CLDP Human Systems Integration Workshop

April 2-4, 2024

COMMERCIAL LEO DEVELOPMENT PROGRAM

Human Performance
Kritina Holden, Ph.D.

- 23 March 2024

Human Performance

- Overview/Background:
  - The most important aspect of human rating is the collection of human performance data to demonstrate that the system is legible, operable, and usable with an acceptable cognitive workload and level of accuracy.
    - If crew cannot read the text or interpret small icons – there is a legibility issue
    - If crew cannot position the cursor where needed or turn the hatch door crank – there is an operability issue
    - If crew have difficulty navigating a display, accessing information, or understanding what is displayed – there is a usability issue
    - If crew have a task that is so complex they cannot monitor for alarms or respond to an important crew question – there may be a workload issue
    - If crew are making too many errors – any/all of the above issues may be the cause

All of these issues increase risk to crew.

Importance of Measuring Human Performance

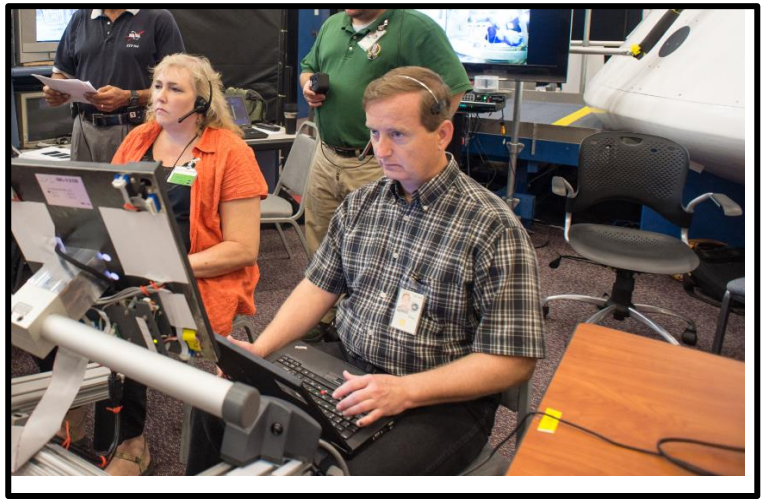
- Unsatisfactory human performance can result in failed verifications and impacts to flight schedules – or waivers that could put crew safety at increased risk
- Poor human performance data indicates that a Human-Centered Design process may not have been followed
  - ❖ Was a **Task Analysis** completed – so you know the system has been designed to perform those specific tasks?
    - Task Analysis should be driving selected test scenarios used in verification
  - Was a **Human Error Analysis** completed – so potential errors have been designed out where possible?
  - Were human-system **requirements** and display standards followed? (first line of defense)
  - Are you performing/planning **iterative developmental tests** with early user feedback (second line of defense)

Human-in-the-Loop (HITL) Tests

- What tasks should be included?
  - Human performance should be evaluated with a representative set of tasks from the **Task Analysis**
  - Choose tasks that are:
    - Frequent
    - Nominal and off-nominal
    - Safety or time-critical
    - Particularly complex
  - Tasks should be integrated into realistic scenarios, using as high a fidelity environment as possible
- What kind of participants are needed?
  - Requirements may specify “crew”, or “crew-like” participants
  - Plan on a minimum of 5 participants for development tests, and a minimum of 10 for verification tests
    - Research shows that the fewer participants tested, the higher the risk that design issues may exist that have not been found

Human-in-the-Loop (HITL) Tests, cont.

- Use multiple measures of performance. There is no one gold standard measure that will capture all aspects of performance with a system. Recommended measures include:
  - Legibility
  - Operability
  - Usability
  - Workload
  - Errors
- Well-planned tests can close multiple verifications with one test
- Make testing a formal data collection event
  - ❖ Have a detailed test plan, and follow it
  - ❖ Control the noise and number of observers
  - Collect data rather than just opinions
  - ❖ Do not discuss verification passing scores with participants



Legibility

- Crew must be able to read and interpret what is on the display in a timely and accurate manner, under flight-like conditions.
- What impacts legibility?
  - Text/character size – this has been a common problem, failing to consider older eyes and spaceflight impacts
  - Color contrast – this is often a problem, particularly with black backgrounds
  - Proximity of other visual objects – display clutter is sometimes a problem
  - Lighting
  - Vibration
  - ❖ Physical obstructions
- Measuring Legibility
  - Simple identification/readability test – ask participants to read certain lines or use a laser pointer to identify items to be read. Measure accuracy.

Operability

- Crew must be able to operate controls with accuracy under flight-like conditions and time constraints.
- What impacts operability?
  - Control shape – cursor control device requires an awkward hand position
  - ❖ Location – hand controller mounted too close to a wall, so can't deflect it all the way
  - Use environment (including acceleration, vibration, suited, pressurized operation, wearing PPE) – can't push the button with a pressurized glove without hitting adjacent controls
  - Feedback – no click or visual indication – did I push the button?
  - Labeling – unclear as to what this is/does
  - Inadvertent actuation protection – keep bumping a switch, resulting in need to rework
  - Display-control movement compatibility – why is cursor moving up when I'm turning knob left?
- Measuring Operability
  - Iterative developmental testing to identify issues early, when they can be resolved

Usability

- Usable interfaces are **effective** (help get the task done), **efficient** (make the job quick and easy), and **satisfying** (enjoyable to use).

❖ Interfaces that are not usable can result in errors, frustration, undesirable workarounds or lack of use.

- What impacts usability?
  - Easy of learning
  - Level of consistency (internal and external)
  - Interface elements
    - Labels, colors, icons, text
    - Layout and organization
    - Information content and terminology
  - Interaction elements
    - Work/task flow
    - Method of navigating the interface
    - Method of interaction (touchscreen, keyboard, voice, etc.)

Measuring Usability

- Objective Test Methods
  - Task success
  - Task completion time (for time constrained tasks)
  - Number and type of errors/issues/requests for assistance
    - All errors and issues should be investigated to understand probable cause
    - Always provide opportunity for comments to explain the issue – written or verbal
      - Think Aloud method, free form questionnaire items, post-test debriefs
- Subjective Test Methods
  - System Usability Scale (SUS)
  - NASA Modified SUS

System Usability Scale (SUS)

- System Usability Scale (SUS; Brooke, 1996)
  - Validated scale used to determine the subjective usability of a system.
    - Appropriate for evaluating “systems”, rather than component-level designs or very short procedures with simple actions – e.g., usability of a screwdriver
  - Maximum score achievable is 100.
  - Heavily validated and widely published
  - Used to measure usability in a variety of environment, with excellent results
  - Should be used in concert with objective data such as task completion time and errors
  - Quick to administer and can be used to measure progress of system design over time.

NASA Modified System Usability Scale (NMSUS)

- Verification testing using the SUS found some crew had issues with the wording of the some of the items
  - ❖ SUS was developed primarily for consumer type products, rather than safety-critical applications
- NASA funded a project to tailor the SUS for better applicability to NASA verification testing
  - NMSUS has 2 fewer items and some wording modifications (based on crew comments during verification)
  - NMSUS was validated through testing and compared to the SUS
    - A brief tech memo available for details on how the NMSUS was developed and validated
  - NMSUS is currently in use across NASA programs
- Administration
  - ❖ Do not discuss passing scores with participants
  - As with any rating scale, follow up extreme or surprising ratings in a post-test debrief.

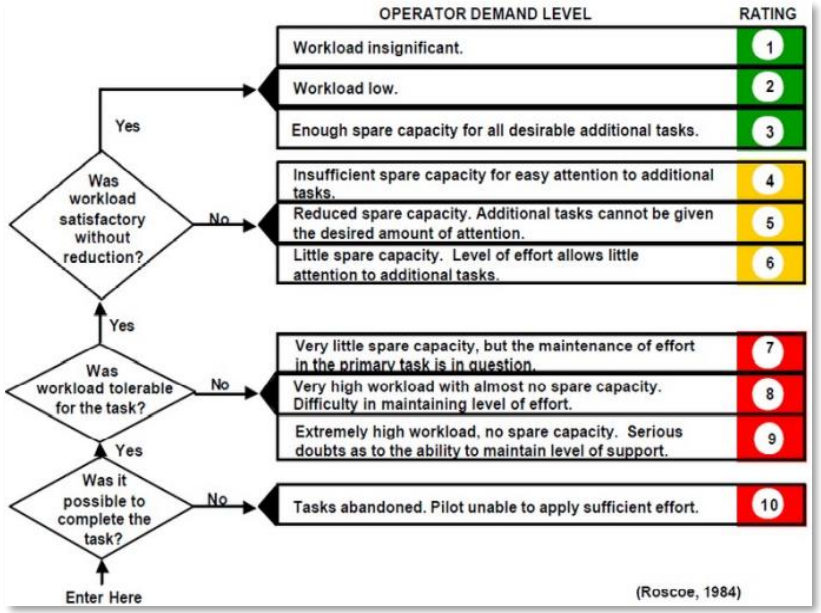
	Strongly Disagree 1	2	3	4	Strongly Agree 5
1. I thought the system was easy to use.	O·	O·	O·	O·	O·
2. I think that I would need technical support to be able to use this system.	O·	O·	O·	O·	O·
3. I found the various functions in this system were well integrated.	O·	O·	O·	O·	O·
4. I thought there was too much inconsistency in this system.	O·	O·	O·	O·	O·
5. I imagine that most trained crewmembers would learn to use this system very quickly.	O·	O·	O·	O·	O·
6. I found the system very cumbersome to use.	O·	O·	O·	O·	O·
7. I felt very confident using the system.	O·	O·	O·	O·	O·
8. I needed a lot of training on this system in order to get going.	O·	O·	O·	O·	O·

Workload

- Workload is a measure of perceived level of mental effort
  - **High workload** can lead to hurried performance, more errors, frustration, fatigue, and poor awareness of surroundings.
  - **Low workload** can lead to boredom, complacency, more errors, frustration, fatigue, and poor awareness of surroundings
- What impacts workload?
  - Task too complex
  - Time/schedule pressure
  - Poor usability
  - Environmental stressors
- While a task may be achievable and pass usability, if it requires a high level of workload, performance could break down over time, under stress, or when crew are deconditioned

Measuring Workload

- Objective Test Methods
  - Physiological measures – EEG, heart rate, pupil dilation, fNIRS
- Subjective Test Methods
  - NASA Task Load Index (TLX; Hart and Staveland, 1988)
    - Diagnostic – measuring: mental, physical, temporal, performance, effort, frustration
    - Good for developmental testing
  - Bedford Workload Scale (Roscoe, 1984)
    - Built on the concept of *spare mental capacity*
    - Preferred for verification because it is familiar to crew, and quick to administer
- Workload measures are often paired with Usability measures in a single HITL
- Workload should be one of the first scales administered after task performance due to the reliance on memory
  - Request rationale for the rating if possible



Bedford Workload Scale

Errors

- Errors can be minor nuisances resulting in rework or can have serious consequences. The goal should be to prevent errors or reduce the impact of errors.
  - **Human Error** – error due to a human failing - forgot training, distracted and missed a message, etc. (unrelated to design)
  - **Design Induced Error** – intentional action that does not reach its intended goal due to design issues
    - Missed or incorrect inputs or selections
    - Display navigation errors
    - Errors due to inadequate hardware component design
    - Errors due to lack of system feedback to user inputs
    - Errors due to design inconsistency or unfamiliar terminology
    - Inability to complete a task step
- If you have many errors per task step, there is likely a design issue with that task step
- If you have a participant that has many errors – there may be a training issue with that participant
- Recoverable design-induced errors still negatively impact crew performance in terms of time and satisfaction – thus they are still included in error calculations

Human Performance Requirements

- Governing/related requirements in CLDP-1130 draft:
  - [R.CLDS.109] CREW INTERFACE USABILITY
  - [R.CLDS.110] CREW INTERFACE WORKLOAD
  - [R.CLDS.150] CREW OPERATION LOADS
  - [R.CLDS.242] DESIGN INDUCED CREW ERRORS
  - [R.CLDS.247] OPERABILITY OF CONTROLS
  - [R.CLDS.261] INTERFACE LEGIBILITY

Final Takeaways

- Make sure your designs are legible, operable, and usable, with acceptable workload and accuracy.
  - Adherence to requirements and standards is the first line of defense
  - Human-in-the-Loop (HITL) testing is the second line of defense
    - This is required for human rating and is the best chance for identifying design issues that can be fixed prior to flight
    - Use enough participants to provide high confidence that a design is solid
    - Use best practice and NASA proven methods for testing

- Crew time is a precious resource - it should be spent on mission tasks, rather than struggling with poor designs, or performing rework resulting from error-prone interfaces that could have been fixed prior to flight.

Thank you!