the system is being measured, then a less rigorous method can be chosen to fit the available testing opportunities.
- **How do operational or safety constraints limit choices?**
  - Can't use observer in a single-seat fighter – is video a viable alternative?
  - How much time can the operator safely devote to a survey?
- **Will it fit?**
  - If the chosen measurement method fits in the planned test – Great!
  - Otherwise, one needs to change – see Decision Flowchart

Outline

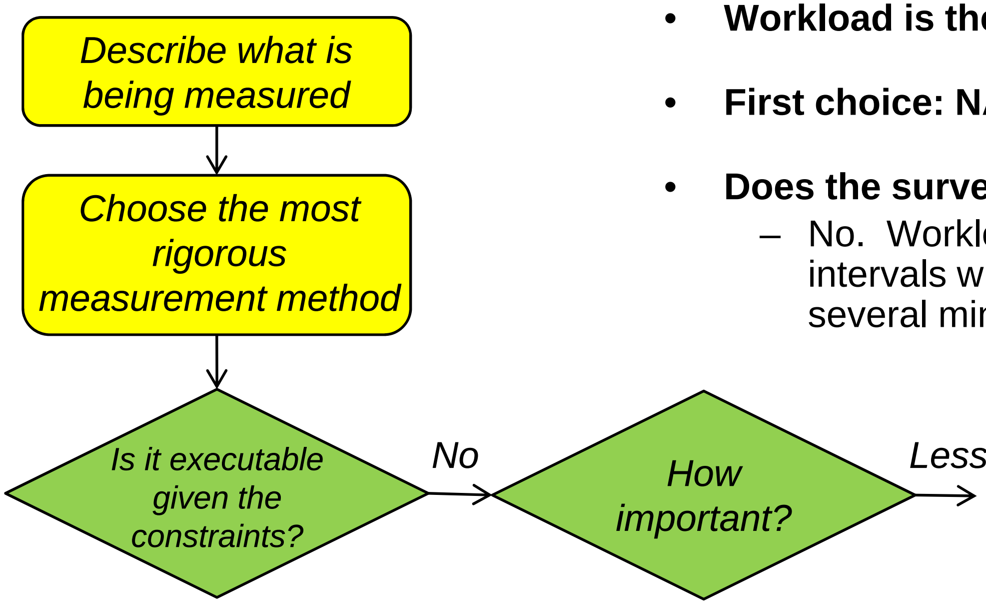
- **How to choose a measurement method**
  - Clearly describe what is being measured
  - Choose the most rigorous measurement method that provides the data required to answer the question being asked
  - Identify the constraints – of the test, of the method, of the environment
  - If the chosen method does not fit the constraints, adjust the test or the method until they do

- **Examples**
  - KC-46 Workload
  - Apache Workload
  - RQ-7BV2 Workload
  - KC-46 Usability
  - KC-46 Diagnostic

KC-46A Workload Example

- **New Aerial Refueling Operator Station**
  - Aerial Refueling Operator (ARO) views aircraft being refueled through 3-D video screens rather than a window
  - Want to understand ARO workload in this environment
- **Choosing a method**
  - Describe what is being measured
    - » What: Workload during specific tasks in a multi-hour mission
    - » Why: To support a workload Measurement of Effectiveness (MOE)
    - » How: Compare factors – operational conditions (e.g., day/night), different receiver aircraft being refueled.
  - Choose the most rigorous method
    - » NASA-TLX – provides diagnostic information and the most sensitivity

KC-46A Workload Example

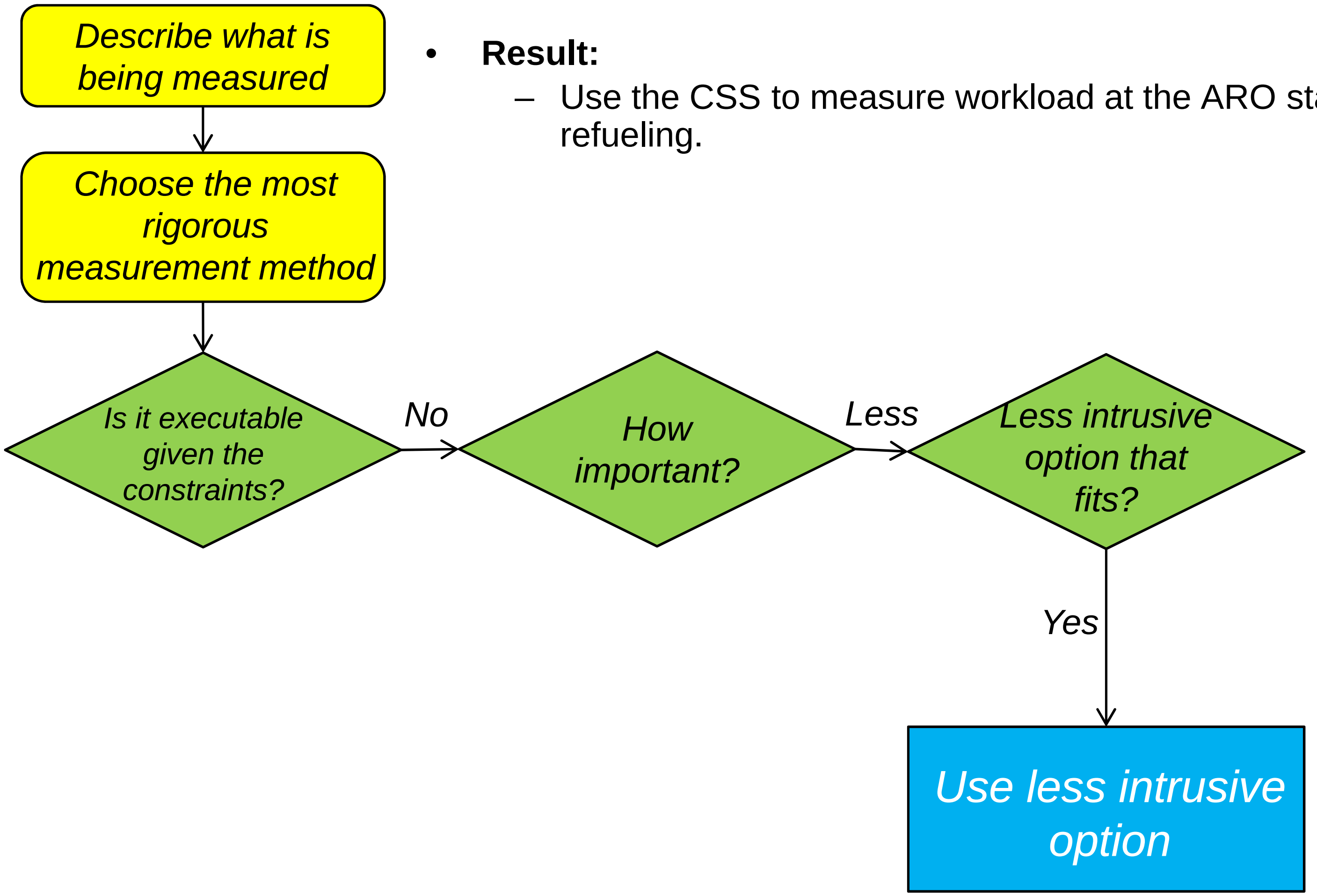


- **Workload is the measurement being made**
- **First choice: NASA-TLX**
- **Does the survey fit?**
  - No. Workload measurements will be taken at frequent intervals while receivers are waiting. May not have several minutes between tasks.

- **Is the measurement important enough to change the test - force the burden on the respondents and possibly lengthen test events?**
  - No. Workload is important, but not a primary response variable

KC-46A Workload Example

- **Is there a less intrusive option that fits?**
  - Yes. Crew Status Survey: Uni-dimensional; takes seconds to complete
- **Result:**
  - Use the CSS to measure workload at the ARO station during aerial refueling.



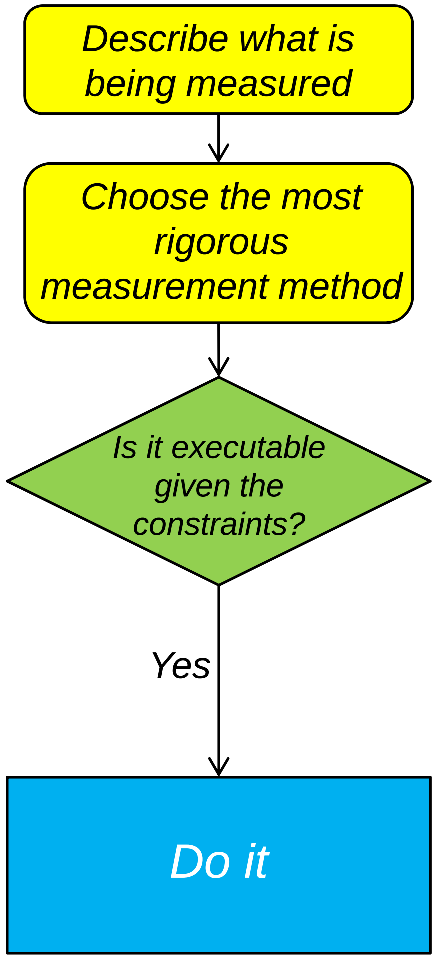
KC-46A Workload Analysis

- **CSS Workload results will be analyzed in several ways**
  - Change in scores will be analyzed to examine effect of experience
  - Workload during different factors will be analyzed
    - Can identify high and low workload scenarios
  - Results will be analyzed with respect to Performance
    - Identify if conflicts exist between user experience and reality, such as low workload with low performance
    - Support performance results with human responses
  - Comments analyzed for problem identification
  - Can't make general comparisons– no current research supports known workload benchmarks in CSS results.

AH-64E Apache Workload Example

- **Lot 4 AH-64E Apache Attack Helicopter FOT&E**
  - Several systems have been upgraded, to include Link 16, upgraded sensors, and new video transfer capability
    - » Expected outcome: improved Joint operations and mission effectiveness
    - » Experiment designed around time to find first target during a mission
    - » Want to measure workload during the mission in conjunction with this primary metric
- **Choosing a method**
  - Describe what is being measured
    - » What: Workload over the entire mission
    - » Why: To support a primary response variable
    - » How: Compare workload in different missions using DOE built for time to find first target
  - Choose the most rigorous method
    - » NASA-TLX – provides diagnostic information and the most sensitivity

Apache Workload Example



- **What is being measured?**
  - Workload
- **First choice: NASA-TLX**
  - Diagnostic, good sensitivity
- **Does it fit?**
  - 3 minutes of time available after mission, before debrief
- **Do it!**

What if it wasn't executable?

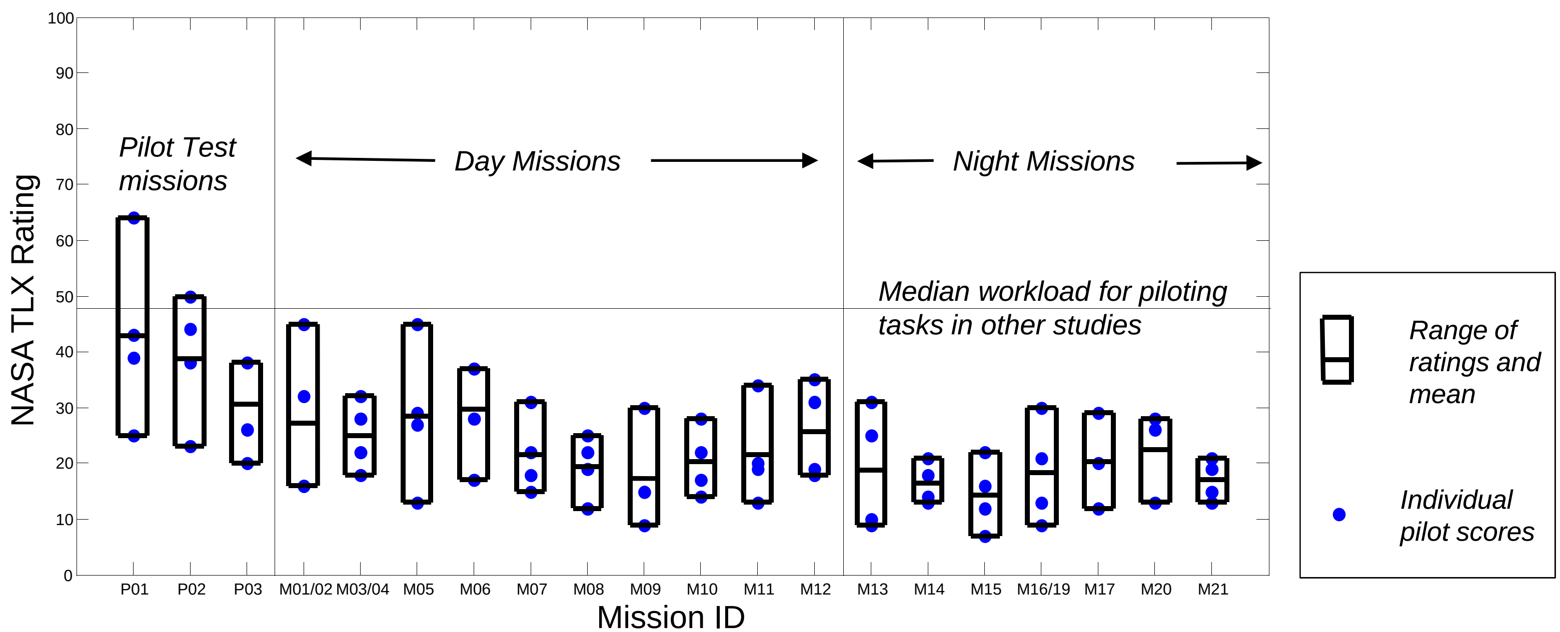
- **In this test, the item of interest was the entire mission**
  - Survey can be administered after the mission is finished
  - Plenty of time for NASA-TLX after mission is complete
- **What about specific tasks within the mission?**
  - To measure tasks within the mission, a survey would have to be used after the specific task, preferably before any other task
  - Unlikely that a minute or two per NASA-TLX would have fit into the flight
  - CSS is a possibility – can administer in flight on kneeboard or via voice question if time permits
  - Other alternatives include physiological measures
    - » Requires equipment and complex analysis, but doesn't take time away from operator

Apache Workload Analysis

- **NASA-TLX survey administered after each mission**
- **Four Factors chosen for primary metric (time to find first target)**
  - Link 16 Targeting Data (yes or no), Battlefield Density (high or low), Light Level (day or night), Pilot Seat Location (front or back)
- **Analysis shows several significant correlations**
  - High Density resulted in higher workload with Link16 (*p* = 0.02)
  - Front seat pilot had higher workload with Link 16 (*p* = 0.10)
  - Night missions were significantly lower workload than day, but all day missions were accomplished first, then night missions. Unclear if results were due to time (experience) or to light level

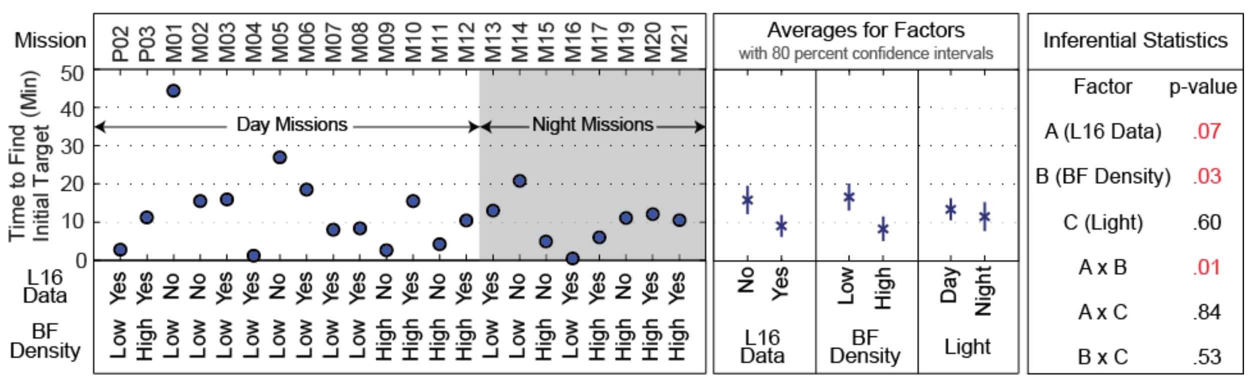
*** 80 % confidence, 10% significance**

Terms	p-value
Link 16 Targeting Data	0.22
Battlefield Density	0.76
Light Level	**0.001**
Pilot Seat Location	0.16
Targeting Data*Battlefield Density	**0.02**
Targeting Info*Light Level	0.73
Targeting Data*Pilot Location	**0.10**
Battlefield Density*Light Level	0.64
Battlefield Density*Pilot Location	0.39
Light Level*Pilot Location	0.33



Apache Workload vs. Performance

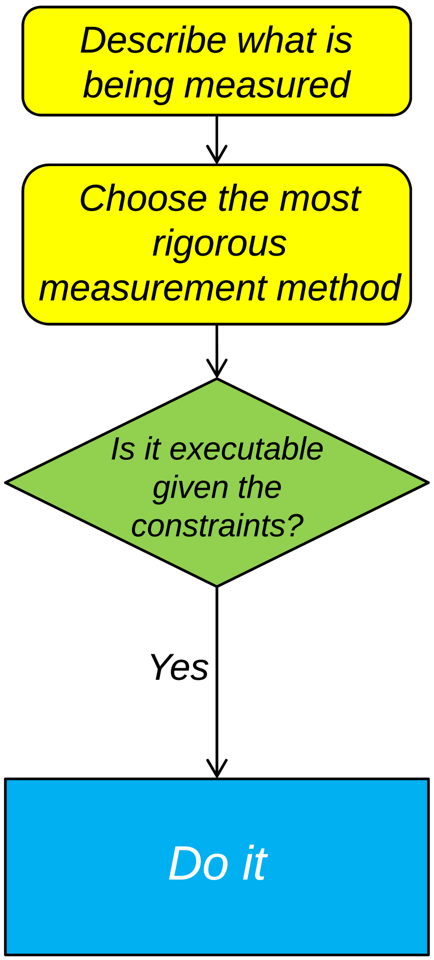
- **Workload differences were found – what do they mean about the mission?**
- **Primary metric – time to find first target**
  - Key finding – Link 16 improved time for low density battlefield (*p* = 0.01)
  - When battlefield density was high – many targets were present – time to find first target was shorter (*p* = .03) whether or not Link 16 was available
- **What does this mean?**
  - Higher effectiveness with Link 16 and low density– no increase in workload
    - » Clear benefit!
  - Higher workload and similar effectiveness with Link 16 and dense battlefield
    - » Correlation, not causation, but potential information for developing TTPs or further testing



Shadow Workload Example

- **RQ-7BV2 Shadow Tactical Unmanned Aerial System (TUAS) FOT&E**
  - Multiple systems improved including a new Universal Ground Control Station (UCGS) with faster processors, improved algorithms, and better ergonomics
    - » Expected outcome: improved mission effectiveness with no greater workload for sensor operator
    - » Free-play exercise – little ability to design the experiment
- **Choosing a method**
  - Describe what is being measured
    - » What: Workload during specific tasks in a multi-hour mission
    - » Why: To support a workload MOE
    - » How: Compare workload across different factors.
  - Choose the most rigorous method
    - » Choose the NASA-TLX, provides diagnosticity and the most sensitivity

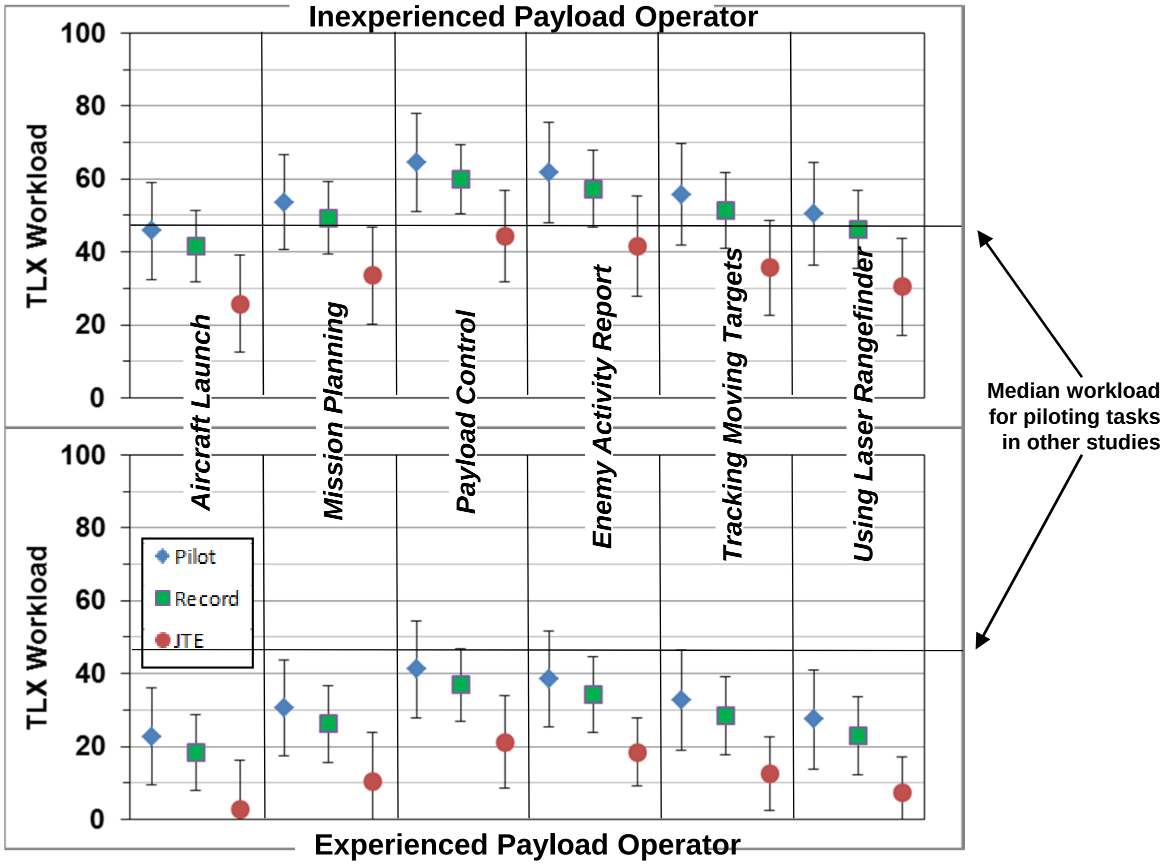
Shadow Workload Example



- **What is being measured?**
  - Workload
- **Workload measurement choices**
  - Need to compare with previous NASA-TLX
    - » Choose NASA-TLX
- **Does it fit?**
  - Yes. Time for questionnaires available after mission before debrief.
- **Do it!**

Shadow Workload Example

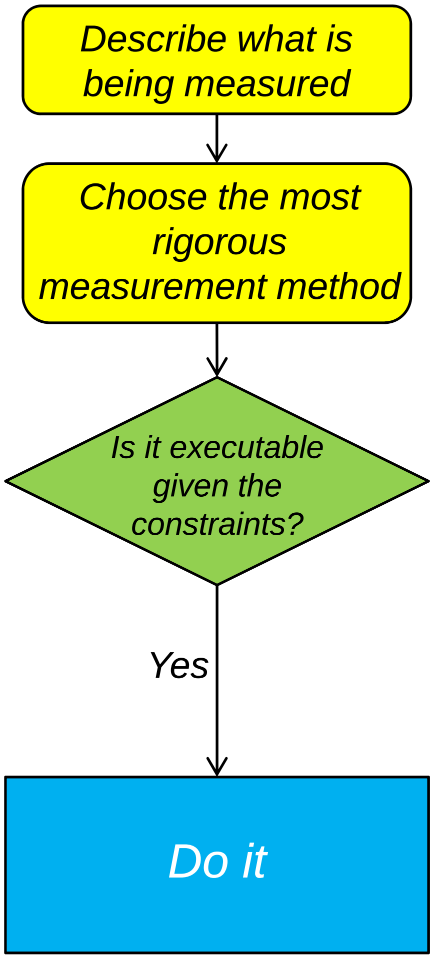
- **Significant effects**
  - Payload operator workload was significantly affected by
    - » operator experience ($p < 0.0001$)
    - » test phase ($p = 0.0019$)
    - » task ($p = 0.0181$)
- Throughout all phases and tasks, inexperienced operators were subject to a higher workload than experienced operators



KC-46A Usability Example

- **KC-46A –Air Refueling Operator Station**
  - Refueling Boom controls and system interface significantly changed from previous designs
  - Expected outcome: improved capability (video feed, IR)
- **Choosing a method**
  - Describe what is being measured
    - » What: Usability of Air Refueling Operator Station
    - » Why: To support “User rating” MOEs
    - » How: General comparison to usability benchmarks, identify problems
  - Choose the most rigorous method
    - » SUS is most rigorous usability option
    - » Use open-ended questions to identify problems throughout test

KC-46A Usability Example



- **What is being measured?**
  - Usability
- **First choice: SUS with an open-ended comment, several times throughout test**
  - Shows effect of experience
  - Comparative ability
  - Problem ID via open-ended comment
- **Does it fit?**
  - Yes, 3 minutes are available at periodic times throughout test period
- **Do it!**

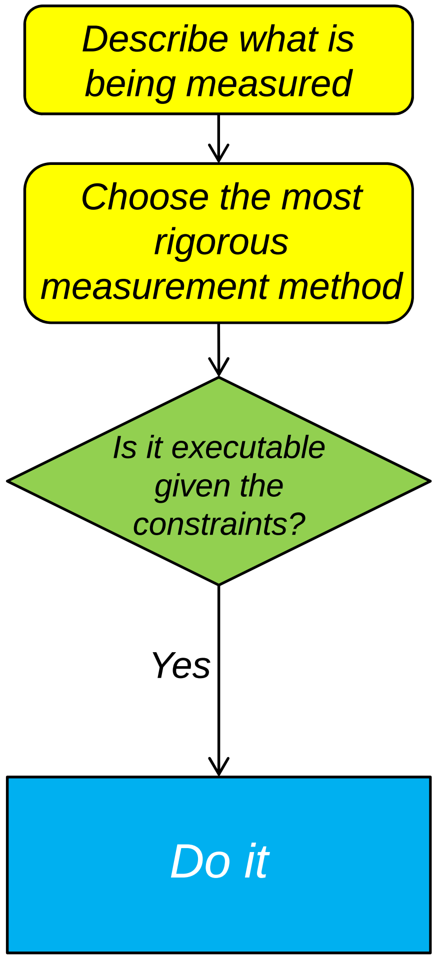
KC-46A Usability Analysis

- **Usability will be analyzed in several ways**
  - Scores will be compared against known ranges for Good, Fair, Poor
  - Change in scores will be analyzed for effect of experience
  - Sample will be analyzed for demographic effects
    - » Do operators with certain backgrounds find the new station easier/harder to use?
  - Results will be compared with Performance
    - » Can identify conflicts in perception and help interpret performance results
  - Comments analyzed for problem identification

KC-46A Custom Question Example

- **Many new features and combinations in the KC-46A cockpit**
  - Some problems will likely show up, but hard to identify all possibilities before testing
  - Desired goal: Use aircrew feedback to identify problems
- **Choosing a method**
  - Describe what is being measured
    - What: The crew is being used as subject matter experts to diagnose problems
    - Why: To identify problems in the system under test
    - How: Problem areas identified for further targeted analysis
  - Choose the most rigorous method
    - Custom open ended questions capture unknown problems
    - A few targeted closed-response questions for areas of particular interest

KC-46A Custom Question Example



- **What is being measured?**
  - Problem identification using crew as SMEs.
- **First choice: A few, targeted, questions plus open-ended comments after every mission, additional targeted questions at end of test or periodically to address identified problems.**
  - Identifies unknown problems and key areas, later questionnaires can be tailored to address specific areas discovered.
- **Does it fit?**
  - Yes, time for written comments after each mission
- **Do it!**

KC-46A Custom question analysis

- **As test progresses, comments monitored for problem areas**
  - Unique combination of events that exposed potential hazards
  - Common complaints that show areas of potential concern
- **Create specific questions to address identified areas**
  - Can support more detailed analysis if needed
  - Questions that aren't needed never get created/asked
  - Requires some intentional flexibility in the test plan

Summary

- **What makes a good survey**
  - Validity, reliability, other psychometric attributes
- **Overview of surveys**
  - Workload, usability, situational awareness, training effectiveness, all analyzed with respect to performance
- **How to choose a measurement method**
  - Pick the most rigorous method that fits the constraints
- **Benefits of empirically vetted surveys**
  - General comparisons for well understood surveys
  - Specific comparisons for empirical surveys in well-designed tests
  - Diagnostic ability when used in conjunction with performance
- **Examples**

Up Next

- Custom-Made Surveys
- ABIS Case Study
- Administration & Analysis
- Air Force DCGS Case Study

Sources

Ames, Lawrence L., and Edward J. George, *Revision and Verification of a Seven-point workload estimate scale.* Air Force Flight Test Center, Edwards AFB, CA. 1993

Bangor, A., P.T. Kortum, and J.T. Miller, *Determining what individual SUS scores mean: Adding an adjective rating scale.* Journal of Usability Studies, 2009. 4(3): p. 114-123.

Bangor, A., P.T. Kortum, and J.T. Miller, *An empirical evaluation of the system usability scale.* Intl. Journal of Human–Computer Interaction, 2008. 24(6): p. 574-594.

Bonner, M.A. and G.F. Wilson, *Heart rate measures of flight test and evaluation.* The International Journal of Aviation Psychology, 2002. 12(1): p. 63-77.

Cinaz, B., et al. *Monitoring of mental workload levels*. in *Proceedings of IADIS eHealth conference*. 2010.

Endsley, M.R. *Situational awareness global assessment technique (SAGAT)*. in *Aerospace and Electronics Conference, 1988. NAECON 1988., Proceedings of the IEEE 1988 National*. 1988. IEEE.

Endsley, M.R. and D.J. Garland, *Situational awareness Analysis and Measurement.* 2000: CRC Press

Gawron, V.J., *Human performance, workload, and situational awareness measures handbook*. 2008: CRC Press.

George, Edward J., *The Psychometric anatomy of two unidimensional workload scales.* Air Force Flight Test Center, Edwards AFB, CA. 2004

Gilmore, M., *Guidance on the Use and Design of Surveys in Operational Test and Evaluation [Memorandum].* 2014: DOT&E

Grier, R., "Situational Awareness in Command and Control Environments" *The Handbook of Applied Perception Research. 2015*

Grier, R., *How High is High? A Meta-analysis of NASA-TLX Global Workload Scales*. 2014, Institute for Defense Analysis: Alexandria, VA.

Sources

Hill, S.G., et al., *Comparison of four subjective workload rating scales.* Human Factors: The Journal of the Human Factors and Ergonomics Society, 1992. 34(4): p. 429-439.

Hart, S. G. and L. E. Staveland. *Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research*. Advances in Psychology, 1988. 52: p. 139-183

Kirkpatrick, D.L., *Evaluating Training Programs: The Four Levels*. 1998, Berrett-Koehler Publishers: San Francisco, CA.

Kruger, J. and D. Dunning, *Unskilled and unaware of it: how difficulties in recognizing one's own incompetence lead to inflated self-assessments.* Journal of personality and social psychology, 1999. 77(6): p. 1121.

Linde, C. and R.J. Shively. *Field Study of Communication and Workload in Police Helicopters: Implications for AI Cockpit Design*. in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*. 1988. SAGE Publications.

Miller, S., *Workload Measures.* 2001.

Phillips, J.J., *Level four and beyond: An ROI model*, in *Evaluating corporate training: Models and issues*. 1998, Springer. p. 113-140.

Reason, J., *Human error*. 1990: Cambridge University press.

Salvendy, G., *Handbook of human factors and ergonomics*. 2012: John Wiley & Sons.

Bedford

- Find a picture or compare to mCH

Why not use the modified Cooper-Harper?

- **The original Cooper-Harper Handling Qualities Rating Scale has been used very successfully by test pilots in the US and other militaries and in industry for decades and is used in MIL-STD-1797B Flying Qualities of Piloted Aircraft.**
  - MIL-STD-1797B explicitly defines the adjectives Satisfactory, Tolerable, and Controllable.
  - Specific tasks are clearly described with explicit definitions for Desired and Adequate performance
    - » Details are export controlled, fine control tasks typically defined in single feet or mils, gross control tasks in tens of feet
  - Tasks accomplished in isolation, are created to be representative of “operational” needs but are not executed in an operational environment
  - Test pilots are highly trained in use of the rating scale, have very broad experience in aircraft of varying handling qualities, and have both theoretical and hands-on training in evaluating and understanding closed-loop control theory as it applies to tasks involved in pilot-vehicle control.

- **Modifications of the Cooper-Harper scale for workload are not used in such a structured environment**
  - Without explicit definitions, operational users are left to come up with their own individual definitions of Satisfactory, acceptable, and similar adjectives.
  - Operational users hesitant to cross “acceptable” cutoff – causing clustering
    - » Linde (1988) saw this when every rating in the study was a 3, Bonner (2002) saw ranges from 2.7 to 3.1 for normal ground and flight ops.
    - » Roscoe (1984) encountered this when crews insisted on entering a 3.5 score – above a 3, but not past the “Acceptable” line.

General Comparative Ability- Workload

[Grier 2014]

- **Range of workloads separated by task area**
  - >1000 NASA TLX scores analyzed
- **Must consider task and performance to identify if workload is acceptable or not**

	Min	Mean (SD)	50%	75%	Max
Daily Activities	7.20	19.34 (8.10)	18.30	25.90	37.70
Card Sorting	16.00	26.77 (8.49)	25.63	27.88	49.80
Mechanical Tasks	20.10	30.52 (8.17)	27.95	33.68	51.03
Navigation	19.72	40.09 (15.50)	37.70	52.74	68.90
Driving Car	15.00	40.59 (13.39)	41.52	51.73	68.50
Process Control	23.90	42.21 (12.49)	42.00	51.83	69.70
Cognitive Activities	13.08	43.89 (13.99)	46.00	54.66	64.90
Classification	8.00	43.92 (18.33)	46.00	51.20	84.30
Computer	7.46	44.39 (21.75)	54.00	60.00	78.00
Pilot Aircraft	16.00	46.29 (11.94)	47.78	54.80	74.00
Memory	6.59	48.01 (20.30)	44.59	66.58	83.50
Command & Control	20.00	48.89 (13.51)	50.55	59.50	75.80
Medical	9.00	48.89 (14.84)	50.60	61.45	77.35
Monitoring	20.00	51.27 (14.15)	52.24	62.63	77.00
Tracking	19.08	51.79 (14.86)	51.00	62.43	88.50
Robot Operation	9.59	52.62 (15.49)	56.00	63.00	80.00
Air Traffic Control	6.21	54.31 (17.30)	52.44	68.32	85.00
Video Game	14.08	54.68 (13.34)	56.50	63.73	78.00
Visual Search	28.98	58.48 (11.52)	57.89	67.74	79.23
Physical Activities	40.83	61.63 (11.07)	62.00	71.83	75.19
Overall	**6.21**	**48.07 (16.11)**	**49.93**	**60.00**	**88.50**