DEVELOPMENT AND VALIDATION OF THE CONTROLLER ACCEPTANCE RATING SCALE (CARS): RESULTS OF EMPIRICAL RESEARCH

Katharine K. Lee, *NASA Ames Research Center*
Karol Kerns, Ph. D, *The MITRE Corporation Center for Advanced Aviation System Development*
Randall Bone, *The MITRE Corporation Center for Advanced Aviation System Development*
Monicarol Nickelson, *Federal Aviation Administration*

Abstract

The measurement of operational acceptability is important for the development, implementation, and evolution of air traffic management decision support tools (DSTs). The Controller Acceptance Rating Scale (CARS) was created at NASA Ames Research Center for the development and evaluation of the Passive Final Approach Spacing Tool. CARS was modeled after a well-known pilot evaluation rating instrument, the Cooper-Harper Scale, and has since been used in the evaluation of the User Request Evaluation Tool, developed by MITRE's Center for Advanced Aviation System Development. This paper provides a discussion of the development of CARS and an analysis of the empirical data collected with CARS to examine construct validity. Evaluations of both DSTs showed that interrater reliability of CARS scores was good to excellent in terms of controller consistency and agreement. Subjective workload data collected in conjunction with the CARS show that the expected set of workload attributes was correlated with the CARS. The analysis also demonstrates that CARS ratings were sensitive to the impact of DSTs on controller operations. Recommendations for future CARS development and its improvement are also provided.

1.0 Introduction

The Federal Aviation Administration's (FAA's) Free Flight Phase 1 (FFP1) Program is currently deploying the core capabilities of several decision support tools (DSTs) at a number of operational air traffic control facilities. As the FFP1 DSTs proceed toward deployment, and as future tools are developed under Free Flight Phase 2, technical guidance on human factors methods and measures is needed to support the evolutionary system development process envisioned by the RTCA [1]. Critical to the success of this evolutionary development process is the definition and application of human factors criteria that are sensitive, accurate, practically relevant, and economical to collect in an operational setting [2].

The work presented in this paper was undertaken to advance the definition and measurement of operational acceptability, an important indicator of satisfactory human-system performance. Operational acceptability, as an air traffic management (ATM) measurement construct, represents the effectiveness and suitability of the total system, including human and automation performance, in the operational environment. There are a number of assumptions underlying the construct of acceptability, including the experienced workload, the effectiveness of the functionality embodied in the equipment, and the suitability for human use in performing tasks in the specified environment. Effectiveness and suitability are generally considered necessary but not sufficient conditions for operational acceptability. System acceptability can be affected if users do not have sufficient understanding of a system, or do not use it according to the designers' intentions [3]. Acceptance is also influenced by less-easily-measured constructs such as trust in the automated system [3], impact on job satisfaction, the comfort level of the operator performing the prescribed duties, and the amount of required training. Thus, operational acceptability is largely a human–centered construct since many variables may combine to create a perception of acceptability within the user.

In considering the validation of a measure of operational acceptability, it is also important to recognize that acceptability should correlate with the extent to which a DST will actually be used. Previous research into the factors that influence automation use found that automation reliability, the operator's trust in automation, and the experienced workload, influenced use [4]. Because the acceptability judgments of different individuals reflect both objective task demands and the operator's response to the task, there may be large individual differences in the judged acceptability of similar operating environments. Research on automation use in ATC environments has further shown that workload extremes in either direction (i.e., overload or under-load) are undesirable and may limit acceptance and use of DSTs [5].

To evaluate acceptability of a new DST, it is critical to assess how the DST influences workload. Workload itself is a concept that has been difficult to measure and validate [6]. One of the complexities in validating a workload measure has been establishing an appropriate criterion variable, i.e., the amount of information-processing resources used during task

performance. Subjective rating scales have emerged as the primary procedure for measuring workload. Although multivariate descriptions of workload are widely used to identify sources of workload, a single number representing the overall demand on information processing resources is also useful to segregate situations that are likely to pose workload problems from those that are not. Similarly, there is a need for a single number measuring whether the workload incurred by the human operator to achieve desired levels of safety and system performance is operationally acceptable and will result in DST use [7].

NASA Ames Research Center and MITRE's Center for Advanced Aviation System Development (CAASD) have been developing DSTs for the terminal area, en route, and traffic management environments. These tool development efforts have necessitated the creation of measures to assess the progress of system development and capture ratings of controller acceptance. This paper describes the validation of a measure of controller acceptance as applied to two DSTs, the Passive Final Approach Spacing Tool and the User Request Evaluation Tool.

1.1 Passive Final Approach Spacing Tool

NASA Ames Research Center has been developing the Center-TRACON Automation System (CTAS), composed of several DSTs that form a suite of automation tools for the controller and the traffic management coordinator. One of the CTAS tools that completed operational evaluation is the Passive Final Approach Spacing Tool (pFAST). Passive FAST is designed to provide advisory information to terminal-area radar controllers for efficient runway balancing and sequencing of arrival traffic. Researchers at NASA Ames conducted several years of controller-in-the-loop simulations for refining the algorithms which drive pFAST. The testing culminated in a six-month field evaluation at Dallas/Ft. Worth (DFW) TRACON in 1996. The engineering data from the pFAST field evaluation showed an increase in throughput of 9-13% when pFAST advisories were used by arrival controllers [8]. Human factors data collected from the pFAST field evaluation are analyzed and presented here to contribute to the validation of the CARS. As reported in Ref. 9, human factors data analyses showed that there was no significant increase in user workload from the addition of pFAST advisories despite the increase in arrival throughput. In addition, the system was deemed acceptable by the controllers. More information regarding pFAST development can be found in Ref. 8 and 9.

1.2 User Request Evaluation Tool

Based on years of collaborative laboratory research to develop en route automation tools, the FAA and MITRE/CAASD have been conducting operational trials of an initial DST for the sector team, called the User Request Evaluation Tool (URET). A URET prototype continues to operate in daily use at Indianapolis and Memphis Air Route Traffic Control Centers (ARTCCs) and is part of the FFP1 deployment. URET has been adapted for primary use by the Radar Associate (or D-controller) position and is designed to provide advisory information for strategic conflict detection and clearance planning. URET also includes interactive trial planning and visualization capabilities that allow the controller to determine whether a trial flight plan modification will create other conflicts. More information regarding the development and evolution of URET capabilities can be found in Ref. 10 and 11.

In the following sections we will: (1) describe the development and use of CARS in measuring controller acceptance during the evaluation of pFAST and URET, (2) present results of analyses aimed at measuring CARS reliability as well as the relationship between CARS and measures of workload and the use of a DST, (3) discuss the results, and the degree to which they support using CARS as a tool for DST evaluation, and (4) recommend further CARS development and validation efforts.

2.0 Development of the CARS Format

This section describes how the CARS was developed, following the model of a previously well-established measure, the Cooper-Harper Scale (CHS). Examples of how CARS was tailored for pFAST and for URET are given, as well as descriptions of the procedures used in administering CARS for the pFAST and URET evaluations described in this paper.

2.1 The Cooper-Harper Scale

The CHS (see figure 1) was developed at NASA Ames Research Center in the 1960's to assess the handling qualities of test aircraft [12]. It has been described as the international standard for pilot evaluations [13, 14, 15]. Pilot evaluation, considered essential to assessments of aircraft handling quality [13], provides the ability to investigate both pilot-vehicle performance and total workload required to achieve an aircraft's intended use.

In the flight testing domain, the CHS was used to capture the fact that handling qualities reflect both the pilot and the aircraft working together. The developers of the CHS and the researchers using the

Cooper-Harper Scale: reproduced from Harper & Cooper, 1986

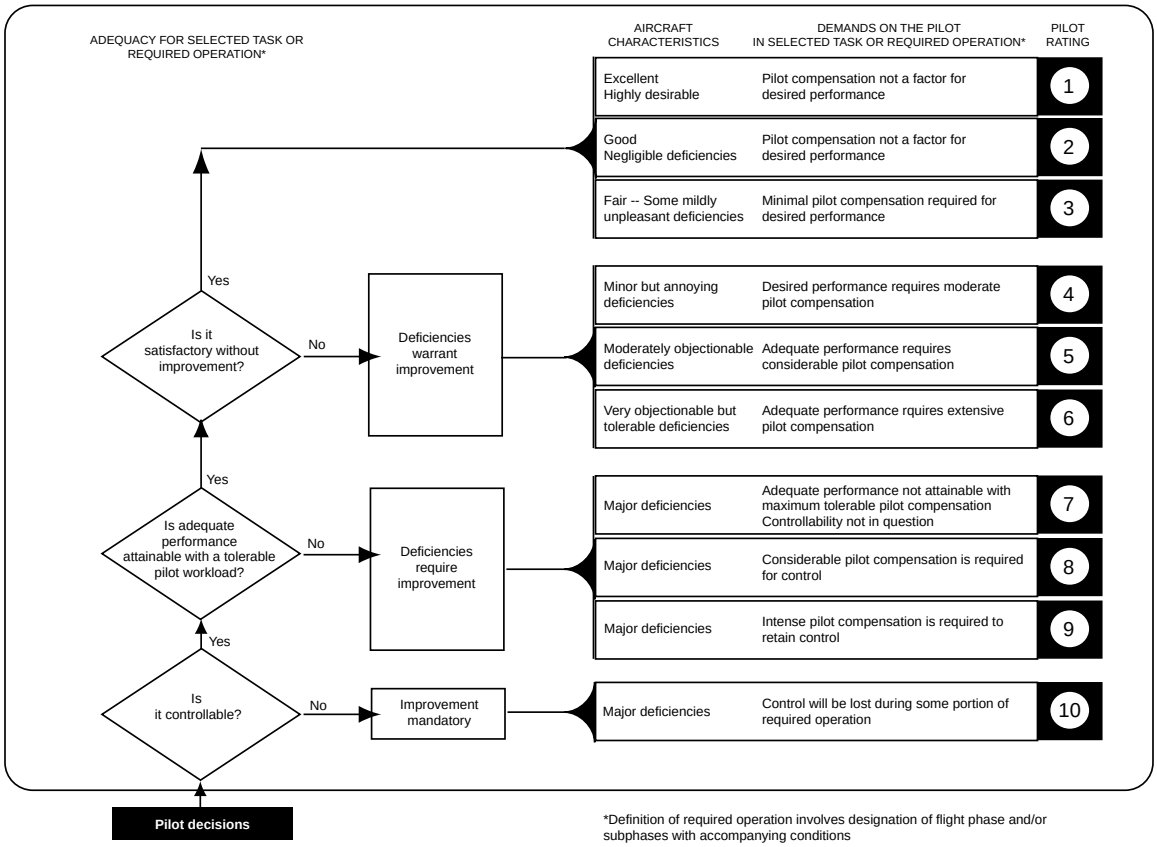

Figure 1. The Cooper-Harper Scale

CHS recognized that how the engineer might view the performance of the aircraft is quite different from how the pilot views it. It was therefore an important objective to achieve a standard set of terminology and definitions for both the pilot and the engineer [13].

The CHS follows a decision-tree structure to help pilots arrive at a rating that best describes the handling qualities of the test aircraft. When the scale is used properly, the raters consider four different rating categories in order of impact on handling qualities: controllability, performance with tolerable workload, aircraft characteristics, and effect on the pilot [14]. By forcing the pilot raters to strictly adhere to the decision-tree structure of the CHS, researchers can also reduce the variability of the pilot ratings [14]. It is also critical to carefully define the meaning of the words used in the scale in order to achieve reliable and meaningful ratings [13,15]. Research evaluating the use of the CHS have also noted the importance of descriptive comments when providing ratings [16].

Since its development in the 1960's, researchers have made modifications to the original CHS to tailor it to system evaluation beyond the pilot-handling qualities domain; for example, the wording in the CHS has been changed to reduce the emphasis on motor skills in the rating process [15]. Experiments with the resulting Modified Cooper-Harper Scale (MCH) showed that the MCH was a statistically reliable indicator of overall mental workload [15].

2.2 CHS Modifications for CARS

Because of successes in a pilot evaluation setting, and its straightforward application and structure, the CHS was chosen as a model for measuring pFAST system acceptance [17]. As described earlier, the operational acceptance of a DST is dependent upon more than just the DST's engineered performance. Further, while controller comments and observations can help to indicate the acceptability of a tool, some means of quantifying this data are also required to demonstrate the consistency of the acceptability criteria. As a result, there was a requirement for developing a measure of acceptance that could be tracked over a period of time, as development progressed. In the

simulation environment, a research goal is to determine when a DST would be ready for operational evaluation. Once in the operational evaluation phase, a research goal is to determine when a DST's performance can be deemed acceptable for daily-use operations [18]. The CHS was modified to help researchers meet these objectives.

A number of cosmetic changes were made to the structure of the CHS in the creation of CARS. The physical direction of the ratings was changed from that of the CHS so that in the CARS, "1" was unacceptable, and "10" was completely acceptable. This direction change was done to make a lower number represent a less desirable rating, and a higher number to represent a more desirable rating. The layout of scale used by the raters was constructed to move from top to bottom, rather than from bottom to top.

The original CHS wording was changed throughout to emphasize the controller's evaluation of an advisory system. While retaining the key categories of controllability, tolerability, satisfaction, and desirability (acceptability), wording changes were also made to describe the "no" response to the major rating categories, with an emphasis on workload in the description of tolerability.

Beginning at the top of the scale, at the "START" label, the rater answers a series of yes-no questions about the system performance of the scenario being evaluated. The response to the yes-no questions leads the rater to the eventual numeric rating that best represented the system performance. After making the numeric rating, the rater then selects a confidence rating and provides comments.

Cooper and Harper [12] first proposed that pilot raters supply a confidence rating that reflected the ratio of information available to the pilot in the simulation to the information necessary to obtain a realistic rating. Confidence ratings can therefore be used to indicate when additional interpretation of the numeric rating is needed, and when the evaluation setting has affected the numeric rating. This can provide valuable feedback in making design decisions. In simulations, the confidence rating can help identify when the simulation environment is not sufficiently realistic and indicate when it could be difficult to extrapolate results to the real-world setting. In an operational setting, the confidence rating can help pinpoint evaluation situations in which the operational setting did not adequately represent the normally anticipated traffic or flight conditions. Such results might then be analyzed separately from the typical results expected under normal operating conditions. As in the CHS model, the CARS confidence ratings were denoted as A, B, or C, for high, moderate, or low confidence, respectively. After the confidence rating, the raters were encouraged to elaborate about their ratings and the system performance through written comments.

Despite the proposed use of the confidence rating in the initial CHS paper [12], the confidence rating was not mentioned in the more recent CHS paper [13]. Further, a cursory examination of recent research into pilot handling qualities using the CHS does not include reporting confidence factors, with the exception of Ref. 19.

Ratings of system acceptability are influenced by how well a user understands a new system, and how well a user is able to utilize a system according to the intentions of the designers [16]. Questions posed to the user must address design intentions; there must also be agreement between the rater and the designer regarding what is being rated, and how to conduct the ratings, as well as defining "adequate" versus "desired" levels of system performance. For example, in the pFAST evaluation, controllers defined adequate performance is "the system performs as well as the current system performs" and desired performance as "the system performs above and beyond the current system performance levels." Definitions of adequate and desired performance which were used in the pFAST evaluation (and which were defined by the pFAST assessment team controllers) are described in Table 1.

Table 1 shows that the CARS definitions that were used in the pFAST evaluation focused on issues of specific interest to the controllers: the impact on coordination, balancing and accuracy of runway assignments, balancing of controller workload, predictability/stability of the advisories, and accounting for aircraft performance.

As the CARS developed for pFAST research showed promise [17], it was then selected for application for URET evaluation. The URET researchers also recognized that accumulating additional empirical data on CARS would help isolate and validate a set of criteria that define acceptability. Once validated, a standardized measure would provide an objective, quantitative index of operational acceptability that could be economically applied to FFP1 and later phases of free flight research and development. The CARS rating descriptors, instructions, and confidence ratings used for pFAST required minimal adaptation

Table 1. Definitions and Guidelines for pFAST Evaluation

Adequate Performance:	**Desired Performance:**
The system performs at least as well as the current system performs.	*The system performs above and beyond the current system performance levels.*
	The system behaves predictably; reacting approximately the same way under the same conditions.
Runways balanced as well as they are currently.	Runways well-balanced, ahead of when normally expected.
Coordination between controllers is similar to what currently is required.	Coordination between controllers is reduced.
Reduced "guesswork" about where aircraft could be going.	Does away with "guesswork" about where aircraft could be going.
Advisories can be reasonably followed.	Advisories are realistic in taking into account aircraft performance. Less sequence swapping close in.
Runway assignments are good, sequence numbers are OK (not "great"). Runway assignments 90% accurate. Sequence numbers 50% accurate.	Runway assignments 90-100% accurate. Sequence numbers 75-80% accurate.
Meeting the advisories doesn't result in excessive pressure.	Workload is well-balanced. Meeting the advisories doesn't increase pressure.
	In simulations: realistic aircraft speeds.

to accommodate URET. The main changes to the URET version included replacing the references to "advisories" with "conflict probe capabilities." The resulting CARS formats for pFAST and URET are depicted in Figures 2 and 3.

3.0 Methods for Construct Validation

The CARS is intended to provide a measure of how well a DST can be used in ATM operations. Although there is some previous research into the factors that influence automation use [4], at present, there is no better known measure of acceptability against which to validate CARS. Nor is there an objective standard against which judgments of acceptability can be compared.

Our approach to validating CARS is based on data from empirical studies that were conducted to support tool design and development decisions. These studies were not focused on establishing the theoretical measurement basis for CARS and we did not design and conduct a comprehensive evaluation of CARS psychometric properties and validity [20]. The first study was a field evaluation of pFAST designed to support a decision to proceed to pFAST daily use. The second study was an experiment designed to assess the effects of URET on flight efficiency and controller performance. We analyzed data collected in these studies to examine (1) CARS reliability, or the extent to which we obtain similar CARS results when different controllers employ the measure under the same operational conditions, and (2) CARS validity, in terms of its relationship to workload factors and other system variables it is expected to assess.

3.1 Reliability Analysis

The reliability analysis was conducted using two measures of reliability, intraclass correlations (ICC) [21, 22, 23] (which are available in two forms: consistency and absolute agreement), and inter-rater reliability (IRR) agreement [24, 25, 26]. Both of these measures yield values that range from zero to one, with one representing perfect reliability. Given the differing nature of the pFAST and URET data, different reliability procedures were used to estimate interrater reliability. For the pFAST data, ICC consistency and IRR agreement measures were used. For the URET data, the ICC consistency and ICC absolute agreement measures were used.

Intraclass Correlations (ICC) Consistency and Absolute Agreement

ICC Consistency is a measure of the extent to which the rank ordering inherent in the controllers' ratings are similar. High *consistency* reliability indicates that the rank ordering of ratings can be very similar or indeed identical while the controllers' mean ratings can be very different. High ICC *absolute agreement* reliability adds to the consistency measurement the differences in the absolute values of the scores themselves. ICC measures are most appropriate when multiple raters evaluate different targets (as in the URET study); while this is a traditionally accepted

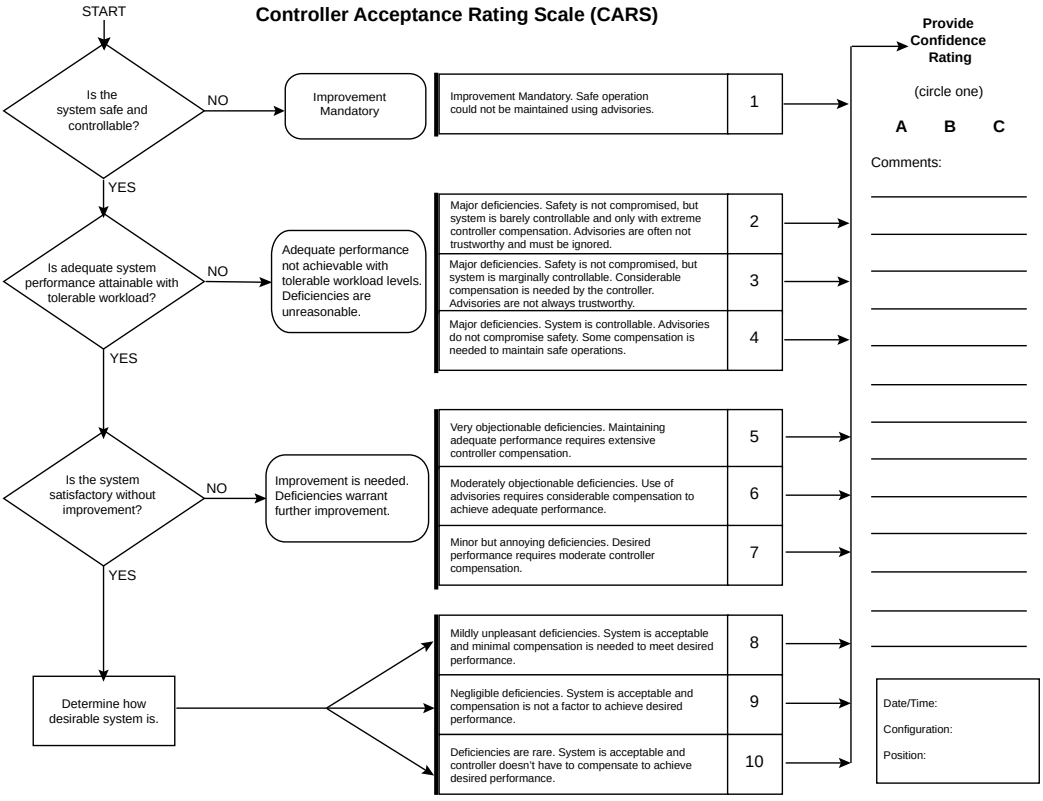

Figure 2. The CARS for pFAST

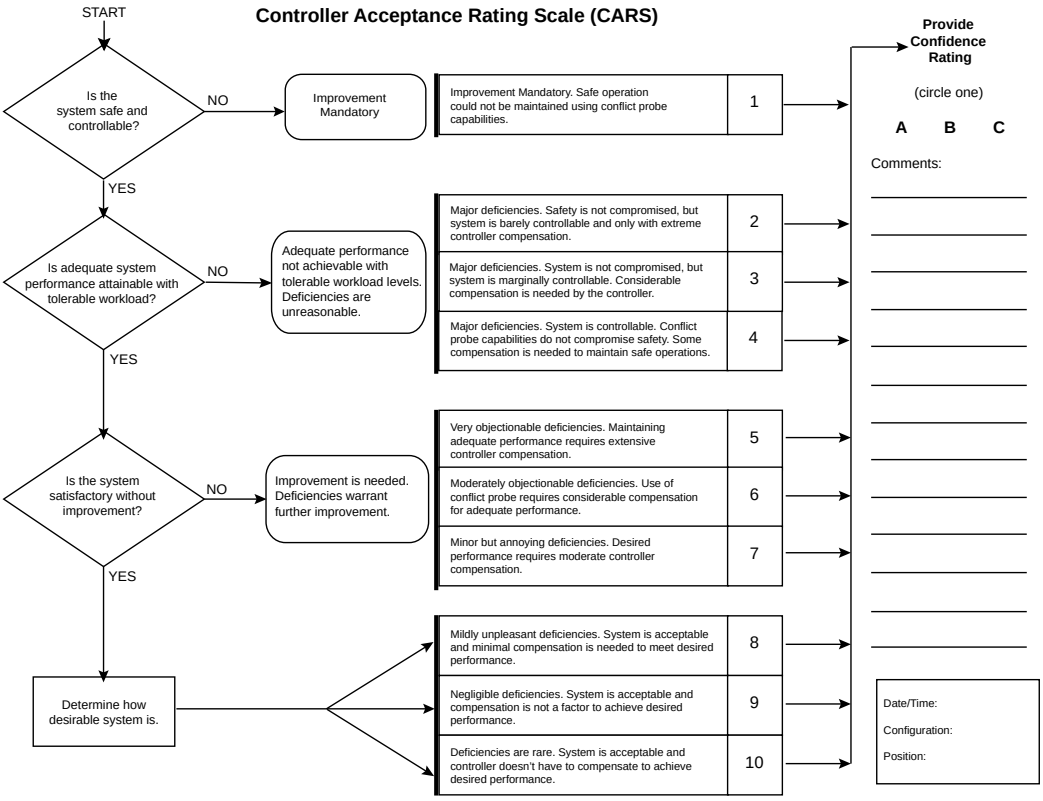

Figure 3. The CARS for URET

measure of reliability, as will be discussed below, it should not be reported as the only measure of reliability when the targets are highly similar (as in the pFAST study).

Interrater Reliability Agreement (IRR)
IRR agreement is a measure of the extent to which the controllers gave the same system acceptability ratings. It is used here to analyze the reliability of the data when the raters are evaluating similar targets. High agreement indicates that controllers, on average, all rated system acceptability similarly across positions. IRR, also designated as $r_{WG(J),}$ is the proportion of measured variance of the judges' scores to random variance, or the ratio of true variance to true variance plus error variance. Random variance for this calculation is the distribution of scores that would be expected if judges were to respond randomly from the available points on the scale. The IRR value depicts the decrease in error variance due to the agreement of judges' ratings.

In the typical case of measurement scales, the IRR calculation assumes random responding across all possible responses. But this should probably not be the case with the CARS scale (as applied to the pFAST evaluation contexts). To complete the CARS form, controllers first examine the left-hand column, with four options. Once an option on the left side of the form has been selected, the range of possible scores is bounded by those categories. As a certain level of developmental maturity in the tools can be expected, assuming the controllers could have randomly selected across the scale from 1 to 10 constitutes an unrealistic expansion of the response range that would tend to inflate any resulting estimates of interrater agreement. Therefore, modifications to the analysis were made to account for this restriction in range.

4.0 Results
The following sections contain detailed descriptions of the statistical analyses conducted on this data. For readers who wish to skip these details, a discussion of the results is found in section 5. For readers unfamiliar with statistical terminology, "significance" is a term that indicates the degree to which an observed effect is considered to be a true effect, rather than one that occurs by chance.

4.1 pFAST Data Analysis
During the pFAST field evaluation, pFAST advisories were presented on arrival controllers' radar displays in 26 different live traffic periods. In each traffic period tested, each controller worked one of seven positions, either on the East or West arrival specialty. Each specialty consisted of one of the parallel runways, a feeder position and a handoff-feeder position. The seventh position was the diagonal runway, which was worked by either East or West controllers, depending on the direction of the arrival flow. Most of the DFW traffic configurations tested were in South flow, so the majority of the diagonal runway operations were on the West side.

While a few controllers worked positions in both specialties, most worked either the East or the West, so for the reliability analysis purposes, the data from the East and West specialties were analyzed separately. The diagonal runway data was grouped with the West specialty data.

Following each traffic period in which pFAST advisories were presented, controllers were asked to fill out numerous questionnaires, including a CARS form and a modified NASA Task Load Index (TLX) scale [27]. The modified TLX was changed from that of the original NASA TLX to reflect the evaluation of workload experienced by controllers in an ATC setting. The original TLX Physical Effort rating was not included in the modified scale as controllers decided that this was not a relevant question. A total of 166 cases were available for the analysis presented here.

4.2 pFAST Results
A group of ten controllers worked the seven arrival positions. Position and controller assignments during the tests were determined by the controller teams and were outside of experimental control; thus not all controllers worked each position an equal number of times (which would have resulted in a complete data matrix). As a result, the CARS scores were spread unevenly across positions and runs, posing a problem for reliability analysis. The number of times controllers worked each position ranged from 1 to 12, with a mean of 4.1. One controller worked only one run and another worked only five runs over four positions; these controllers' data were not included in the reliability analysis. East-side controllers worked the east positions a minimum of three times, so there were controller means for every controller/position for that analysis.

Overall, the mean CARS rating, averaged over all the traffic periods and all the controllers was 7.8 (SD = 1.1). Rounded to 8, this mean rating corresponds to the description, "Mildly unpleasant deficiencies. System is acceptable and minimum compensation is needed to meet desired performance."

4.2.1 Reliability

Mean CARS scores were calculated for each controller on each position. Examination of the controller data determined that the pattern of scores for one West specialty controller were suggestive of a multivariate outlier. Anecdotal evidence determined that this controller was relatively less experienced than other members of the evaluation team, so reliability analyses were done without this controller's responses.

4.2.1.1 ICC Consistency

A two-way, mixed effects model, average measure reliability, ICC(3,k) [21] was used for all inter-rater reliability consistency analyses. For the East specialty, the reliability was ICC(3,4) = .91, $p < .01$. For the West specialty, which included the diagonal runway position, reliability was ICC(3,4) = .49, $p = .21$. Removing the diagonal runway scores from analysis of the West specialty data resulted in ICC = .81, $p = .11$.

4.2.1.2 IRR Agreement

As alluded to above, ICC measures of reliability are more appropriate in situations in which multiple raters evaluate different targets. In the pFAST evaluation, there is a restriction of the overall range of ratings when similar targets are rated. For the IRR analysis, a range of 5 to 10 was assumed for the random measure part of the IRR calculations. As in the ICC measures, mean CARS ratings were calculated for each controller on each position. The Spearman-Brown prophecy formula was applied to the basic IRR equation and yielded a multiple-item estimate of $r_{wg(7)} = .94$ for the West specialty, $r_{wg(7)} = .996$ for the East specialty, and $r_{wg(7)} = .96$ when combining the specialties. There is no significance test available for IRR calculations. As with other reliability measures, higher values represent better interrater agreement, with an upper bound of 1.00.

4.2.2 Construct Validity

Validity was analyzed by examining the relationship between CARS scores and data from controller questionnaires. As reported in Ref. 9, the CARS results were significantly, positively correlated with the controllers' self-reported agreement with the runway advisories and significantly, negatively correlated with how often the controllers considered the sequence numbers to be in error. The CARS ratings were also significantly negatively correlated with the controllers' self-reported ratings of the amount of effort required to accomplish the controlling tasks, and significantly negatively correlated with the difficulty of managing and controlling the traffic feed.

4.2.2.1 Relationship between CARS Ratings and Workload Measures

The relationship between CARS and the TLX workload factors was also analyzed. A multiple regression was performed on the CARS data using the five modified TLX factors as predictors. The analysis showed a significant relationship between the TLX factors and the CARS score ($F_{5,160} = 19.2$, $p < .0001$). The R^2 was .38; the adjusted R^2 was .36, suggesting that the modified TLX factors explain about 36% of the variance in the CARS' numerical ratings.* The standardized regression coefficients for each predictor and the associated *t* test values are shown in Table 2.

The TLX is composed of three types of scales: task-related (which reflect the objective demands imposed on the operator/rater), behavior-related (which reflect the subjective evaluations of effort that the raters exerted to satisfy their task requirements) and subject-related (which reflect the psychological impact on the raters) [27]. As Table 2 shows, the pFAST results showed that the subject-related scale (Satisfaction/Frustration) and the behavior-related scale (Overall Effort) were significant contributors to controller acceptance.

These results are consistent with the expected behavior of CARS in this study context. The TLX subject-related scales reflect more of the psychological impact on the operator and the effort required to perform and satisfy task demands. The objective demands (traffic levels) were not manipulated in the pFAST evaluation, thus significant variability was not observed in the controller ratings of the task-related workload elements of mental and temporal demand.

4.2.2.2 Confidence Ratings

Of the 166 CARS numeric ratings that were collected, 145 of these (87%) included confidence ratings. A statistically significant correlation was found between the CARS numeric ratings and the confidence ratings, $R = -0.43$, $p < .01$. The

* R^2, the squared multiple correlation, is a measure of effect size representing the proportion of variance in the dependent variable (in this case the CARS scores) that is accounted for by the linear combination of independent variables (in this case the weighted TLX variables). Its value ranges from zero to one, with higher values meaning more accurate prediction of the CARS scores by the TLX values. Because a large number of independent variables and small sample size can artificially inflate the value of R^2, an "adjusted R^2" value is calculated to account for such effects. More details on this statistic can be found in any basic multivariate statistics textbook.

Table 2. Results of Multiple Regression Analysis of CARS Scores on Five Modified NASA-TLX Factors

Predictors	Standardized Regression Coefficients	t statistic	Significance
Satisfaction/Frustration	-.347	-4.48	< .0001
Overall Effort	.204	2.07	.0397
Performance Support	-.121	-1.65	
Mental Demand	.120	1.01	
Temporal Demand	.087	.76	

correlation result demonstrates that the higher CARS ratings were associated with higher confidence levels; the mean CARS rating when high confidence was rated (A) was 8.2. Table 3 shows the distribution of the confidence ratings, together with the associated mean CARS ratings.

Table 3. Distribution of Confidence Ratings

Confidence Level	Mean CARS	SD	Count
A – high	8.2	1.0	102
B – moderate	7.2	.90	37
C – low	6.7	1.4	6

The distribution of the confidence ratings also show that 61% of the CARS ratings were accompanied by a high confidence in the rating (A); thus in 61% of the runs in which CARS data and confidence data were available, the controllers reported that their rating was appropriate given their knowledge of the traffic situation, performance of the software, and performance of the equipment, and they felt that their rating sufficiently represented an evaluation of pFAST performance. The controllers also rated very few instances of low confidence (C), suggesting that there were few instances in which they felt that their acceptability ratings could be in question, and might not be related to the performance of the software. These confidence ratings results add credibility to the higher CARS ratings, and suggest that the evaluation environment was sufficient to make an evaluation that was not compromised by the testing environment or unforeseen operational factors.

4.3 URET Results

An experiment was conducted in the dynamic simulation (DYSIM) training facility at Indianapolis Center to identify and quantify benefits associated with use of the URET in the current and emerging unstructured traffic environments [28]. The experiment used a within-subjects design to measure the effects of URET and traffic conditions. The URET variable was defined by two levels—on or off. Traffic condition was defined by three levels—structured, unstructured, and high-volume unstructured. For the structured condition, scenarios were created from actual recorded flights from the Center, reflecting a moderate traffic volume. For the unstructured condition, scenarios were created using the same set of flights as the structured condition, but all of the aircraft were placed on direct routes between the origin and destination. For the high-volume, unstructured condition, scenarios were created by adding flights to the unstructured scenarios until the traffic volume increased by 25%. Combining these two independent variables resulted in six test conditions. Twelve participants were divided into six Radar (R) and Radar Associate (D), controller teams for the six test sessions. Dependent measures included acceptability, measured by the CARS, and workload, measured by the NASA TLX. The CARS was adapted for URET conditions as shown in Figure 3. For test conditions without URET, the term "system" was substituted for the term "conflict probe" in the CARS descriptors. The mean CARS rating averaged over the URET test conditions for all controllers was 8.3 (S.D.= 1.1), corresponding to a description, "Mildly unpleasant deficiencies. System is acceptable and minimum compensation is needed to meet desired performance." The mean overall TLX ratings were below 50 on the 100-point scale, indicating light to moderate workloads were experienced under all test conditions.

4.3.1 Reliability

Because URET was expected to affect R and D controller ratings differently, we analyzed CARS scores for the R and D positions separately. As discussed above, ICC consistency and absolute agreement tests were used in the URET analysis.

4.3.1.1 Consistency

A two-way mixed effect model, intraclass correlation showed that the controller ratings of the same conditions were highly consistent. The average measures of intraclass correlation were significant for the R and D controllers ($ICC[3,6] = .78$, $p < .01$ and $ICC[3,6] = .81$, $p < .01$, respectively). These results indicate that the CARS was effective in allowing controllers to consistently discriminate among the

different levels of acceptability represented by the operational conditions.

4.3.1.2 Agreement

We then examined a two-way mixed effects model for an average measure of intraclass correlation with absolute agreement. Under this definition, the average measures of intraclass correlation were lower but also significant (ICC[3,6] =.74, $p < .01$ for R controllers and ICC[3,6] =.77, $p < .01$ for D controllers). These results indicate that there was good agreement among controllers in terms of the precise scores assigned to each condition.

4.3.2 Construct Validity

Validity was analyzed by examining the relationship between CARS scores and TLX scores and by examining the effect of URET DST use on CARS scores.

4.3.2.1 Relationship Between CARS Ratings and Workload

A multiple regression analysis was run using the CARS as the criterion and the six TLX subscales as predictors, all entered simultaneously into the analysis. Overall, the full model containing all six predictors accounted for 20% of the variability in CARS ratings, $F_{6,65} = 2.78$, $p \leq .01$, adjusted $R^2 = 0.13$. The standardized regression coefficients for each predictor and the associated *t* test values are shown in Table 4. Only one subject-related scale, Frustration, was a significant predictor of the CARS rating. Two task-related scales, Mental Demand and Temporal Demand, were marginally significant predictors. Negative regression coefficients for frustration level and mental demand indicated that acceptability was higher when frustration and mental demand were lower. A positive regression coefficient for temporal demand indicated that acceptability was higher when the temporal pace of tasks was faster. Because light to moderate traffic loads were experienced in this study, it is possible that the positive relationship between CARS and temporal demand reflects the expected effect of under-load on acceptability.

4.3.2.2 Relationship Between CARS and URET

Although analyses of R and D controller TLX data did not reveal any main effect for URET, the analyses of R and D controller CARS ratings revealed that URET significantly improved the operational acceptability of the unstructured traffic conditions. However, this effect was observed only for the D controller. Figure 4 shows the CARS scores for the D controllers.

Table 4. Results of Multiple Regression Analysis of CARS Scores on Six Workload Factors

Predictors	Standardized Regression Coefficients	t statistic	Significance
Frustration	-.550	-3.37	.001
Mental Demand	-.668	-1.89	.064
Temporal Demand	.641	1.90	.062
Effort	.216	.84	
Performance	-.078	-.60	
Physical Demand	-.015	-.06	

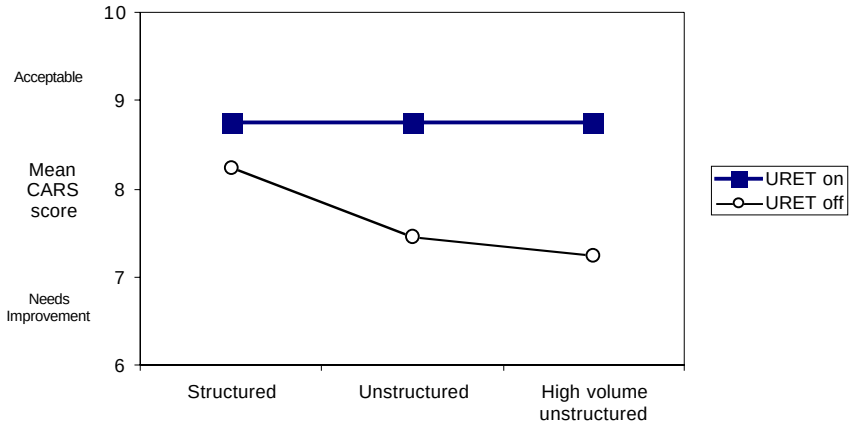

Figure 4. CARS Scores for D controllers

The Friedman test indicated a significant difference in the test conditions, $\chi^2(5) = 10.69$, $p \leq .05$, indicating an interaction of URET with the traffic conditions. While acceptability was essentially equivalent in structured conditions with or without URET, there was a continuing drop in acceptability under the unstructured and high-volume unstructured conditions without URET. Thus the introduction of URET improved the operational acceptability under the free routing test conditions.

5.0 Discussion

The results reported in this paper are drawn from relatively small sample sizes, which created difficulties in the analysis. Some of the sample size issues are attributable to the problems inherent in a field test setting. In the pFAST evaluation, the controller subjects were primarily responsible for controlling live traffic and the evaluation of the pFAST advisories was of secondary importance. Therefore, staffing of controller positions during the pFAST test, as well as the frequency with which a controller worked a position, was left up to the discretion of the area supervisor, rather than strictly determined by experimental design. The reliability analysis is therefore affected by the missing data and resulting small sample sizes. In addition to lack of control over staffing, loss of data due to subjects neglecting to fill out the CARS forms also occurred (9% of the cases had missing CARS scores).

In the case of the URET data, the experimental setting allowed a measure of control over the conditions tested and may account for the high consistency among CARS ratings of the same conditions by different controllers. At the same time, the small sample size relative to the number of predictor variables constrained the analyses that could be performed and accounts for the magnitude of the shrinkage observed in the adjusted R^2.

Problems noted in both data sets include the limitation of the actual range of ratings (and subsequent low standard deviations). Random error may have also been elevated by the time span of the pFAST data collection period (6 months).

All of this considered, the obtained CARS scores showed relatively good scale reliability. From previous research in the evaluation of instructor/evaluator agreement in assessments of aircraft simulator proficiency checks [29], an average interrater correlation of $r_{WG(J)} = .54$ was considered acceptable and agreement of $r_{WG(J)} = .80$ was considered high. With larger sample sizes we would expect to see respectable ICC values, although probably lower than the highest values in our analyses (i.e., ICC = .91; IRR [$r_{WG(7)}$] = .96 obtained for pFAST and ICC = .78 and ICC = .81 for URET).

In view of the different research contexts and differences in the DSTs evaluated, the level of agreement between the studies is encouraging. The pattern and direction of relationships observed in the data accurately reflects the behavior expected of measures of the acceptability construct. The pFAST results suggest that CARS is capturing elements of controller satisfaction and frustration, as well as overall effort. Consistent with these results, the URET results also showed that CARS was related to controller satisfaction and frustration, as well as perceived levels of mental and temporal workload. Comparing the results of the two studies, there were some inconsistencies with regard to the relative importance of various TLX subscale predictors. Additional data are needed to investigate whether these inconsistencies are artifacts of the data collection environments or limitations in the CARS itself. Finally, both studies further suggest that workload accounts for a significant but limited portion of the variance in CARS ratings, 36% with the (modified) TLX in the pFAST study, and 20% with the TLX in the URET study. Presumably, the remaining variance in CARS is attributable to non-workload factors which influence controller acceptance. Results from the URET study are consistent with this explanation. In that study, CARS was sensitive to the effects of introducing URET while the TLX was not.

5.1 Lessons Learned

Two general categories of lessons learned arise from the analyses presented in this paper: improvements in data collection and improvements towards the application of the CARS itself. Improvements to either category would enhance future attempts to assess CARS reliability and validity.

5.1.1 Data Collection

The collection of subjective data invariably poses risks for obtaining adequate, representative sample sizes. The lack of control in the field further compromises the data collection process. Researchers could attempt to apply more control over the test environment through counterbalancing controller participation; this would help in reducing unidentified sources of systematic variance into the data from any controller-controller or controller-position interactions. In the absence of *a priori* counterbalancing, attempts could be made in the latter stages of a field evaluation to fill in empty cells of the test matrix. Further, to the extent that the

controller participants can be selected ahead of time, it would also be advantageous to try and keep experience levels (in terms of overall controller experience and experience within a specialty) as homogeneous as possible.

In addition, it is important that the controllers that assess DST performance be appropriately trained in the use and the purpose of the tools; at least a uniform level of experience for the controllers should be expected. It is often impossible to guarantee this type of uniformity due to constraints on controller participation.

5.1.2 Improvements in the Application and Validation of CARS

In addition to sampling controllers, validation should be concerned with sampling operational conditions. The data presented in this paper represent a relatively limited set of operational contexts; more data with a wider range of operational conditions, as well as different DSTs, is needed to further examine the relationships between CARS and other variables thought to influence acceptability. With a sufficient controller sample, additional predictor variables including measures of automation reliability, controller comfort level, controller proficiency and understanding of the DST functionality, as well as TLX factors, could be examined. Finally, some measure of actual DST use is also needed to serve as a criterion.

Improvements in the implementation of the CARS can also be made by refining the terminology/ guidelines used in the scale. The pFAST data, for example, was collected over the course of approximately six months. While the controller team that participated in the pFAST test helped to define all the elements used in the scale, it might have been valuable to review the CARS guidelines and definitions mid-way through the testing period. This would have enabled the raters to raise any questions or concerns about their interpretation of the scale, and would have probably helped to insure that the CARS was being interpreted consistently.

While the confidence ratings have received relatively little attention in the research literature, it remains an interesting area for further exploration. Future research into CARS validity could devise a method of weighting the CARS scores with the confidence ratings.

While CARS does need further refinement, we have shown that it is a useful tool for researchers to use to evaluate controller acceptance. Its methodology and application lend itself to post-experiment administration, and it has a straightforward approach toward obtaining a rating. Careful definition of the guidelines and terminology used in the scale is critical in the use of the scale as well as the interpretation of the results. We are seeking to validate CARS with new candidate decision-support tools and technologies, particularly those being developed in the European Community. In doing so, we would be interested in providing assistance in applying the CARS.

CARS should also be further researched in order to determine if it can be made into a more generic format for evaluating multiple DSTs without having to tailor the scale for each tool being developed. This more general CARS could then be used much as the CHS is used, though its application would still require that significant time be spent on defining the tasks and the guidelines by which the system performance would be judged.

6.0 Conclusions

CARS was used to measure controller acceptance of two different controller tasks, for two different DSTs, pFAST and URET, in two different study contexts. Our results suggest that the CARS (1) allows controllers to consistently discriminate among the different levels of acceptability represented by operational conditions, (2) accurately measures selected facets of workload that influence the controller's use of, and satisfaction with, the DSTs, and (3) is more sensitive to the psychological effects of introducing DSTs, such as satisfaction and effort, than a task-related workload measure.

In both studies, CARS was shown to be a simple measure to implement and use for data collection. However, before CARS can be used, it requires significant investment on the part of the researcher to clearly define and train the users on how the scale is structured, as well as to reach a clear definition of the elements that comprise the scale descriptors.

7.0 References

1. RTCA. (1998, August) Government/Industry Operational Concept for the Evolution of Free Flight Addendum 1: Free Flight Phase 1 Limited Deployment of Select Capabilities. Washington, DC: RTCA, Inc.
2. Deaton, J.E., & Morrison, J.G. (1999). Aviation Research and Development: A Framework for the Effective Practice of Human Factors, or "What Your Mentor Never Told You About a Career in Human Factors." In D. Garland, J.A. Wise, and V.D. Hopkin (Eds.), Handbook of Aviation Human Factors (pp. 15-32), Lawrence Erlbaum Associates, Mahwah, NJ.

3. Center for Aviation/Aerospace Research, Embry-Riddle Aeronautical University (1994). Human Factors Challenges in Traffic Flow Management: Preliminary Issues. CAAR-15418-94-101, Daytona Beach, FL.
4. Parasuraman, R. and Riley, V. (1997). Humans and Automation: Use, Misuse, Disuse, Abuse. Human Factors, 39(2), 230-253.
5. Hilburn, B.G., Jorna, P.G., Byrne, E.A., and Parasuraman, R. (1997). The Effect of Adaptive Air Traffic Control Decision Aiding on Controller Mental Workload. In M. Moula and J.M. Koonce (Eds.) Human-Automation Interaction: Research and Practice (pp. 84-91), Lawrence Erlbaum Associates, Mahwah, NJ.
6. Hendy, K., Hamilton, K., and Landry, L. Measuring Subjective Workload: When is one Scale Better than Many? (1993) Human Factors, 35(4), 579-602.
7. Green, D. L., Andrews, H., & Gallagher, D. W. (1993). Interpreted Cooper-Harper for broader use. Paper presented at the Piloting Vertical Flight Aircraft: A Conference on Flying Qualities and Human Factors, 221-234.
8. Davis, T.J., Robinson, J.E. III, Isaacson, D.R., den Braven, W., Lee, K.K., & Sanford, B.D. (1997). Operational Test Results of the Final Approach Spacing Tool. Proceedings of the 8th IFAC Symposium on Transportation Systems, June 16-18 1997, Chania, Greece.
9. Lee, K.K., & Sanford, B.D. (1998). Human Factors Assessment: The Passive Final Approach Spacing Tool (pFAST) Operational Evaluation. NASA Technical Memorandum 208750.
10. Brudnicki, D.J. & McFarland, A.L. (1997). User Request Evaluation Tool (URET) Conflict Probe Performance and Benefits Assessment, MP97W112, The MITRE Corporation, McLean, VA.
11. Kirk, D.B., Heagy, W.S., McFarland, A.L., & Yablonski, M.J. (2000). Observations about Providing Problem Resolution Advisories to Air Traffic Controllers. Proceedings of the 3rd USA/Europe Air Traffic Management R&D Seminar, Naples, Italy.
12. Cooper, G.E., & Harper, R.P. (1969). The Use of Pilot Ratings in the Evaluation of Aircraft Handling Qualities. NASA TN-D-5153. Moffett Field, CA: NASA Ames Research Center.
13. Harper, R.P., & Cooper, G.E. (1986). Handling Qualities and Pilot Evaluation. Journal of Guidance, 9(5), 515-529.
14. Mitchell, D.G., & Aponso, B.L. (1990). Reassessment and Extensions of Pilot Ratings with New Data. AIAA-90-2823-CP.
15. Wierwille, W.W., & Casali, J.G. (1983). A Validated Rating Scale for Global Mental Workload Measurement Applications. Proceedings of the Human Factors Society 27th Annual Meeting, 1983 (pp. 129-133). Santa Monica, CA: Human Factors Society.
16. Wilson, D.J., & Riley, D.R. (1989). Cooper-Harper Pilot Rating Variability. AIAA 89-3358-CP. AIAA Atmospheric Flight Mechanics Conference, Boston, MA, Aug. 14-16, 1989, Washington, DC, American Institute of Aeronautics and Astronautics, p. 96-105.
17. Lee, K.K., & Davis, T.J. (1996). Development of the Final Approach Spacing Tool (FAST): A Cooperative Controller-Engineer Design Approach. Control Engineering Practice, 4(8).
18. Schick, F. (1998). Methods and Measurements for the Evaluation of ATM Tools in Real-Time Simulation and Field Tests. Proceedings of the 2nd USA/Europe Air Traffic Management Research and Development Seminar. Orlando, FL.
19. Schroeder, J.A., & Chung, W.W.Y. (2000). Simulator platform Motion Effects on Pilot-Induced Oscillation Prediction. J. of Guidance, Control, and Dynamics. 23 (3), pp. 438-444. May-June 2000.
20. Campbell, D.T., & Fiske, D.W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. Psychological Bulletin, 56, 81-105.
21. Shrout, P.E., & Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. Psychological Bulletin 86(2): 420-428.
22. McGraw, K.O., & Wong, S.P. (1996). Forming inferences about some intraclass correlation coefficients. Psychological Methods 1(1): 30-46.
23. Yaffee, R.A. (1998). Enhancement of reliability analysis: application of intraclass correlations with SPSS/Windows v.8. [On-Line]. Available: http://www.nyu.edu/acf/socsci/Docs/intracls.html .
24. James, L.R., Demaree, R.G., & Wolf., G. (1993). r-sub(wg): An assessment of within-group interrater agreement. Journal of Applied Psychology 78(2): 306-309.
25. Law, J.R., & Sherman, P.J. (1995). Do raters agree? Assessing inter-rater agreement in the evaluation of air crew resource management skills. International Symposium on Aviation Psychology, 8th, Columbus, OH, Ohio State University: 608-612.
26. Williams, D.M., Holt, R.W., & Boehm-Davis, D.A. (1997). Training for inter-rater reliability: baselines and benchmarks (for aviation instructor/evaluators). International Symposium on Aviation Psychology, 9th, Columbus OH, Ohio State University: 514-520.
27. Hart, S.G., & Staveland, L.E. (1988) Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In P.A. Hancock & N. Meshkati (Eds.), Human Mental Workload (pp. 239-250). Amsterdam: North Holland Press.
28. Kerns, K. (2001). An Experimental Approach to Measuring the Effects of a Controller Conflict Probe in a Free Routing Environment. IEEE Transactions on Intelligent Transportation Systems, 2(2), June 2001, 81-91.
29. Holt, R. W., Meiman, E., & Seamster, T. L. (1996). Evaluation of aircraft pilot team performance. Human Factors and Ergonomics Society, 40th Annual Meeting, Philadelphia, PA, Human Factors and Ergonomics Society: 44-48.

Author Biographies

Katharine K. Lee is a human factors researcher in the Terminal Area Air Traffic Management Research Branch at NASA Ames Research Center in Moffett Field, California. Since 1993, she has conducted human factors research and evaluations of the Center/TRACON Automation System (CTAS) decision support tools. She is currently co-leading the research and development of the Multi-Center Traffic Management Advisor (McTMA) under NASA's Advanced Air Transportation Technologies (AATT) Project and previously led the human factors evaluations of pFAST simulation and field testing. She received an M.A. in Psychology from San Jose State University and received her B.A. degrees in Psychology and Biophysics from the University of California at Berkeley.

Karol Kerns currently serves as a Principal Scientist with the MITRE Corporation, Center for Advanced Aviation System Development (CAASD). She recently returned from a two-year interagency assignment as the Human Factors Lead for the Federal Aviation Administration (FAA) Aeronautical Data Link Product Team where she coordinated human factors research, development, and implementation activities on The Controller Pilot Data Link Communications Program. Prior to joining the FAA, Dr. Kerns has been with the MITRE CAASD since 1983. At MITRE, she had responsibility for the application of the human factors principles and methods to the development of next generation air traffic control systems, including advanced communications and decision support tools. Her principal duties included development of operational concepts detailing the human-computer function allocation, prototyping the human computer interface portion of the system, design and conduct of human-in-the loop operational simulations, benefits assessment, and participation in government-industry standards committees. Dr. Kerns received her B.A. degree from LaSalle University in 1974 and her M.S. and Ph.D. degrees in experimental psychology from Saint Louis University in 1976 and 1980, respectively.

Randall Bone earned a M.S. in Engineering Psychology in 1998 from the University of Illinois at Urbana-Champaign. He has worked in the field of aviation operations, human factors, and safety at the Federal Aviation Administration, National Transportation Safety Board, University of Illinois, United Airlines, and the Aircraft Owners and Pilots Association Air Safety Foundation. Mr. Bone is an instrument flight instructor as well as an advanced and instrument ground instructor. He is currently an Operations / Human Factors Specialist at the MITRE Center for Advanced Aviation System Development.

Monicarol Nickelson is a Human Factors Research Engineer/Psychologist for the Federal Aviation Administration (FAA) at NASA Ames Research Center. She was the first member of the FAA's new Technology Transfer team, whose task is to assist NASA in the research and development of new decision-support tools and then to provide the expertise and guidance necessary for the FAA to implement and maintain those tools for operational use. Ms. Nickelson is a former controller and private pilot. She received her B.A. and M.A. degrees from Wichita State University and is currently a Doctoral Candidate in Human Factors Psychology there. At NASA, she is presently working on the development of McTMA and Direct-To DSTs, and is conducting research on Dynamic Density and Distributed Air-Ground Traffic Management.