Measuring the Effectiveness of Human Autonomy Teaming

Thomas Z. Strybel, Jillian Keeler, Natassia Mattoon, Armando Alvarez, Vanui Barakezyan, Edward Barraza, James Park, Kim-Phuong L. Vu
California State University Long Beach

Vernol Battiste
San Jose State University Foundation at NASA Ames Research Center

Human-Autonomy Teaming (HAT)

- The interdependent coupling between human operators and autonomous systems
  - *requires collaboration and coordination to accomplish system and task goals*
- Proposed tenets of HAT
  - *Bi-Directional communication*
  - *Transparency*
  - *Operator-directed interface*
- Overcome issues of
  - *Out-of-loop problems*
  - *Miscalibrated trust*
  - *Automation brittleness*

Evaluating HAT Effectiveness

- System
  - *Effectiveness in achieving mission goals*
  - *System outcomes (safety, efficiency)*
- Operator
  - *Operator performance*
  - *Operator situation awareness, workload*
  - *Operator trust and reliance on automation*
- Automation
  - *Automation performance*
  - *Automation situation awareness*
- Collaboration
  - *Shared awareness*
  - *Transparency*
  - *Bi-Directional communication*

Evaluating HAT Effectiveness

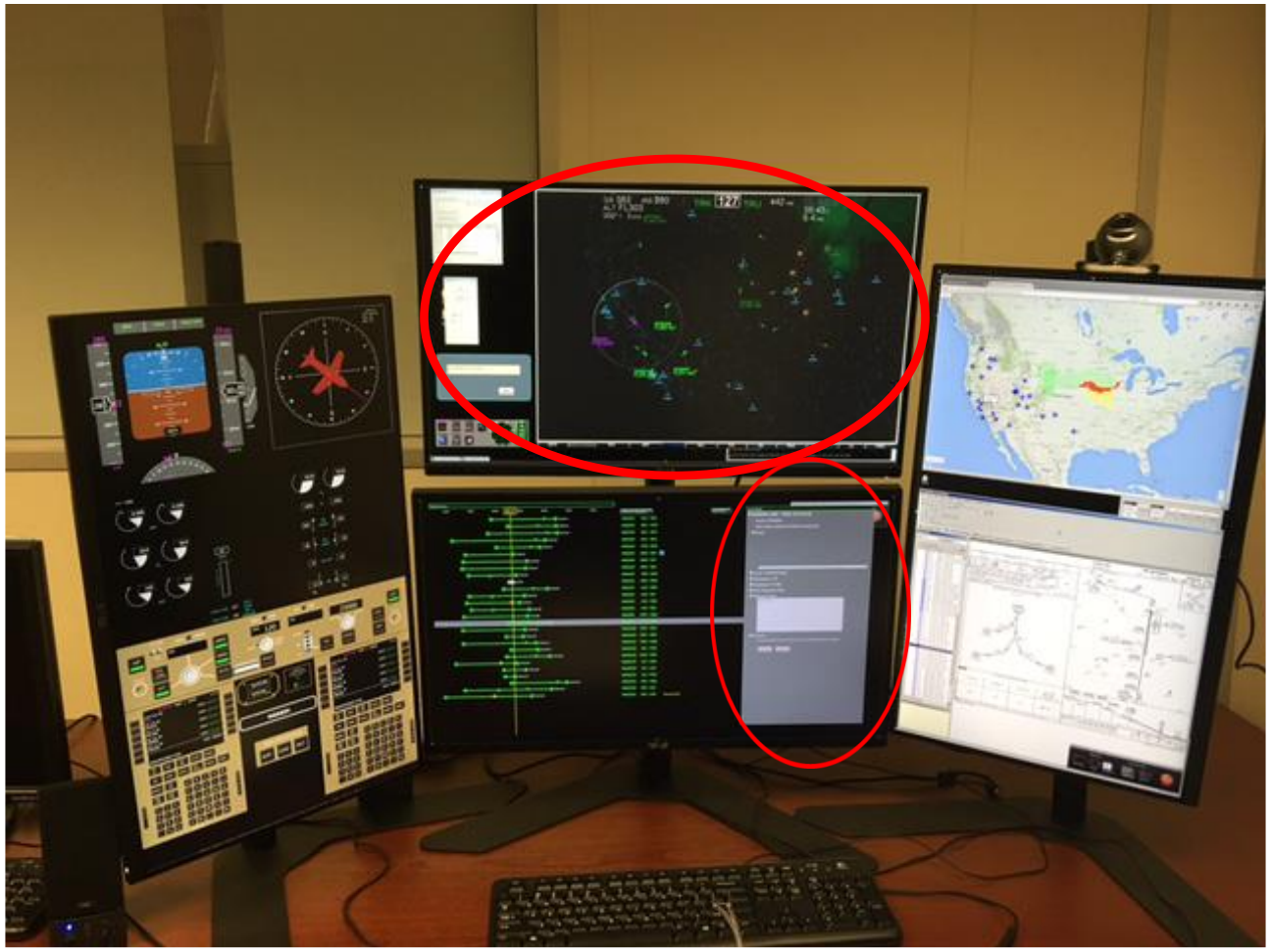
- Evaluations must include
  - *Operators with varying skill levels (re: automation and operations)*
  - *Nominal and off-nominal situations*
  - *Long term usage*
- Evaluations must have sufficient statistical power for detecting changes in performance and behaviors
  - *Simulations, scenarios, tasks*
- Measures of HAT effectiveness must include
  - *Subjective responses*
  - *Behavior (operator and automation)*
  - *Performance (system, operator, automation)*
- Compare alternative designs and concepts

Purpose

- Work with NASA HAT Lab (Brandt et al., previous presentation) to look for metrics for comparing operator performance and behavior with and without HAT tools
  - *Subjective workload during the scenario - determine if HAT interactions changed operator workload*
  - *Resolution time and eye gaze - determine the extent to which operators actually used the HAT tools*

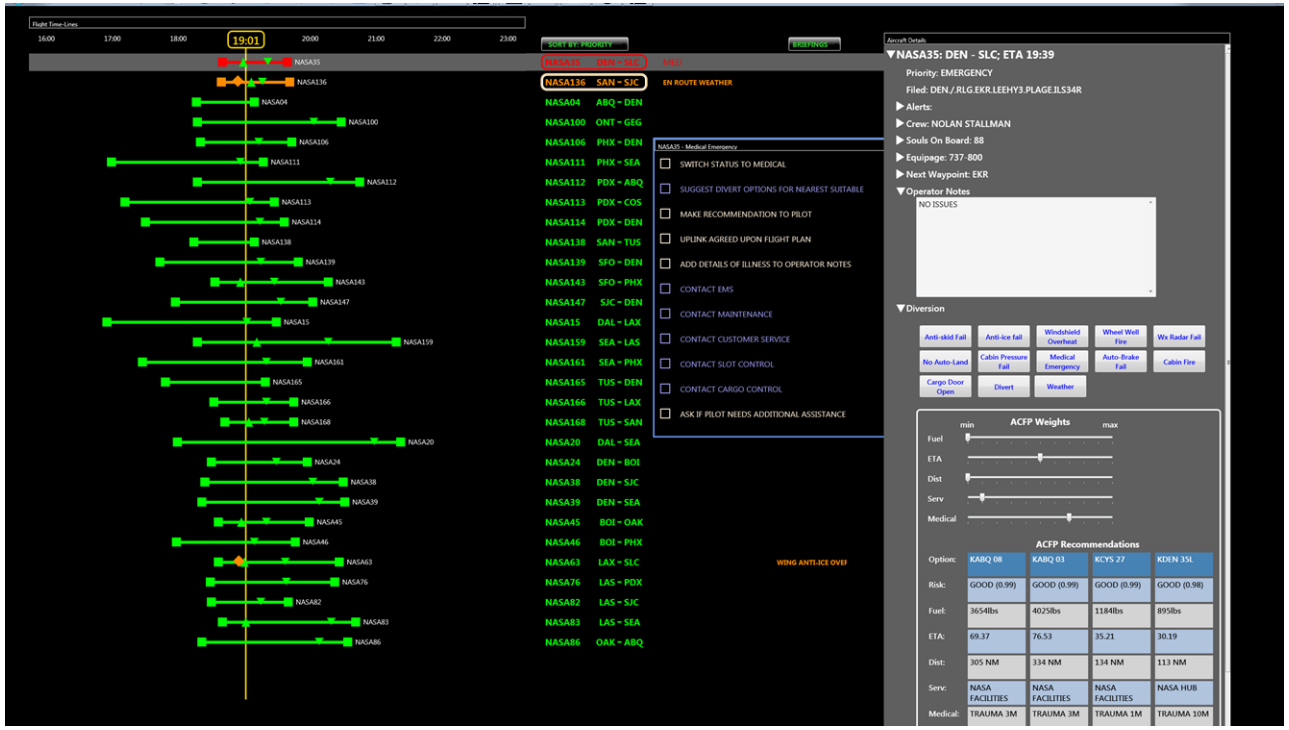
- Six participants assumed the role of an “advanced dispatcher.”
  - *Flight-following task supporting aircraft with single onboard pilots operating in high workload and off-nominal situations.*
  - *Two scenarios (HAT and No HAT), containing approximately 30 aircraft, and 6 off-nominal events.*

Method: Ground Station Layout



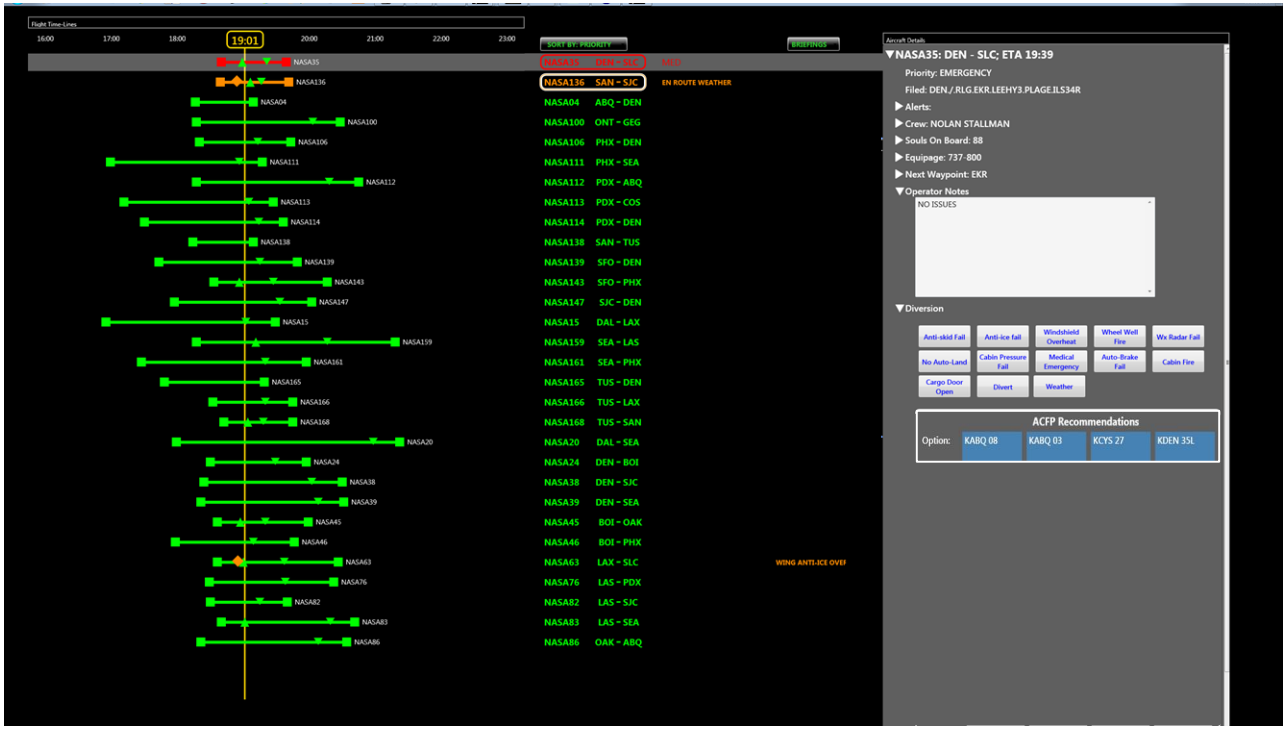
Tested two configurations of ACFP and TSD: HAT and No HAT

HAT ACFP Information and Display



- Plays
- Automated Checklist
- ACFP Recommendations
  - Factors Involved in Recommendation
  - Adjustable Weights

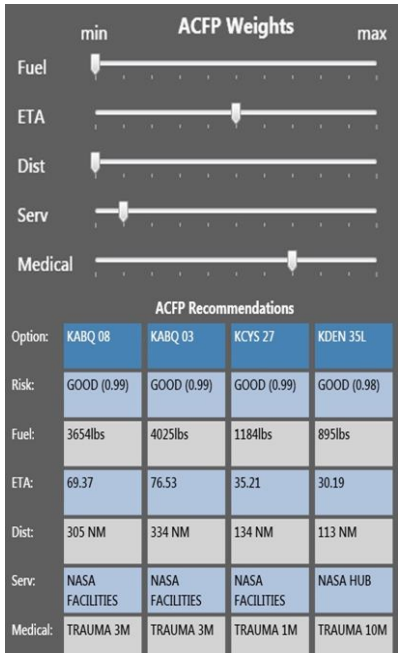
No HAT ACFP Information and Display



- Plays
- ACFP Recommendations
  - No reasoning
  - No automated checklist
  - No weight adjustments

ACFP HAT vs. No HAT

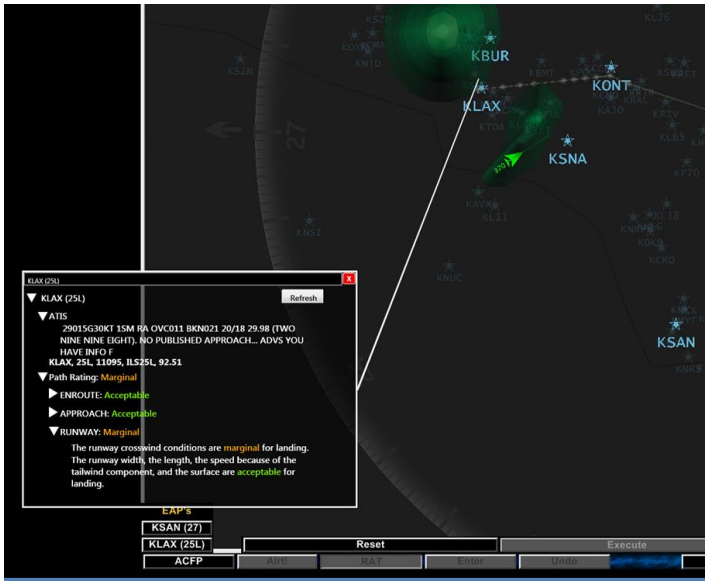
HAT



No HAT

TSD: HAT vs. No HAT

HAT: – ATIS and reasoning



No HAT – ATIS only

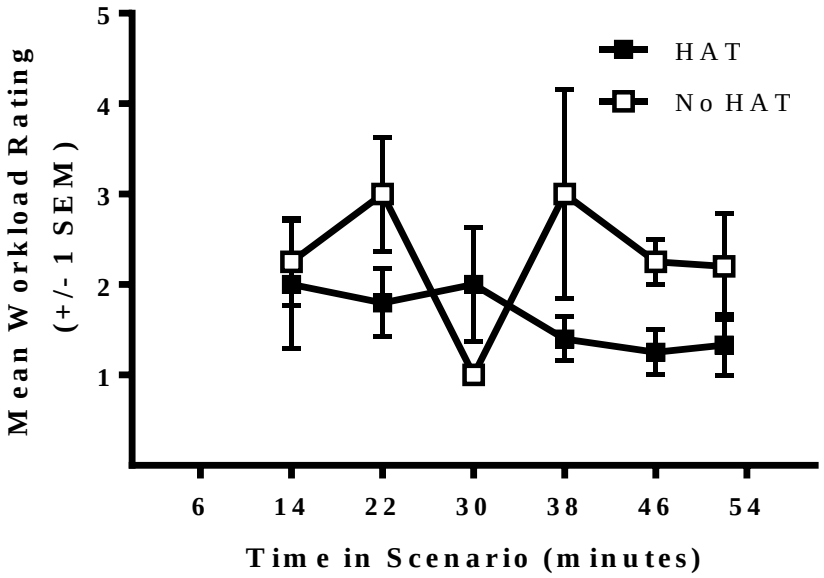
Scenario Events Requiring Dispatcher Assistance

Aircraft Event	Description
Fire in Lavatory	Fire ignited in the lavatory; immediate landing is likely.
Airport Weather	Weather at destination airport is near or below minimums. If below, pilot must divert to a suitable nearby airport.
Wheel Well Fire	Fire detected in the main wheel well shortly after takeoff.
Medical Emergency	Passenger onboard an aircraft requires medical attention and possibly immediate landing based on severity of condition.
Anti-Skid Inoperative	Antiskid prevents wheels from skidding during braking by minimizing speed difference between wheel speed and aircraft speed.
Windshield Overheat	Windshield heating system has malfunctioned and may cause damage to the windshield.
Aft Cargo Door Open	One or more cargo doors are not closed and secure; detrimental if aircraft is above 8,000 ft.
Weather Radar Fail	Failure prevents cockpit crew from viewing weather near the aircraft.

Metrics for Operator Behavior and Performance

- Subjective workload ratings obtained at regular intervals throughout the scenario (1=very low workload, 5=very high workload)
- Resolution time – measured for those events in which a definite starting and ending point could be seen
- Eye gaze duration - measured throughout the scenario by means of cameras mounted on each display
  - *Time spent on ACFP and TSD (HAT vs. No HAT) might indicate the extent to which the operator is using the HAT tools*
  - *Time spent on other displays (HAT vs. No HAT) might indicate the extent to which operator is verifying the ACFP recommendations*
- Slider use – weight adjustments

Results: Mean Workload Ratings as a Function of Time in the Scenario



Resolution Time and Time Spent on Ground Station Displays

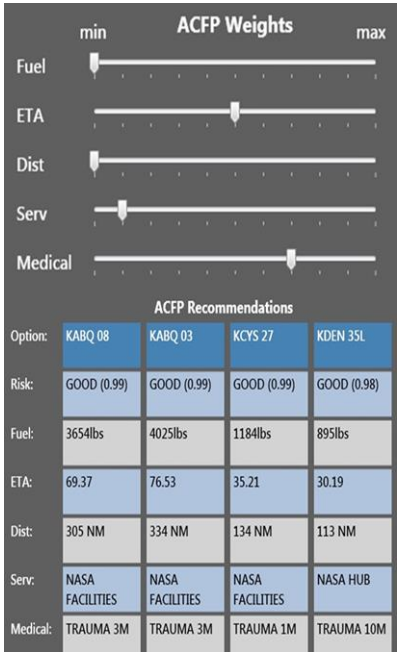
Summary Statistic	HAT	No HAT	p*
Mean (SEM) number flight plan changes	3.3 (0.8)	3.0 (0.6)	.78
Mean (SEM) time per flight plan change (s)	101.2 (19.4)	63.7 (14.9)	.026
Mean (SEM) gaze time on ACFP (s)**	53.6 (9.5)	27.5 (3.5)	.009
Mean (SEM) gaze time on TSD (s)	38 (14.0)	20.7 (11.0)	.10
Mean (SEM) gaze time on other displays (s)	9.6 (4.5)	15.5 (6.3)	.06

**probability of obtained difference based on repeated measures t test (df = 5)*

** Mean (SEM) gaze time on ACFP when sliders not moved in HAT condition (s)	41.9 (8.7)	27.5 (3.5)	.08

Did the Operators Adjust the Factor Weights in HAT Condition?

Event	N Participants (out of 6)	Factor(s) Adjusted
Fire in Lavatory	5	Distance
Airport Weather	2	Distance, ETA
Wheel Well Fire	1	ETA
Medical Emergency	1	Distance, ETA



Preliminary Results Summary

- Operator workload was lower in the HAT condition and decreased with time in the scenario
  - *Consistent with Brandt et al. post scenario workload ratings*
- Operators took more time to uplink flight-plan recommendations in the HAT condition
  - *HAT: relatively more time looking at HAT displays (TSD, ACFP)*
  - *No Hat: relatively more time looking at other displays (flight instruments, JEP Charts, CONUS)*
- Did additional time result in better resolutions?

Conclusion

- Measuring HAT effectiveness in terms of performance and behavior is necessary but challenging

- Design simulations and scenarios to elicit differences in behavior and performance
  - *Identify behaviors*
  - *Identify performance metrics*

- One possible solution – a testbed for testing HAT concepts and designs
  - *Generic, airspace/aviation related*
  - *Scenarios should be easily manipulated*
  - *Should be sensitive to changes in operator and system performance*

Thank You

Resolution Time and Time Spent on Ground Station Displays

Summary Statistic %	HAT	No HAT	p*
Mean (SEM) number flight plan changes	3.3 (0.8)	3.0 (0.6)	.78
Mean (SEM) time per flight plan change (s)	101.2 (19.4)	63.7 (14.9)	.026
% gaze time on ACFP	59%	50%	.15
% gaze time on TSD	32%	29%	.56
% gaze time on other displays	9%	21%	.06

**probability of obtained difference based on repeated measures t test (df = 5)*