# IDENTIFICATION OF STABLE ASYMPTOTIC PERFORMANCE ON COMPUTER-BASED COGNITIVE TESTS

Byeong-cheol Hwang, Robert E. Schlegel, and Randa L. Shehab
School of Industrial Engineering
University of Oklahoma
Norman, Oklahoma

Examining whether human cognitive performance is affected by environmental conditions requires stable performance measures prior to stressor exposure. This study evaluated the stability and reliability of six computer-based cognitive performance tasks. A Microsoft Excel Visual Basic for Applications (VBA) macro program, the Stability and Reliability Analysis System (SRAS), was developed to evaluate performance of the cognitive tests using three approaches for identifying stability: Graphical Analysis, Learning Curve Fitting, and Statistical Analysis. The results of the comparative evaluation indicated that the SRAS macro program was effective in determining differential stability for the various tasks and measures. Across all tasks, the use of a compound graphical analysis approach was better than a single graph method in terms of providing a more reliable estimation of task stability. Learning curves were fit to each performance measure. For most tasks, the best-fit models were power and logarithmic models. The statistical analysis methods provided conservative estimates of task stability.

## INTRODUCTION

Human cognitive capabilities are affected by a variety of stressors and risk factors. The utility of laboratory-based human cognitive performance assessment in predicting human performance under various conditions has not been fully explored (Turnage and Kennedy, 1992). Prior to the development of computer-based cognitive performance test batteries, human performance was measured using self-reports or paper-and-pencil tests. However, the methods used were limited because subjects could provide false answers on the self-reports and the data from the paper-and-pencil tests were not reliable (Turnage and Kennedy, 1992). However, due to advances in computer technology, complex human cognitive performance can now be assessed using computer-based task batteries.

Schlegel and Gilliland (1992) discussed the need for theories of human cognition and performance in order to define both theoretical and practical limits of mental work capacity. Computer-based test batteries represent one means by which the limits of human work capacity can be identified. The Performance Assessment Workstation (PAWS), developed for the United States National Aeronautics and Space Administration (NASA) Second International Microgravity Laboratory (IML-2) space shuttle mission, is one computer-based task battery designed to investigate the effects of the space environment (Schlegel, Shehab, Gilliland, Eddy, and Schiflett, 1995).

During a series of space shuttle launches, NASA funded researchers to study the effects of microgravity on cognitive performance. As part of the IML-2 payload, the NASA–PAWS experiment studied the effects of microgravity on astronaut cognitive performance. A ground-based study was conducted prior to the mission to provide comparative data important for the detection of microgravity effects (Schlegel et al., 1995). In addition, the ground-based study examined the stability and reliability of the assessment measures.

The study presented here evaluated the differential stability of computer-based cognitive performance measures collected as part of the NASA-PAWS experiment. Differential stability is the consistency of individual differences when a repeated-measures task is performed (Bittner, 1979; Jones, Kennedy, and Bittner, 1981; Kim and Miller, 1995). Graphical methods of examining differential stability include mean and standard deviation plots, superdiagonal correlation plots, and sessions after base session (SABS) correlation plots. Learning curve models can also be developed to graphically describe data patterns. In addition, statistical evaluation of stability can be performed with both ANOVA applied to inter-session correlations and Lawley tests.

Mean and standard deviation plots can be used to visually identify the session in which performance becomes stable. According to Jones et al. (1981) and Kim and Miller (1995), one indicator of graphical stability for a cognitive performance measure is that the mean becomes asymptotic and the standard deviation among subjects remains relatively constant from session to session. Superdiagonal correlation plots indicate an *unstable* variable when the correlations between successive sessions show "superdiagonal form," that is, the slope of the successive correlation plot is significantly positive or negative. Graphical analysis techniques also include SABS correlation plots. These plots are constructed by selecting a base session of interest and plotting the correlations of all succeeding sessions with that session. To indicate stability from a specific session onward, the slope of the SABS plot for that session must be flat and should overlay plots for succeeding sessions.

Learning curves are also important tools to graphically investigate changes in performance. As an individual becomes more practiced, learning curves become asymptotic. The point where the curve becomes asymptotic represents the point where stability is attained. Many learning curves can be described by mathematical models (Damos, 1991) such as linear, logarithmic, parabolic, power and exponential curves.

Jones (1979) defined stability as the point when practice effects no longer appear. This definition forms the statistical basis for the Early vs. Late Session Correlation ANOVA Test and the Lawley Test of Correlation Equality (Bittner, 1979), both of which examine the stability of the data.

Previous work by Jones-Parra (1993) resulted in a macro-based analysis tool to evaluate differential stability for computer-based human cognitive performance measures. The macro examined several techniques and concluded that the use of correlation plots combined with Early vs. Late correlation ANOVA was effective in determining whether a task had reached stability.

The primary objective of the current work was to provide a more generalized analysis tool for identifying stable asymptotic performance on computer-based human cognitive assessment tasks. In order to achieve this objective, several analytical methods for identifying stability in human cognitive performance were implemented in a computer-based evaluation tool. These methods were used to identify the particular session within a series of sessions where differentially stable, asymptotic performance is attained on measures from human cognitive performance tests. The database from the NASA ground-based study mentioned above was used to evaluate the software tool.

## METHOD

A macro program, the Stability and Reliability Analysis System (SRAS), was developed in Microsoft Excel to implement various analysis methods for the determination of data stability. Three types of analyses were implemented in the SRAS: Graphical Analysis, Learning Curve Fitting, and Statistical Analysis. Three graphical analysis techniques, the Mean and Standard Deviation Plot, the SABS Correlation Plot, and the Superdiagonal Correlation Plot, are used to investigate overall performance patterns and differential stability of the task measures. Five different learning curve models are fit to the overall means. Two statistical analysis methods, the Early vs. Late Correlation ANOVA and the Lawley Test of Correlation Equality, are also applied to the data. The SRAS program generates an output file that includes graphs, the correlation coefficient matrix, five learning curve models, ANOVA tables, and the Lawley statistic.

The SRAS macro program was used to analyze the NASA-PAWS ground-based data with respect to identifying the session corresponding to stable performance. SRAS analyses were conducted on performance measures from each of six tasks. To determine a unified estimate of task stability from the alternative analyses, a task stability criterion was developed. Since the NASA-PAWS data were collected to identify and verify performance decrements caused by microgravity conditions in space, task stability was determined using the most conservative conclusion from among the methods.

The NASA-PAWS ground-based data were collected from 96 subjects (34 females and 62 males) in studies conducted at Brooks Air Force Base and the University of Oklahoma. Subject age ranged from 17 to 52 years, with a mean age of 25.8 years and a standard deviation of 8.1 years. Six performance tasks from the NASA-PAWS experiment were examined in this study, including Critical Tracking (TRK), Spatial Matrix (MTX), Sternberg Memory Search (STN), Continuous Recognition (CRC), Switching (MAN and MTH) and Dual Task (DUL). There were 28 total sessions: 8 training sessions, 15 practice sessions, and 5 simulated mission sessions.

As an example, Figure 1 illustrates the mean and standard deviation plot for the Tracking (TRK) task Maximum Lambda (LM) measure. This plot presents the group mean and standard deviation across subjects as a function of session number. This method determines an asymptotic point by using visual inspection of the plot created by the SRAS macro program. From visual inspection of the mean and standard deviation plot, the TRK task acquired stability at Session 22.

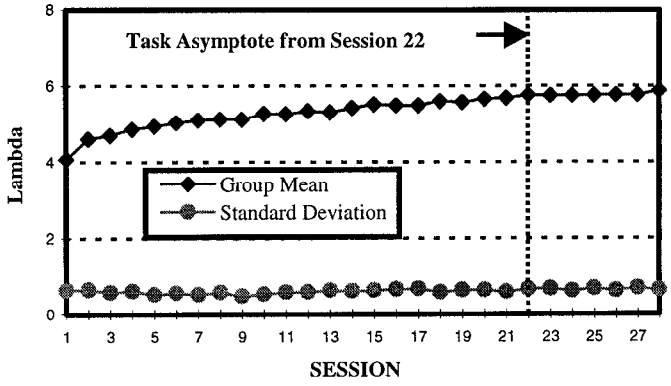


**Figure 1**. Mean and Standard Deviation Plot for Tracking Task Maximum Lambda.

As shown in Figure 2, the SABS correlation plot for TRK LM is constructed by selecting the row of the session-by-session correlation matrix corresponding to the base session of interest and plotting the correlations on that row to the right of the base session. From the SABS correlation plot, differential stability can be confirmed when the slopes of the correlation plots are approximately zero and plots using succeeding base sessions overlay one another. Figure 2 suggests that task stability occurred starting at base session 15.

Another method to assess the stability of a performance task is the superdiagonal correlation plot. When a variable has not stabilized, the slope of correlations between successive sessions are significantly positive or negative. Figure 3 illustrates a superdiagonal correlation plot of the data for the TRK LM measure. As shown in Figure 3, the task stability for the TRK LM measure can be confirmed from Session 7.

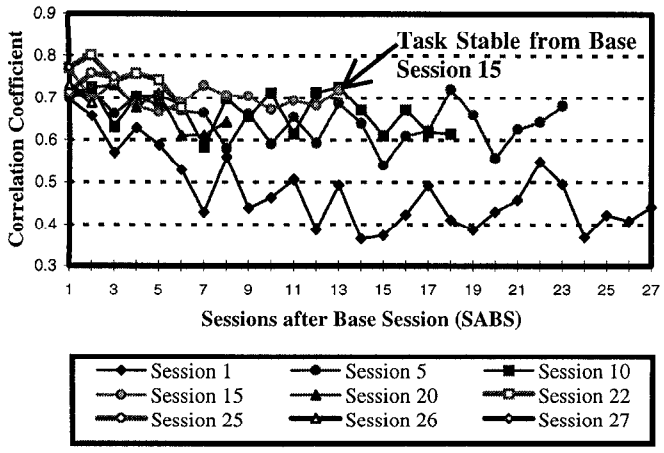


**Figure 2.** SABS Correlation Plot for Tracking Task Maximum Lambda.

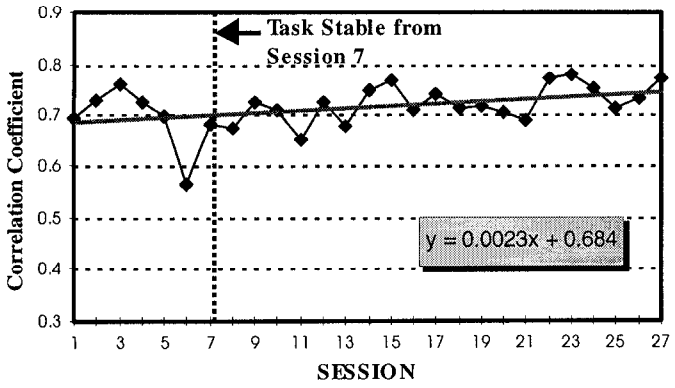


**Figure 3.** Superdiagonal Correlation Plot for Tracking Task Maximum Lambda.

To assist in modeling human performance learning, five different models of learning were fit to the data and the R-squared ($R^2$) values for each model were compared to identify the best-fit equation. Figure 4 shows a Power learning curve for the TRK LM measure as an example of the best-fit learning curve. This learning curve plot provides the same means plot to visually identify the point of stability, and also provides a numerical value for the performance asymptote.

Two statistical methods, ANOVA and Lawley's test, were used to provide objective statistical analysis tools for determining the onset and extent of stability. Table 1 represents the SRAS output for the ANOVA test of the TRK LM measure. If stabilization occurs, a correlation matrix can be divided into pre-stabilization and post-stabilization partitions such that the correlations between any of the later sessions and one of the earlier sessions is constant and the correlation between any two later sessions is the same. Orthogonal contrasts are used to determine the linear component of the column factor. If this component is non-significant, there is no significant difference among the correlations of the later sessions with a specific early session, and it can be concluded that the task is stable from the first session of the "later" sessions group. As seen in Table 1, Session 25 was examined as the first session of the post-stabilization partition. Based on this segmentation, the linear contrast was not significant, thus indicating that the measure was stable as of Session 25.

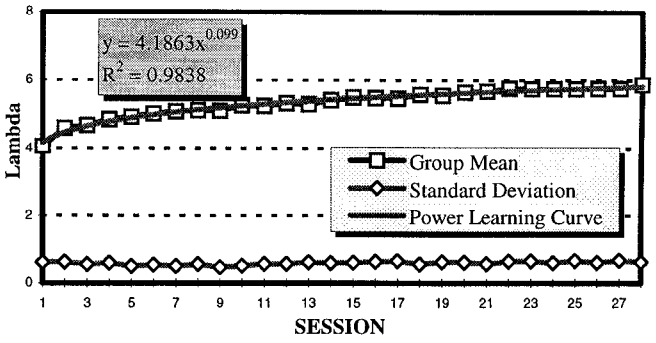


**Figure 4.** Power Learning Curve for Tracking Task Maximum Lambda.

**Table 1.** ANOVA Results Table for Tracking Task Maximum Lambda.

**ANOVA TEST**

Session # 25
Significance Level = 0.05

| Source | SS | df | MS | F | P-value | F crit | |
|---|---|---|---|---|---|---|---|
| Rows | 0.531 | 23 | 0.023 | 15.249 | 0.000 | 1.687 | Significant |
| Columns | 0.005 | 3 | 0.002 | 1.075 | 0.365 | 2.737 | Not Significant |
| Linear | 0.001 | 1 | 0.001 | 0.491 | 0.486 | 3.980 | Not Significant |
| Residual | 0.004 | 2 | 0.002 | 1.367 | 0.262 | 3.130 | Not Significant |
| Error | 0.105 | 69 | 0.002 | | | | |
| Total | 0.641 | 95 | | | | | |

Table 2 shows an example of the Lawley test implemented with the SRAS program. Lawley's test examines the equality of all correlations among a proposed group of "later" sessions. This test is an approximation of a likelihood-ratio test (Morrison, 1990). If the result is non-significant, indicating that all correlations among the "later" sessions are similar, the measure is deemed stable from the first session of the "later" group.

**Table 2.** Lawley Test Results Table for Tracking Task Maximum Lambda.

**LAWLEY'S TEST**

Session # 22
Significance Level = 0.05
Degrees of Freedom = 20
Chi-squared = 21.35038
P = 0.376774 Not Significant

## RESULTS

The results for all task measures are summarized in Table 3.

### Critical Tracking

The Critical Tracking task requires that the subject maintain an unstable target in the center of a horizontal line by manipulating a control device to nullify the input disturbance. TRK performance was assessed in terms of the maximum lambda during a session (LM), the mean of the lambdas associated with each control loss during the session (ML), the number of control losses (CL), and the Root Mean Square error (RM).

Visual inspection using the graphical analysis methods confirmed a stable asymptotic level for all measures (LM, ML, CL and RM) beginning with Session 22. The standard deviation plots for LM, ML and RM suggested that the measures were stable across all sessions. On the other hand, the CL measure had high variability during the training sessions. Based on the SABS and superdiagonal correlation plots, stability for all measures occurred by the end of the practice sessions (Sessions 22 and 23).

The Lawley test agreed with the graphical analyses better than did the ANOVA test. The Lawley and ANOVA tests confirmed that TRK measures stabilized at the end of practice (Session 23) or during the mission sessions (Session 25), respectively. Both the power and logarithmic models provided a good fit for the TRK task with high $R^2$ values for most measures. In summary, the TRK task reached an asymptotic level by the end of the practice sessions.

### Spatial Matrix

The Matrix task involves paired comparisons of five-by-five matrices with five illuminated cells. A subject must decide if a second matrix is identical to the first except for a 90° rotation. MTX performance was assessed in terms of the mean reaction time for correct responses (RT), the percentage of correct responses (PC), and throughput (TP: combines both RT and PC). The RT and TP measures for the MTX task never achieved differential stability, while the PC measure was stable from Session 3 based on the statistical analyses.

### Sternberg Memory Search

The Sternberg task requires that subjects respond as rapidly (RT) and accurately (PC) as possible to visually presented letters. The SRAS statistical analysis indicated that RT and TP measures for the STN task were stable from Session 26 and Session 22, respectively. However, task stability for the PC measure was achieved by Session 3 due to a ceiling effect typical for this measure of this task.

### Continuous Recognition

The Continuous Recognition task was scored using the same measures as MTX and STN and the results were similar to the STN results. The RT and TP measures for the CRC task were stable from Session 26 and Session 24, respectively. However, task stability for the PC measure was achieved at Session 12 due to the ceiling effect for the measure.

### Switching Task

For the Switching task, RT and PC were assessed separately for the Manikin (MAN) and Mathematical Processing (MTH) portions of the task. The RT measures of both subtasks produced similar results (stable at Sessions 26 and 24, respectively). Also, the PC measures of both subtasks produced similar results (stable at Sessions 20 and 18, respectively).

### Dual Task

The Dual Task combines the TRK task (with CL and RM measures) with the STN task (with RT, PC and TP measures). Due to a change in task difficulty during the practice sessions, only the first eight training sessions were analyzed for stability in this study. With the exception of the PC measure, all measures were stable from Session 6. The PC measure was stable from Session 3 based on the statistical analysis.

## DISCUSSION

To date, most research has focused on visual examination of group means and standard deviations to determine a performance asymptote and task stability. The mean and standard deviation plot provides the experimenter with a graphical view of the overall group behavior of subjects across sessions. However, it is not possible to determine differential stability by examining plots of group means.

The SABS correlation plot provides visual information about task stability, but it is often difficult to interpret (refer to Figure 2). The superdiagonal correlation plot allowed determination of task stability in most cases. However, the conclusion from this method can only be regarded as a possible starting session of task stability. The superdiagonal correlation plot typically identified stability during the early sessions, while the SABS correlation plot identified stability at later sessions. Therefore, the SABS correlation plot provided more reliable results.

For most NASA-PAWS tasks, power and logarithmic learning curve models provided the best fit. The parabolic model was effective for the Dual Task. However, long-term prediction of performance was poor with the parabolic model.

This study revealed that the statistical approaches were the most conservative among the analysis methods used. In most cases, the statistical analyses identified differential stability and stable correlations during later sessions (Sessions 24 through 28). In many cases, the early vs. late correlation ANOVA provided objective numerical information. On the other hand, this method showed inconsistent results as a stand-alone analysis tool. This phenomenon existed in the analysis of later sessions due to the smaller number of data points included in the analysis. In such cases, simultaneous use of the SABS correlation plot with the ANOVA results may help the researcher determine if stability has been achieved. Although this study revealed that the graphical analyses and the Lawley test were typically in agreement regarding points of stability, the ANOVA test often produced different results. However, with Percent Correct (PC) measures, the ANOVA test was more effective in confirming the onset of differential stability.

This study demonstrated that although the mean and standard deviation plot may show a stable learning trendline, the use of alternative analysis techniques may yield discrepant

results. Therefore, this study recommends that a compound analysis approach utilizing all analysis techniques available in the SRAS, be considered for more accurate interpretation of data.

## REFERENCES


Bittner, A.C., Jr. (1979). Statistical tests for differential stability. *Proceedings of the 23rd Annual Meeting of the Human Factors Society* (pp. 541-545). Santa Monica, CA: Human Factors Society.

Damos, D.L. (1991). Examining transfer of training using curve fitting: A second look. *The International Journal of Aviation Psychology, 7*, 73-85.

Jones, M.B. (1979). Stabilization and task definition in a performance test battery. *Proceedings of the 23rd Annual Meeting of the Human Factors Society* (pp. 536-540). Santa Monica, CA: Human Factors Society.

Jones, M.B., Kennedy, R., and Bittner, A.C., Jr. (1981). A video game for performance testing. *American Journal of Psychology, 94*, 143-152.

Jones-Parra, L.C. (1993). *Identification and statistical verification of performance stability for cognitive tasks.* Unpublished Project Report. Norman, OK: School of Industrial Engineering, University of Oklahoma.

Kim, H.T., and Miller, J.C. (1995). Statistical stability of the ReadyShift fitness-for-duty driving test. In A.C. Bittner, Jr., and P.C. Champney (Eds.), *Advances in Industrial Ergonomics and Safety VII* (pp. 501-508). London, UK: Taylor and Francis.

Lawley, D.N. (1963). On testing a set of correlation coefficients for equality. *Annals of Mathematical Statistics, 34*, 149-151.

Morrison, D.F. (1990). *Multivariate statistical methods* (3rd Ed.). New York: McGraw-Hill.

Schlegel, R.E., and Gilliland, K. (1992). *Development of the UTC-PAB normative database* (AL-TR-92-0145). Wright-Patterson AFB, OH: USAF Armstrong Laboratory.

Schlegel, R.E., Shehab, R.L., Gilliland, K., Eddy, D.R., and Schiflett, S.G. (1995). *Microgravity effects on cognitive performance measures: Practice schedules to acquire and maintain performance stability* (AL/CF-TR-1994-0040). Brooks AFB, TX: USAF Armstrong Laboratory, Crew Technology Division.

Turnage, J.J., and Kennedy, R.S. (1992). The development and use of a computerized human performance test battery for repeated-measures applications. *Human Performance, 5*, 265-301.


**Table 3**. Summary Table of Stability As Identified from SRAS Analyses.

| Task | Measure | SRAS Stability Analysis Methods | | | Conclusion |
|---|---|---|---|---|---|
| | | Graphical | Statistical[1] | Best-Fit Learning Model | |
| Critical Tracking (TRK) | LM | Session 22 | Session 25 | Power | Session 25 |
| | ML | Session 22 | Session 25 | Logarithmic | Session 25 |
| | CL | Session 18 | Session 26 | Power | Session 26 |
| | RM | Session 23 | Session 24 | Logarithmic | Session 24 |
| Spatial Matrix (MTX) | RT | Session 9 | Not Confirmed | Power | Not Stable |
| | PC | Session 13 | Session 3 | Power | Session 3 |
| | TP | Session 9 | Not Confirmed | Parabolic | Not Stable |
| Sternberg Memory Search (STN) | RT | Session 9 | Session 26 | Power | Session 26 |
| | PC | Not Confirmed | Session 3 | Parabolic | Session 3 |
| | TP | Session 9 | Session 22 | Power | Session 22 |
| Continuous Recognition (CRC) | RT | Session 9 | Session 26 | Logarithmic | Session 26 |
| | PC | Session 18 | Session 12 | Logarithmic | Session 12 |
| | TP | Session 3 | Session 25 | Power | Session 25 |
| Switching Task (MAN and MTH) | MAN RT | Session 5 | Session 26 | Logarithmic | Session 26 |
| | MAN PC | Not Confirmed | Session 20 | Logarithmic | Session 20 |
| | MTH RT | Session 4 | Session 24 | Logarithmic | Session 24 |
| | MTH PC | Not Confirmed | Session 18 | Logarithmic | Session 18 |
| Dual Task (DUL)[2] | CL | Session 4 | Session 6 | Parabolic | Session 6 |
| | RM | Session 1 | Session 6 | Logarithmic | Session 6 |
| | RT | Session 5 | Session 6 | Parabolic | Session 6 |
| | PC | Not Confirmed | Session 3 | Parabolic | Session 3 |
| | TP | Session 2 | Session 6 | Power | Session 6 |

**NOTE:** 1. Determination of stability using statistical methods was based on the later session reported by the two techniques. However, for the PC measures, the earlier session was selected due to ceiling effects.
2. Due to a change in task difficulty after Session 9, only the first eight training sessions were analyzed for stability in this study.