DETECTION OF OPERATOR PERFORMANCE BREAKDOWN AS AN AUTOMATION TRIGGERING MECHANISM

Hyo-Sang Yoo, San Jose State University / NASA Ames Research Center, Moffett Field, CA

Paul U. Lee, NASA Ames Research Center, Moffett Field, CA

Steven J. Landry, Purdue University, West Lafayette, IN

Abstract

Performance breakdown (PB) has been anecdotally described as a state where the human operator "loses control of context" and "cannot maintain required task performance." Preventing such a decline in performance is critical to assure the safety and reliability of human-integrated systems, and therefore PB could be useful as a point at which automation can be applied to support human performance. However, PB has never been scientifically defined or empirically demonstrated. Moreover, there is no validated objective way of detecting such a state or the transition to that state. The purpose of this work is: 1) to empirically demonstrate a PB state, and 2) to develop an objective way of detecting such a state. This paper defines PB and proposes an objective method for its detection.

A human-in-the-loop study was conducted: 1) to demonstrate PB by increasing workload until the subject reported being in a state of PB, and 2) to identify possible parameters of a detection method for objectively identifying the subjectively-reported PB point, and 3) to determine if the parameters are idiosyncratic to an individual/context or are more generally applicable. In the experiment, fifteen participants were asked to manage three concurrent tasks (one primary and two secondary) for 18 minutes. The difficulty of the primary task was manipulated over time to induce PB while the difficulty of the secondary tasks remained static. The participants' task performance data was collected. Three hypotheses were constructed: 1) increasing workload will induce subjectively-identified PB, 2) there exists criteria that identifies the threshold parameters that best matches the subjectively-identified PB point, and 3) the criteria for choosing the threshold parameters is consistent across individuals. The results show that increasing workload can induce subjectively-identified PB, although it might not be generalizable—only 12 out of 15 participants declared PB. The PB detection method based on signal detection analysis was applied to the performance data and the results showed that PB can be identified using the method, particularly when the values of the parameters for the detection method were calibrated individually.

Introduction

Anecdotally, most people are familiar with the sensation where, during a task with very high workload, a state is reached where the operator goes "hands off" and completely drops the primary task. Such an extreme state is referred to here as performance breakdown (PB). It is important to prevent such a state from being reached, particularly in a safety critical system that requires a human operator to assure the safety and reliability of the system's operations. If PB can be detected in advance, then it can be prevented from occurring by allowing the automation system to intervene and assist or replace the human operator. However, PB has been only anecdotally described in past research, such as PB occurs when task demand exceeds resource capacity [1]. Also, PB has never really been scientifically identified or empirically demonstrated in an experimental setting. The work described in this paper contributes to filling those gaps and could potentially provide the ground work for future work on PB and its method of detection.

This paper is organized in the following way: 1) a definition of Performance Breakdown an objective method to detect it, 2) the method used for the human-in-the loop study, 3) the results obtained from conducting the study, 4) the discussion of the results, and 5) the conclusion of the study.

Performance Breakdown (PB)

The PB detection method distinguishes data into a binary form (PB vs. Non-PB) by setting the threshold on the selected measure for monitoring the human operator's state changes. The following describes the method in more detail, which could be used as a framework for detecting transition for other cognitive states as well.

PB occurs when the human operator fails to maintain minimally acceptable performance in a primary task for some minimum duration or longer.

$$(p < p_{crit}) \cap (\Delta t > \varepsilon) \quad (1)$$

In the equation (1) above, p refers to the human operator's performance on a specific task. p_{crit} is a minimally acceptable performance level for the task. ε indicates a maximum duration of time allowed for adjusting performance to maintain performance above the minimum performance level (p_{crit}). Δt is the continuous duration of time that an operator fails to maintain the minimum performance level (p_{crit}). Parameters (p_{crit}, ε) are most likely task specific, and may need to be defined by subject matter experts or be empirically determined.

In certain tasks, performance can also be computed as an error rate, i.e. the number of correct or incorrect responses during a fixed duration of time. In such cases, the equation can be modified accordingly. For example, the operator is asked to respond to twenty stimuli that are presented every two minutes. The total duration of the operation is thirty minutes. The operator's performance can be evaluated for every two-minute period by computing the error rate during that period. If the error rate exceeds the critical threshold value for an indicated duration of time, then PB is said to occur for that time period.

In addition to error rate, performance can also be evaluated based on error occurrences. For instance, the compliance of a pilot with a specified flight path could be considered the pilot's performance. In such a case, PB would be indicated if the pilot failed to keep the aircraft on the target route beyond the minimally acceptable deviation for a minimum period of time.

Previous work has indicated the potential sensitivity issues associated with using the threshold approach for detecting changes in the human's state [2][3]. Hence, three evaluation criteria are identified, which can be used to evaluate the efficacy of parameters (p_{crit}, ε) in detecting PB. The three evaluation criteria are: sensitivity, specificity, and delay time to detection. These criteria are commonly used parameters in signal detection analysis [4][5][6].

The sensitivity was computed using the following equation [7]:

$$Sensitivity = \frac{Total\ duration\ of\ true\ positive}{Total\ duration\ of\ true\ positive + Total\ duration\ of\ false\ negative} \quad (2)$$

In the equation above (2), the total duration of true positive (TP) indicates the time period that PB is correctly diagnosed as PB. The total duration of false negative (FN) represents the period when PB is incorrectly identified as not being PB (Non-PB). In the rest of the document, False Positive Rate (FPR) and Sensitivity are used interchangeably.

The specificity was calculated using the following equation [7]:

$$Specificity = \frac{Total\ duration\ of\ true\ negative}{Total\ duration\ of\ true\ negative + Total\ duration\ of\ false\ positive} \quad (3)$$

In the equation above (3), the total duration of true negative (TN) is the period that the Non-PB condition is correctly identified as Non-PB. The total duration of false positive (FP) is the period when Non-PB is incorrectly identified as PB. In the rest of the document, True Positive Rate (TPR) and (1 – Specificity) are used interchangeably.

Figure 1 depicts a nominal example of the false negative situation. In Figure 1, a tracking task with increasing task performance over time results in PB, shown as the red dotted line after 500 seconds. Once PB occurs in a task with increasing task difficulty, it should continue as long as no resolution action is made. However, from 700 seconds to 727 seconds, it is identified that there is Non-PB. This duration represents a false negative.

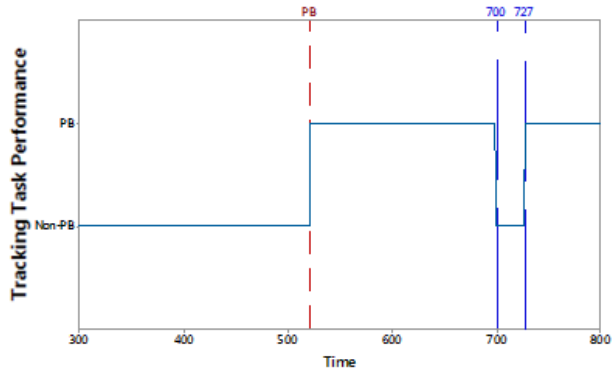

Figure 1. Nominal Example of False Negative

An example of a false positive is presented in the figure below (Figure 2). PB is shown to occur after 500 seconds. However, there is a time period (from 380 seconds to 399 seconds) that is identified as PB, even though the task difficulty would have been lower compared to the subsequent periods leading up to PB. That preceding time period represents a false positive.

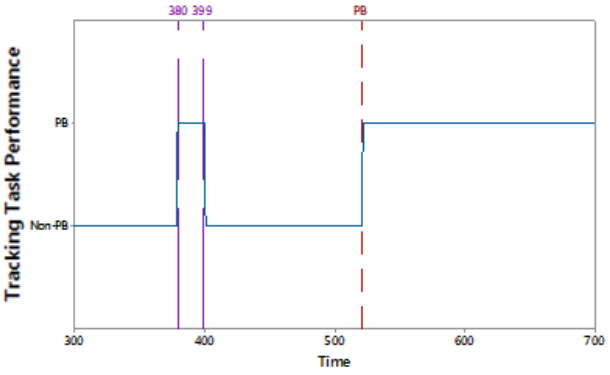

Figure 2. Nominal Example of False Positive

The delay time to detection is the period of time it takes from the point when PB occurs to the time the PB detection method detects PB. Having a large value for $\mathcal{E}$ is one of the major contributors for having a large delay time. When it is ambiguous to determine which values work the best for the parameters, this delay time could be used to identify the parameters.

A Receiver Operating Characteristic (ROC) curve can be constructed to investigate how various threshold values affect PB detection. The ROC curve helps determine the optimal threshold values that effectively balance the TPR and FPR [4]. Figure 3 shows an example of ROC curve. The curve is plotted by showing the TPR against FPR at various different combinations of threshold parameters ($P_{crit,}$ ε). Ideally, the optimal parameters would maximize TPR while guaranteeing the minimum FPR, which could be placed on the left top corner. In the figure, the numbers in the right upper corner indicate different values for P_{crit} and the number on top of each dot in the graph represents the value that has been tested for ε.

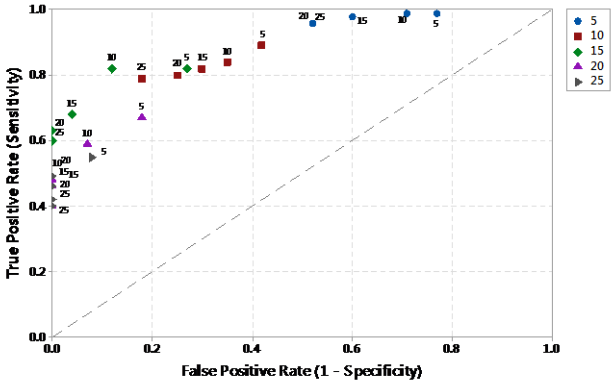

Figure 3. Nominal Example of ROC curve

In the ROC curve graph (Figure 3), the threshold values found with the shortest Euclidian distance to the left upper corner are sought to balance the competing characteristics most optimally (i.e., maximizes the TPR while minimizing FPR), which is referred to as *Criteria 1*. This could be applied in the system where the false detection and missed detection are equally important. In the figure (Figure 3) above, P_{crit} = 15, ε =10 are identified based on *Criteria 1*.

The combination of the threshold values that detect PB more conservatively can be also selected. The condition that shows the minimum FPR but had the highest TPR will be referred to as *Criteria 2* for the rest of the paper. *Criteria 2* could be applied to the situation where the impact of the missed detection is critical. In Figure 3, P_{crit} = 15, ε = 20 satisfy such criteria. The following sections present the human-in-the-loop study that was conducted to demonstrate PB and examine the proposed method for detecting PB.

Method

Participants

There were a total of 15 participants (13 male and 2 female). The age range of the participants was 23 – 34 years old. The participants had no prior experience performing the tasks.

Experimental tasks

The study required participants to perform three tasks concurrently (see Figure 4), which were the system monitoring task, the resource management task, and the tracking task from the latest version of Multi-Attribute Task Battery-II (MATB-II) [8]. These tasks are designed in a way that mimics the general operations of a pilot's tasks in the cockpit environment, which all required perceptual attention.

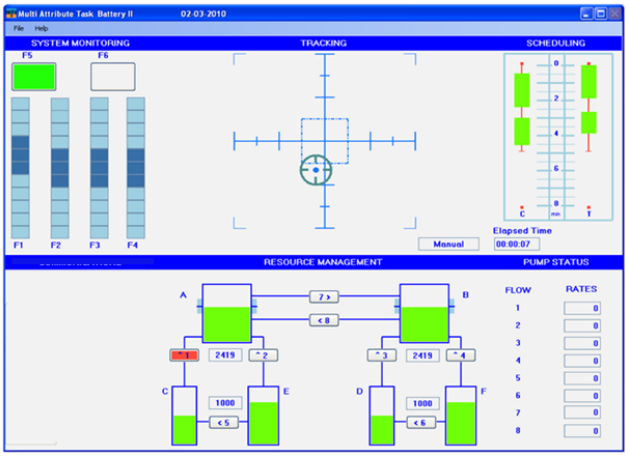

Figure 4. Screen Shot of MATB-II

Independent Variables

In this study, there were nine (3 X 3) different levels of difficulty of the primary task that increased in steps to induce PB. The task difficulty was determined by the combination of two parameters: 1) the target movement, and 2) the joystick response sensitivity level. The target update rate varied based on the amount of random target movement per update cycle and the joystick response sensitivity levels varied based on the amount of influence the joystick movement had on target movement per update cycle.

Table 1 shows the nine conditions that were created to induce a step-wise increase in task difficulty. It was determined that high response sensitivity required more effort than the medium or low level for the participants, as they tend to overshoot. It was determined that the medium sensitivity level provides the most comfortable sensitivity out of the three levels for the participants. Task difficulty was designed to increase every two minutes to provide sufficient time for the participants to realize the change in task difficulty.

Table 1. The Nine Levels of Task Conditions

Task difficulty level	Target update rate	Response sensitivity
1	Low	Medium
2	Low	Low
3	Low	High
4	Medium	Medium
5	Medium	Low
6	Medium	High
7	High	Medium
8	High	Low
9	High	High

Each update cycle of the tracking task is 100 ms (i.e., 10 Hz). Figure 5 shows all possible directions for the next movement of the target in the tracking task. The target always starts at the center position (5). At every update cycle, the current position of the target is evaluated and random numbers are generated to determine whether to stay at the current position or to move towards one of the other states.

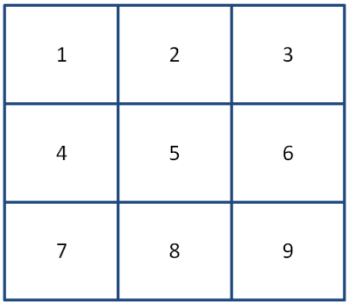

Figure 5. The Target States of the Tracking Task

Dependent Variables

There were three dependent variables: 1) time of PB that the participant verbally indicated, 2) Root mean square error (RMSE) of the tracking task (pixel

unit), and 3) errors in the secondary tasks (resource management task and system monitoring task).

During the experimental run, the participants were asked to subjectively identify the PB point, and that time was recorded.

In the tracking task, the target continuously deviated from the center point. The participants' goal was to keep the target at the center point. The target positions were sampled twenty times per second and the root mean square deviation (RMSD) values were recorded at every one-second interval. The following equation (4) was used to compute RMSD.

$$RMSD = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(0 - x_i)^2 + (0 - y_i)^2} \quad (4)$$

The system monitoring task required the operator to monitor and respond to simulated warning lights and gauges. The minimum response time was set for all stimuli in this task. If participants failed to respond within five seconds, each failure was counted as an error. The participants were required to respond by pressing the corresponding function key. Both response time (RT) and the number of errors were recorded. An equal number of stimuli (a total of sixteen stimuli) were presented at random points within every 2–minute period.

In the resource management task, fuel levels in two primary tanks (A & B) had to be maintained at a target level (2,500 units). Deviations from the target level were recorded every ten seconds. The sum of absolute deviation from the target level in both tanks A and B were computed for the analysis.

Hypotheses

First, the following hypothesis was examined to determine whether an increase in workload induces PB.

Hypothesis A: Increasing workload will induce subjectively-identified PB.

As mentioned earlier, the PB detection method is task-specific since what gets measures depends on the type of task. The method has been modified to detect PB on the collected tracking task performance data.

$$(RMSD > RMSD_{crit}) \cap (\Delta t \geq \varepsilon) \quad (5)$$

The equation above indicates that PB is identified when the deviation (RMSD) of the target for the tracking task exceeds the minimally acceptable performance level ($RMSD_{crit}$) for longer than a specified duration(ε). Time values of 5, 10, 15, 20, and 25 seconds were used as the values of each parameter $(RMSD_{crit}, \varepsilon)$. The following hypothesis was constructed to test whether there is a criterion for choosing the combination of the parameters that identifies the subjectively-identified PB point.

Hypothesis B: There exist criteria ($RMSD_{crit}$ and ε) such that the point in time corresponding to $(P < P_{crit}) \cap (\Delta t \geq \varepsilon$) matches the subjectively-identified performance breakdown point.

Next, the following hypothesis was constructed to identify whether the criterion that was found to detect the subjectively-identified PB point is consistent across participants.

Hypothesis C: The criterion from Hypothesis B is consistent across individuals.

Results

Overview

The following are the results of the hypothesis testing:

Hypothesis A: Increasing workload can induce subjectively-identified PB, although it might not be generalizable.

Hypothesis B: There were criteria that exhibited good performance in detecting the subjectively-identified PB point.

Hypothesis C: However, there were no criteria that were consistent among participants.

Hypothesis A

A total of 12 (10 male + 2 female) participants indicated that they experienced PB, which supports Hypothesis A (see Table 2). Table 2 also shows that there are large individual differences in how the participants performed the tracking task.

Table 2. Summary of the Tracking Task Performance

Participant	Mean	SD	Median	PB
1	26.8	18.4	22.5	Yes
2	18.4	11.7	15.6	Yes
3	29.1	19.3	24.6	No
4	22.7	12.5	20.7	Yes
5	25.8	22.1	20.6	Yes
6	33.5	21.8	28.5	Yes
7	28.3	18.8	23.8	Yes
8	19.5	10.5	17.4	No
9	19.7	11.7	17.4	No
10	40.4	28.2	33.8	Yes
11	22.3	13.4	18.9	Yes
12	22.9	15.7	18.9	Yes
13	23.3	15.0	20.2	Yes
14	24.9	15.3	21.8	Yes
15	26.4	16.5	23.4	Yes

An additional analysis was conducted on the tracking task performance. Figure 6 shows the histogram of the tracking task.

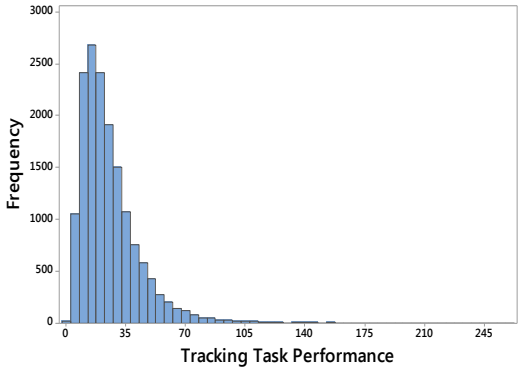

Figure 6. Histogram of the Tracking Task Performance

In Figure 6, it is observed that the distribution of the tracking task performance has a left skew with a long right tail.

Hypothesis B

The detection method has been applied to determine whether subjectively-identified PB can be sensitively detected. A ROC curve was constructed (See Figure 7) for each participant individually to investigate how various threshold values affect PB detection. In the figure, it is found that there is no combination of the threshold parameters that perfectly identified PB, but there is a threshold combination that performs better in terms of detecting the subjectively-identified PB point than the other combinations for each participant.

Additionally, in Figure 7, the relationships between different parameters were observed. It was observed that the duration of false detection is inversely related to the value of $RMSD_{crit}$. It was also identified that as the value of ε increases, the false detection rate decreases. It was also found that the duration of missed detection of PB increases as the values of $RMSD_{crit}$ and ε increase.

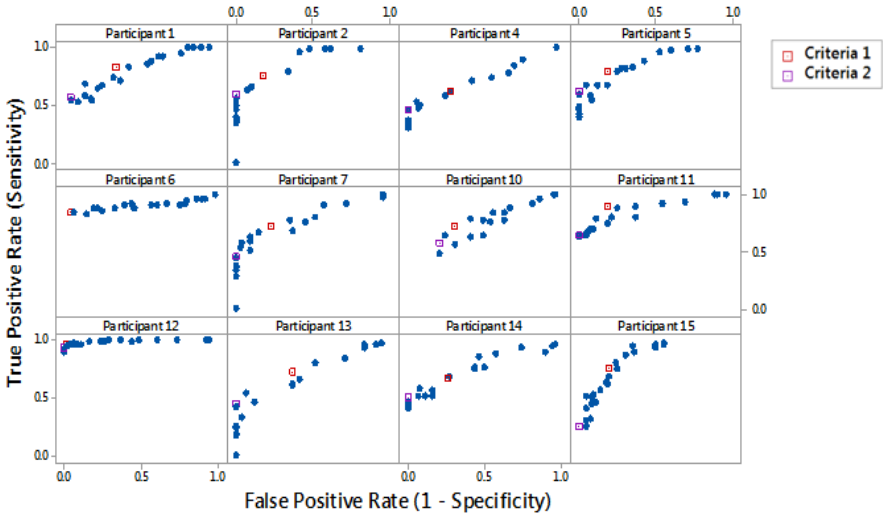

Figure 7. ROC Curves: Evaluation of the Parameters

Hypothesis C

The next analysis was conducted to further verify whether there was consistency in the criterion for detecting PB among the participants.

Table 3 indicates how the average duration of false detection, missed detection, delay time, FPR, and TPR changed due to use of the different threshold values. The values in Table 3 indicate that there was no unambiguous criterion for choosing the optimal threshold values that perform consistently among the participants.

The identified threshold values based on *Criteria 1* are presented in Table 4. As can be seen, there was no consistency among the participants in the threshold values that met *Criteria 1*.

Table 5 below contains the threshold values that were identified based on *Criteria* 2 for each participant. Again, it can be seen that there was no consistency in the threshold values among the participants. Also, there were some participants with threshold values that achieved no (= zero) FPR.

Table 3. The Average Effect of the Parameters

$RMSD_{crit}$	ε (sec.)	FPR	TPR	Delay (sec.)
5	5	0.90	0.99	0.0
5	10	0.87	0.99	2.0
5	15	0.84	0.99	2.5
5	20	0.81	0.98	3.3
5	25	0.79	0.98	4.2
10	5	0.67	0.99	14.6
10	10	0.56	0.89	7.5
10	15	0.48	0.85	10.7
10	20	0.42	0.81	19.2
10	25	0.36	0.77	37.2
15	5	0.42	0.81	3.4
15	10	0.27	0.73	15.3
15	15	0.19	0.67	79.8
15	20	0.19	0.62	65.7
15	25	0.11	0.58	192.5
20	5	0.26	0.70	10.3
20	10	0.14	0.62	86.4
20	15	0.10	0.51	172.1
20	20	0.06	0.49	105.8
20	25	0.05	0.46	187.2
25	5	0.14	0.62	21.8
25	10	0.08	0.52	112.7
25	15	0.04	0.45	201.7
25	20	0.03	0.44	282.9
25	25	0.02	0.41	140.7

Table 4. The Parameters Selected Based on *Criteria 1*

Partic-ipant	$RMSD_{crit}$	ε (sec.)	FPR	TPR	Duration of Missed Detection (sec.)
1	15	10	0.3	0.8	0.0
2	10	10	0.2	0.8	45.1
3	No report of PB				
4	10	25	0.3	0.6	74.0
5	15	10	0.1	0.8	5.0
6	25	25	≈ 0.0	0.9	87.0
7	15	10	0.2	0.7	0.0
8	No report of PB				
9	No report of PB				
10	25	10	0.3	0.7	0.0
11	10	25	0.2	0.9	0.0
12	20	15	≈ 0.0	1.0	8.0
13	10	15	0.4	0.7	3.0
14	20	5	0.3	0.7	0.0
15	10	15	0.2	0.8	0.0

Table 5. The Parameters Selected Based on *Criteria 2*

Partic-ipant	$RMSD_{crit}$	ε (sec.)	FPR	TPR	Duration of Missed Detection (sec.)
1	25	10	≈ 0.0	0.6	0.0
2	10	20	0.0	0.6	90.0
3	No report of PB				
4	20	15	0.1	0.5	26.0
5	15	20	0.0	0.6	163.1
6	25	25	≈ 0.0	0.9	87.0
7	20	15	0.0	0.5	265.1
8	No report of PB				
9	No report of PB				
10	20	20	0.2	0.6	0.0
11	20	15	0.1	0.7	290.1
12	25	15	≈ 0.0	0.9	9.9
13	20	10	0.0	0.5	251.6
14	20	15	0.0	0.5	201.9
15	15	25	≈ 0.0	0.5	8.9

Discussion

The study was conducted to empirically demonstrate PB and to evaluate the method developed to objectively detect such an extreme state. After running the study, it was determined that increasing workload can induce subjectively-identified PB. However, it was observed that it is not generalizable as only 12 out of 15 participants indicated that they experienced PB.

The PB detection method was applied to the performance data to identify how effectively it could detect PB. There were some indications that PB could be detected using the PB detection method, particularly when the parameters of the detection method were calibrated per individual, as there was no criterion that was consistent for all participants.

Although clear instructions were given to the participants that the goal was to keep the target at the center point, participants performed at different levels, which may have caused the lack of consistent criteria among participants. The variance in performance may be due to natural causes and may have contributed to these differences. In order for the PB detection method to work effectively, the participants must show good tracking task performance when they control the task. However, some of the participants did not show or maintain this performance throughout the whole study. Future research could establish criteria for determining which participants are good candidates for applying the PB detection method. One possible approach for determining qualified participants is by setting a minimum required performance level and applying the PB detection method only with the participants who can maintain their performance within the minimum required level as long as they possess control of the task.

Conclusion

In the past, PB has been only anecdotally described as a state where the operator "loses control of the context" and "cannot maintain task performance." The past works on PB description do not have specific definitions. In addition, PB has not been empirically demonstrated. There is no validated objective way of detecting PB or the transition into such state. An objective way of detecting PB

transition is needed for a system to determine when to intervene and assist human operators to prevent PB from occurring.

In this work, a definition of PB is given. PB was successfully induced in a controlled setting and the characteristics of PB were reported. The criteria from the PB definition detected PB and it was shown that increasing workload can induce subjectively-identified PB, although this might not be generalizable. This suggests that the parameters of the PB detection method may have to be calibrated per individual. Future work could evaluate whether such calibrated parameters could be re-used over time.

The parameters of the PB detection method were calibrated to match the subjectively declared PB point. Currently, the only available way of identifying PB is through subjective identification. There are, however, ambiguity issues with such subjectively declared PB points. Hence, other indicators of PB using other measures should be investigated. The redundancy that could potentially be provided by multiple indicators could help improve the reliability of PB detection.

Also, in order to prevent operators from experiencing PB, an effort should be made to look for reliable precursors to PB. Such precursors can be used to preemptively prevent PB from occurring.

References

[1] Wickens, Christopher D 2008, "Multiple resources and mental workload." *Human Factors: The Journal of the Human Factors and Ergonomics Society* 50, no. 3: 449-455.

[2] Lagu, Amit V., Steven J. Landry, and Hyo-Sang Yoo 2013, Adaptive function allocation stabilization and a comparison of trigger types and adaptation strategies. *International Journal of Industrial Ergonomics* 43, no. 5, 439-449.

[3] Yoo, Hyo-Sang 2012, Framework for designing Adaptive Automation. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, SAGE Publications, vol. 56, no. 1, pp. 2133-2136.

[4] Bradley, Andrew P. 1997, The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern recognition* 30, no. 7, 1145-1159.

[5] Kuchar, James K. 1996, Methodology for alerting-system performance evaluation. *Journal of Guidance, Control, and Dynamics* 19, no. 2, 438-444.

[6] Parasuraman, Raja, Thomas B. Sheridan, and Christopher D. Wickens. 2000, A model for types and levels of human interaction with automation. *Systems, Man and Cybernetics, Part A: Systems and Humans, IEEE Transactions on* 30, no. 3, 286-297.

[7] Swets, John A. 2014, *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Psychology Press.

[8] Comstock, James R., and Ruth J. Arnegard 1992. *The multi-attribute task battery for human operator workload and strategic behavior research*. Hampton, VA: National Aeronautics and Space Administration, Langley Research Center.

Acknowledgements

The authors would like to thank the lab personnel (particularly, Nancy Smith) at Airspace Operations Laboratory (AOL) at NASA Ames Research center.