Interface Consistency:
Phase I Results & Phase II Status

Amanda Smith[1], Kritina Holden[2], Ian Robertson[1], John Karasinski[3], Satyajit Upasani[1], Joschka Monsonyi[4], Shu-Chieh Wu[5], Megan Parisi[3], Katie McTigue[3], and Ryan Lange[4]

[1]KBR, [2]Leidos, [3]NASA Ames Research Center, [4]Aegis Aerospace, [5]San Jose State University

2025 Human Research Program
Investigators’ Workshop

Characterizing the Problem

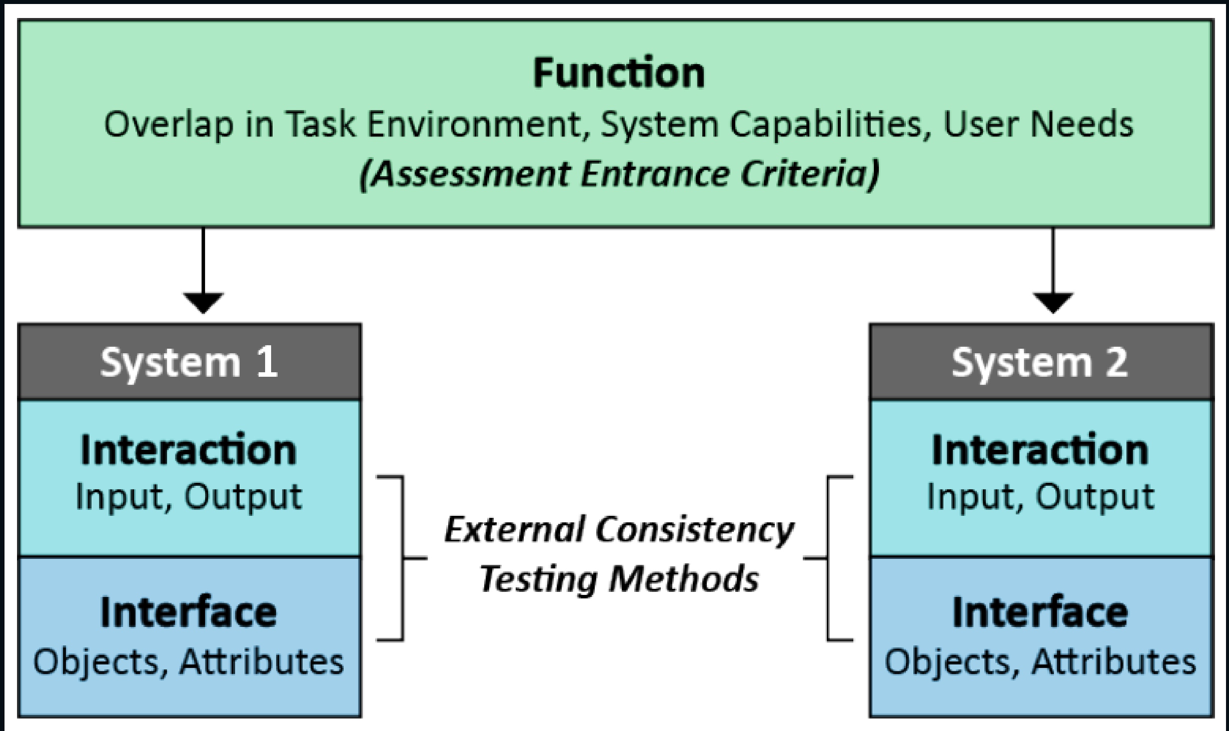
- Multiple internal and commercial Providers are designing vehicle systems for exploration missions, likely leading to considerable design diversity
- Literature yields mixed results (e.g., workload, errors, time); variability in operational definitions, stimuli and manipulations
- Human performance risks related to inconsistency between systems are poorly understood
- “Gold Standard” tools quantifying interface consistency have not been established

Phase I: Characterization & Measurement

Phase I: Consistency as a Construct

Consistency Definition:

The degree to which an interface supports a ***familiar*** *interpretation and interaction pattern through use of* ***design features*** *and* ***interaction styles*** *to achieve a* ***task previously learned and performed*** *on a different interface.*



Function
Overlap in Task Environment, System Capabilities, User Needs
(Assessment Entrance Criteria)

System 1
Interaction
Input, Output
Interface
Objects, Attributes

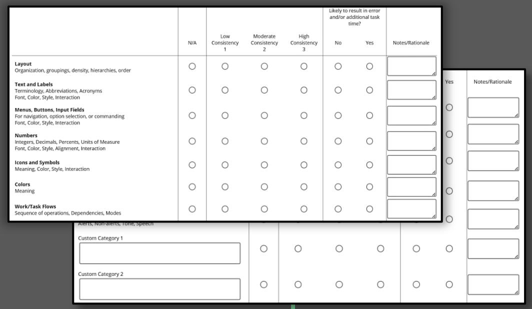
External Consistency Testing Methods

System 2
Interaction
Input, Output
Interface
Objects, Attributes

Consistency taxonomy applied to multiple system interfaces.

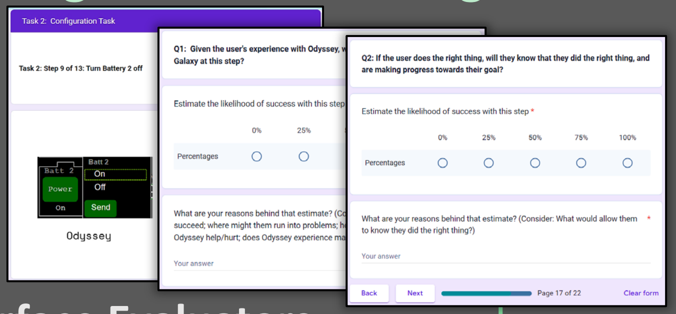
Phase I: Tool Development

Checklist



	N/A	Low Consistency 1	Moderate Consistency 2	High Consistency 3	Likely to result in error and/or additional task time? No	Yes	Notes/Rationale
Layout Organization, groupings, density, hierarchies, order	○	○	○	○	○	○	
Text and Labels Terminology, Abbreviations, Acronyms Font, Color, Style, Interaction	○	○	○	○	○	○	
Menus, Buttons, Input Fields For navigation, option selection, or commanding Font, Color, Style, Interaction	○	○	○	○	○	○	
Numbers Integers, Decimals, Percents, Units of Measure Font, Color, Style, Alignment, Interaction	○	○	○	○	○	○	
Icons and Symbols Meaning, Color, Style, Interaction	○	○	○	○	○	○	
Colors Meaning	○	○	○	○	○	○	
Work/Task Flows Sequence of operations, Dependencies, Modes	○	○	○	○	○	○	
Alerts, Non-alerts, Tone, Speech							
Custom Category 1	○	○	○	○	○	○	
Custom Category 2	○	○	○	○	○	○	

Cognitive Walkthrough



Task 2: Configuration Task

Task 2: Step 9 of 13: Turn Battery 2 off

Q1: Given the user's experience with Odyssey, w Galaxy at this step?

Estimate the likelihood of success with this step

	0%	25%
Percentages	○	○

What are your reasons behind that estimate? (Co succeed; where might them run into problems; h Odyssey help/hurt; does Odyssey experience ma

Your answer

Q2: If the user does the right thing, will they know that they did the right thing, and are making progress towards their goal?

Estimate the likelihood of success with this step *

	0%	25%	50%	75%	100%
Percentages	○	○	○	○	○

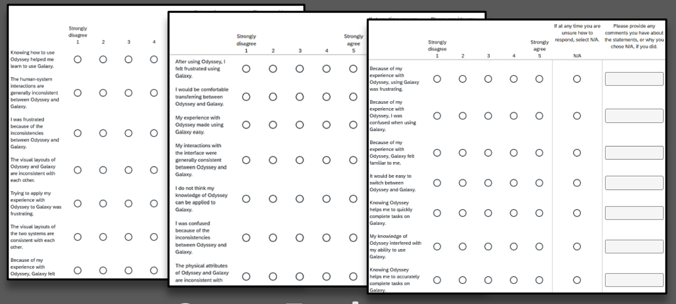
What are your reasons behind that estimate? (Consider: What would allow them to know they did the right thing?) *

Your answer

Back Next Page 17 of 22 Clear form

Expert Interface Evaluators

Intersystem Consistency Scale



	Strongly disagree 1	2	3	4
Knowing how to use Odyssey helped me learn to use Galaxy.	○	○	○	○
The human-system interactions are generally inconsistent between Odyssey and Galaxy.	○	○	○	○
I was frustrated because of the inconsistencies between Odyssey and Galaxy.	○	○	○	○
The visual layouts of Odyssey and Galaxy are inconsistent with each other.	○	○	○	○
Trying to apply my experience with Odyssey to Galaxy was frustrating.	○	○	○	○
The visual layouts of the two systems are consistent with each other.	○	○	○	○
Because of my experience with Odyssey, Galaxy felt	○	○	○	○

	Strongly disagree 1	2	3	4	Strongly agree 5
After using Odyssey, I felt frustrated using Galaxy.	○	○	○	○	○
I would be comfortable transferring between Odyssey and Galaxy.	○	○	○	○	○
My experience with Odyssey made using Galaxy easy.	○	○	○	○	○
My interactions with the interface were generally consistent between Odyssey and Galaxy.	○	○	○	○	○
I do not think my knowledge of Odyssey can be applied to Galaxy.	○	○	○	○	○
I was confused because of the inconsistencies between Odyssey and Galaxy.	○	○	○	○	○
The physical attributes of Odyssey and Galaxy are inconsistent with	○	○	○	○	○

	Strongly disagree 1	2	3	4	Strongly agree 5	If at any time you are unsure how to respond, select N/A. N/A	Please provide any comments you have about the statements, or why you chose N/A, if you did.
Because of my experience with Odyssey, using Galaxy was frustrating.	○	○	○	○	○	○	
Because of my experience with Odyssey, I was confused when using Galaxy.	○	○	○	○	○	○	
Because of my experience with Odyssey, Galaxy felt familiar to me.	○	○	○	○	○	○	
It would be easy to switch between Odyssey and Galaxy.	○	○	○	○	○	○	
Knowing Odyssey helps me to quickly complete tasks on Galaxy.	○	○	○	○	○	○	
My knowledge of Odyssey interfered with my ability to use Galaxy.	○	○	○	○	○	○	
Knowing Odyssey helps me to accurately complete tasks on Galaxy.	○	○	○	○	○	○	

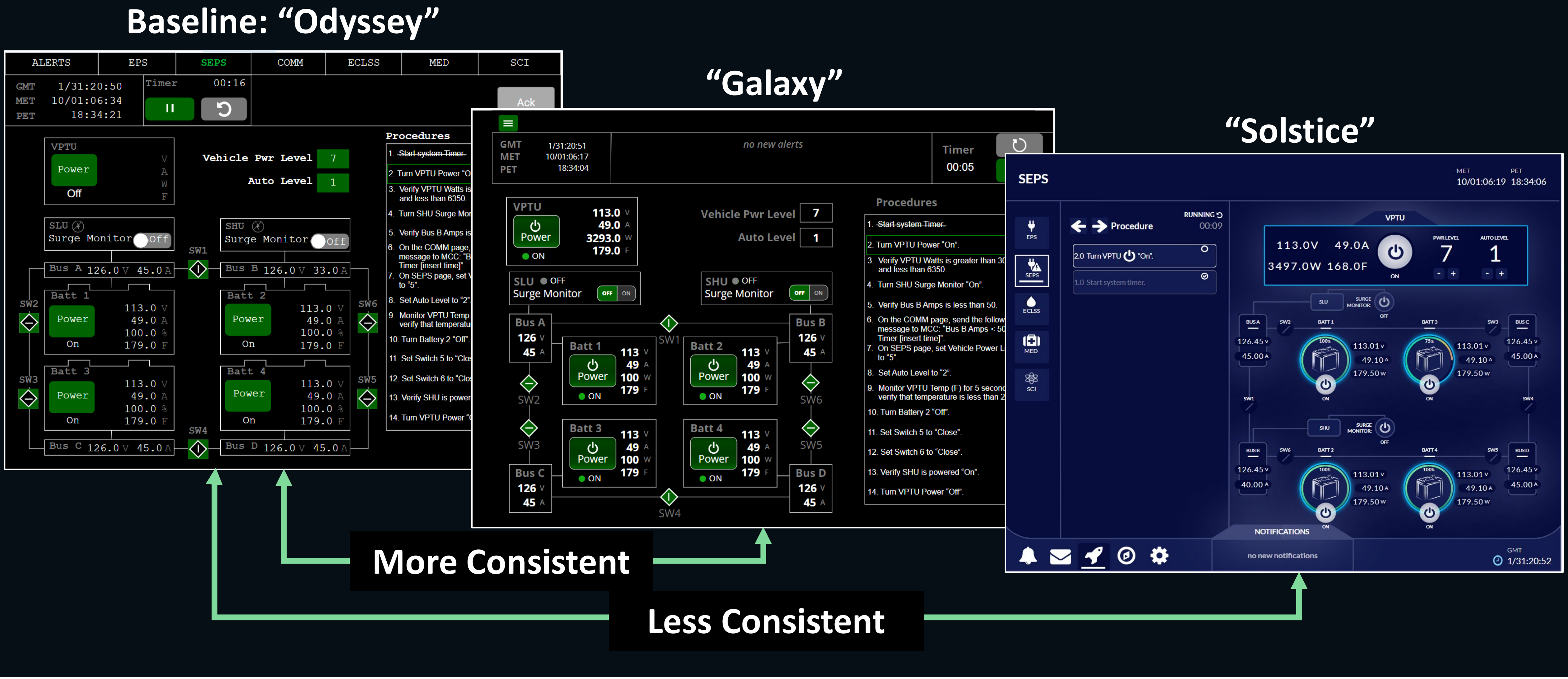
Crew Evaluators

- Desired quick, usable methods with clear criteria to yield quantitative scores and rationales
- Taxonomy used as a framework for the development of three assessment tools:
  - **Checklist:** Consistency ratings (1 = low to 3 = high) averaged across 12 categories consistency; option for up to two additional custom categories
  - **Cognitive Walkthrough:** Likelihood of success (0, 25, 50, 75, 100%) ratings averaged across 15 action tasks for 1) knowing what to do and 2) will the user know they did the right action
  - **Intersystem Consistency Scale:** Consistency ratings averaged across 21 comparison statements (1 = strongly disagree to 5 = strongly agree)

Phase I: Pilot Study

- Pilot study conducted to evaluate **performance** and **subjective impressions** of the three tools
- Participants recruited from the JSC Human Test Subjects Facility and word of mouth
    - **Checklist:** *N* = 6
    - **Cognitive Walkthrough:** *N* = 6
    - **ICS:** *N* = 13
- Participants completed the same tasks (i.e., electrical system configuration, chat, alerts, procedures) across all three prototypes
- Prototypes reflected features representative of “Artemis-like” interfaces
- Attributes were manipulated to be more or less consistent relative to a baseline design (e.g., color, layout, icons, symbols, system navigation)

Phase I: Prototypes



Baseline: “Odyssey”

GMT 1/31:20:50
MET 10/01:06:34
PET 18:34:21
Timer 00:16

VPTU Power Off
Vehicle Pwr Level 7
Auto Level 1
SLU Surge Monitor Off
SHU Surge Monitor Off
Bus A 126.0 V 45.0 A
Bus B 126.0 V 33.0 A
Batt 1 Power On 113.0 V 49.0 A 100.0 % 179.0 F
Batt 2 Power On 113.0 V 49.0 A 100.0 % 179.0 F
Batt 3 Power On 113.0 V 49.0 A 100.0 % 179.0 F
Batt 4 Power On 113.0 V 49.0 A 100.0 % 179.0 F
Bus C 126.0 V 45.0 A
Bus D 126.0 V 45.0 A

Procedures

1. ~~Start system Timer.~~
2. Turn VPTU Power “O
3. Verify VPTU Watts is and less than 6350.
4. Turn SHU Surge Mon
5. Verify Bus B Amps is
6. On the COMM page, message to MCC: “B Timer [insert time]”.
7. On SEPS page, set V to “5”.
8. Set Auto Level to “2”
9. Monitor VPTU Temp verify that temperatu
10. Turn Battery 2 “Off”.
11. Set Switch 5 to “Clos
12. Set Switch 6 to “Clos
13. Verify SHU is power
14. Turn VPTU Power “

“Galaxy”

GMT 1/31:20:51
MET 10/01:06:17
PET 18:34:04

no new alerts

Timer 00:05

VPTU Power ON 113.0 V 49.0 A 3293.0 W 179.0 F
Vehicle Pwr Level 7
Auto Level 1
SLU OFF Surge Monitor OFF ON
SHU OFF Surge Monitor OFF ON
Bus A 126 V 45 A
Bus B 126 V 45 A
Batt 1 Power ON 113 V 49 A 100 W 179 F
Batt 2 Power ON 113 V 49 A 100 W 179 F
Batt 3 Power ON 113 V 49 A 100 W 179 F
Batt 4 Power ON 113 V 49 A 100 W 179 F
Bus C 126 V 45 A
Bus D 126 V 45 A

Procedures

1. ~~Start system Timer.~~
2. Turn VPTU Power “On”.
3. Verify VPTU Watts is greater than 30 and less than 6350.
4. Turn SHU Surge Monitor “On”.
5. Verify Bus B Amps is less than 50.
6. On the COMM page, send the follow message to MCC: “Bus B Amps < 50 Timer [insert time]”.
7. On SEPS page, set Vehicle Power L to “5”.
8. Set Auto Level to “2”.
9. Monitor VPTU Temp (F) for 5 second verify that temperature is less than 2
10. Turn Battery 2 “Off”.
11. Set Switch 5 to “Close”.
12. Set Switch 6 to “Close”.
13. Verify SHU is powered “On”.
14. Turn VPTU Power “Off”.

“Solstice”

SEPS

MET 10/01:06:19
PET 18:34:06

Procedure
RUNNING 00:09

2.0 Turn VPTU “On”. ○
1.0 Start system timer. ✓

VPTU 113.0V 49.0A 3497.0W 168.0F ON
PWR LEVEL 7
AUTO LEVEL 1

SLU SURGE MONITOR: OFF
SHU SURGE MONITOR: OFF
BUS A 126.45 V 45.00 A
BATT 1 100% 113.01 V 49.10 A 179.50 W ON
BATT 3 75% 113.01 V 49.10 A 179.50 W ON
BUS C 126.45 V 45.00 A
BUS B 126.45 V 40.00 A
BATT 2 100% 113.01 V 49.10 A 179.50 W ON
BATT 4 100% 113.01 V 49.10 A 179.50 W ON
BUS D 126.45 V 45.00 A

NOTIFICATIONS
no new notifications

GMT 1/31:20:52

More Consistent

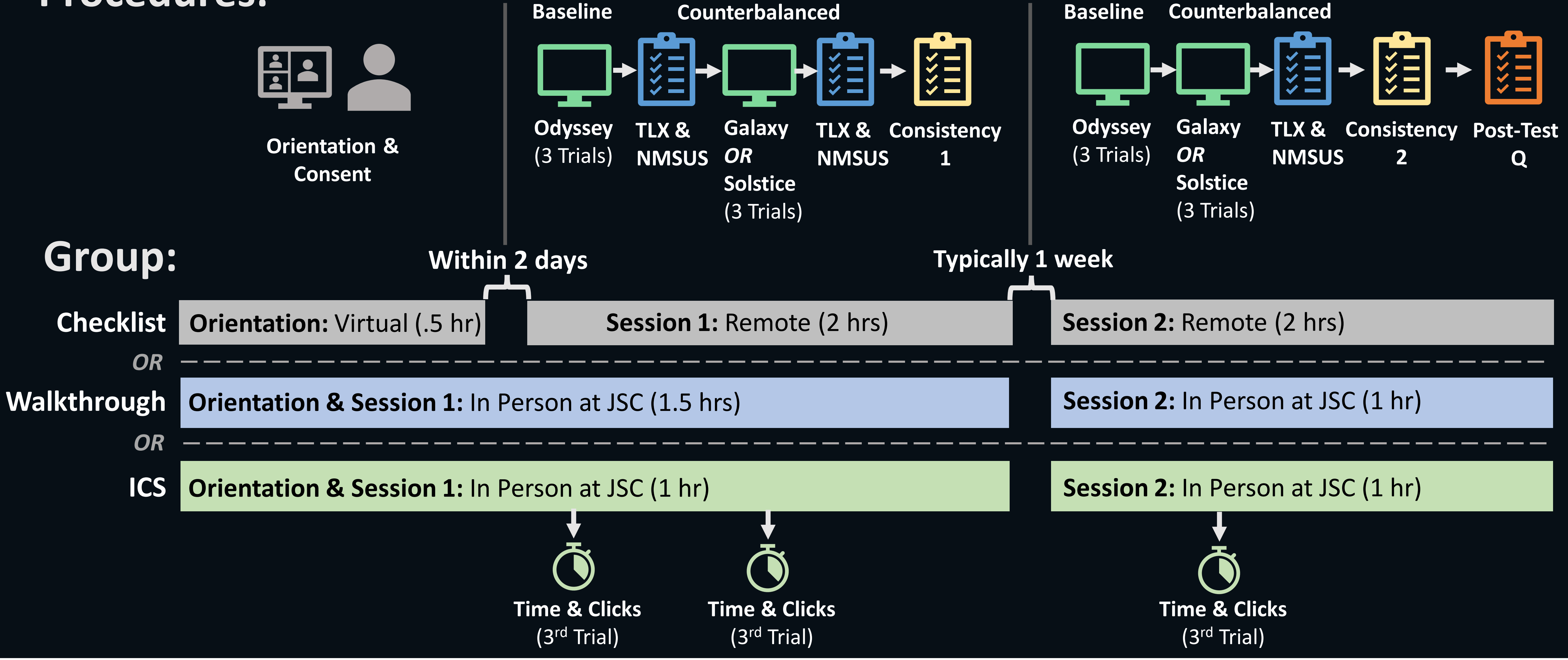
Less Consistent

Amanda Smith | amanda.l.smith-1@nasa.gov

Phase I: Pilot Study

- **Prototype** was the within-subjects factor
- The **assessment tool** was the between-subjects factor
- **Dependent Measures:**
  - **Subjective prototype usability**: NASA Modified System Usability Scale (NMSUS)
  - **Subjective workload** to complete tasks on each prototype: NASA-Task Load Index (NASA-TLX)
  - **Consistency scores for prototype pairs:**
    - Checklist, *OR*
    - Cognitive Walkthrough, *OR*
    - Intersystem Consistency Scale (ICS)
  - **Performance measures:**
    - Time on Task
    - Click Frequency (used to calculate click errors, or deviation from optimal clicks)
  - **Post-Test Survey** to assess usability and utility of the assessment tool

Procedures:



Orientation & Consent

Baseline: Odyssey (3 Trials) → TLX & NMSUS → Counterbalanced: Galaxy *OR* Solstice (3 Trials) → TLX & NMSUS → Consistency 1

Baseline: Odyssey (3 Trials) → Counterbalanced: Galaxy *OR* Solstice (3 Trials) → TLX & NMSUS → Consistency 2 → Post-Test Q

Group:

Within 2 days

Typically 1 week

Checklist: **Orientation:** Virtual (.5 hr) | **Session 1:** Remote (2 hrs) | **Session 2:** Remote (2 hrs)

OR

Walkthrough: **Orientation & Session 1:** In Person at JSC (1.5 hrs) | **Session 2:** In Person at JSC (1 hr)

OR

ICS: **Orientation & Session 1:** In Person at JSC (1 hr) | **Session 2:** In Person at JSC (1 hr)

Time & Clicks (3rd Trial)

Time & Clicks (3rd Trial)

Time & Clicks (3rd Trial)

Phase I: Pilot Study High-level Results

Assessment Tools Overall

- All tools were rated user-friendly by participants
- Participant ratings with each tool reflected the same trends: that Galaxy (high consistency) had a higher degree of similarity to Odyssey (Baseline) than did Solstice (low consistency)
- Intentional design differences were identified in rationale statements with varying degrees of specificity
- Participants provided feedback to improve the tools

Phase I: Pilot Study High-level Results

Checklist

- **Interrater Reliability:** Participants demonstrated moderate agreement in their consistency ratings for Odyssey-Solstice, but poor agreement for Odyssey-Galaxy.
  - Odyssey-Solstice: ICC(A,6) = 0.69 (95% CI 0.33-0.90), $F = 3.79$ (11, 36.3), $p = 0.001$.
  - Odyssey-Galaxy: ICC(A,5) = 0.30 (95% CI 0 – 0.77), $F = 1.41$ (11, 58.6), $p > 0.05$.
- **Validity:** compared to Odyssey, participants rated Galaxy (Mdn = 2.67) as significantly more consistent than Solstice (Mdn = 1.70), $W = 74.50$, $p < 0.01$.
  - The checklist correctly discriminated between intentional high/low consistency manipulations.

Phase I: Pilot Study High-level Results

Cognitive Walkthrough

- **Interrater Reliability:** Participants demonstrated moderate agreement in their likelihood for success ratings for Odyssey-Galaxy, but poor agreement for Odyssey-Solstice.
    - Odyssey-Galaxy: ICC(A,6) = 0.55 (95% CI 0.25, 0.76), $F = 2.26$ (29, 149), $p < 0.001$.
    - Odyssey-Solstice: ICC(A,6) = 0.22 (95% CI -0.18, 0.55), $F = 1.39$ (29, 74.7), $p > 0.05$.
- **Validity:** compared to Odyssey, participants rated likelihood of success as significantly higher for Galaxy (Mdn = 0.98) than for Solstice (Mdn = 0.88), $W = 28.00$, $p < 0.001$.
    - The cognitive walkthrough correctly discriminated between high/low consistency manipulations.

Phase I: Pilot Study High-level Results

Intersystem Consistency Scale

- On average, participants rated Galaxy as more consistent with Odyssey (*M* = 4.18, *SD* = 0.66) than Solstice (*M* = 3.22, *SD* = 0.79)
- **Reliability:** Both the consistency and inconsistency scales met the minimum recommendation for an internally reliable scale (α ≥ .70)
- **Discriminant Validity:** Both the consistency and inconsistency scales demonstrated sensitivity to similarities and differences, respectively, in prototype designs
  - **Consistency Scale Scores:** *t*(12) = 5.25, *p* < .001, Cohen's *d* = 1.46
  - **Inconsistency Scale Scores:** *t*(12) = 5.25, *p* < .001, Cohen's *d* = 1.46

Phase I: Pilot Study High-level Results

- **Usability**
  - Median NMSUS scores were all within “good” equity thresholds (Bangor et al., 2009)
  - Odyssey and Galaxy were at the higher end and Solstice scores were at the lower end of this range
- **Workload**
  - Compared to Odyssey, participants generally reported lower workload (NASA-TLX scores) to complete tasks with Galaxy and higher workload for Solstice
  - Significant differences observed for mental, effort, frustration and performance dimensions
- **Performance**
  - Compared to Odyssey, participants generated significantly more click errors when using Solstice than when using Galaxy
  - Time on task was essentially the same across all prototypes, though Solstice was designed to require fewer clicks (and hypothetically less time) to complete the task

Phase II: Risk Assessment, Standards & Guidelines

Phase II: Risk Assessment, Standards & Guidelines

Phase II builds on work completed in Phase I, with the focus of “*when does inconsistency matter*”?

- **Aim 1: Refine assessment tools** using feedback and results obtained in Phase I and provide practitioner documentation (user guides).
- **Aim 2: Refine tasks and prototype stimuli** using Program human interface and training Subject Matter Expert inputs.
- **Aim 3: Complete a risk assessment study** to understand differences in subjective feedback and human performance (including transfer of training) related to each specific type of inconsistency (i.e., interface, interaction).
- **Aim 4: Propose a set of standards and guidelines** assessing aspects of inconsistent design that appear to have the most impact on performance.

Phase II: Aim 1

Assessment Tool Updates

- Post-processed user comments to assess coverage of intentional design manipulations
- **Checklist** selected as the expert evaluator assessment tool
  - Increased clarity of categories
  - Added global rating scale
  - Added method for annotating screenshots to increase comment specificity
  - **Next Steps:** Assess performance of the updated tool and create practitioner guide
- **ICS updates**
  - Reduced redundant items
  - Added global severity scale
  - **Next Steps:** Crew interview to obtain feedback on changes; create practitioner guide once complete

Phase II: Aim 2

Task and Prototype Refinement

- Interviewed two Artemis training SMEs and four Program display SMEs for awareness of tasks shared across vehicles, interface characteristics, and cross-training expectations
- Finalized study task set to include audio configuration, alerts, procedures, and electrical system configuration
- Ensured common interface attributes are represented across the prototype: labels, readouts, interactive widgets, icons/symbols, color-coding, schematics, menus, timers
- Incorporated guidance from the Artemis GUI standard
- **Next steps:** finish design, coding and test usability of each prototype

Phase II: Aim 3

Risk Assessment Study

- Goal is to assess differences in our DVs associated with prototypes that reflect different levels of consistency, relative to our baseline prototype
- Five total prototypes, manipulated at the taxonomy level (interface, interaction)
  - **Interface manipulations:** color, symbols/icons, schematics/layout
  - **Interaction manipulations:** information architecture, workflow, procedures, input style

Prototype A (Baseline)	Interface Similar	Interface Dissimilar
Interaction Similar	Prototype B	Prototype C
Interaction Dissimilar	Prototype D	Prototype E

Phase II: Aim 3

Risk Assessment Study

- Study design will be similar to Phase I but with some key differences:
  - Collecting data with “crew-like” participants and using the ICS only
  - Between-subjects (single pair of prototypes for comparison)
  - In addition to workload, usability, time and clicks, we plan to add eye tracking measures as a DV
  - Adding an activity to assess time and accuracy for comprehending low-level interface attributes (e.g., color coding, icons, symbols, etc.)
  - Interested in baseline performance and performance after switching between prototypes

- **Next steps:** finalize study design and start the recruitment process

Phase II: Aim 4

Standards and Guidelines

- A set of consistency-related standards and guidelines were collected from government, agency, and industry sources during Phase I
- Results from the Risk Assessment Study will provide insights to understanding which types of consistency are associated with greater performance differences
- Study results will inform selection of recommended standards and guidelines suitable for submission for NASA-STD-3001 and other program documents

Thank you!

Contact: Amanda.L.Smith-1@nasa.gov