Extended Abstract/Draft Paper: ATM 2001

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

Katharine K. Lee, *NASA Ames Research Center*
Karol Kerns, *The MITRE Corporation Center for Advanced Aviation System Development*
and Randall Bone, *The MITRE Corporation Center for Advanced Aviation System Development*

Abstract

The measurement of operational acceptability is important for the development, implementation, and evolution of air traffic management decision support tools. The Controller Acceptance Rating Scale was developed 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. In this paper, we provide a discussion of the development of CARS and an analysis of the empirical data collected with CARS to examine construct validity. Results of intraclass correlations indicated statistically significant reliability for the CARS. From the subjective workload data that were collected in conjunction with the CARS, it appears that the expected set of workload attributes was correlated with the CARS. As expected, the analysis also showed that CARS was a sensitive indicator of the impact of decision support tools on controller operations. Suggestions for future CARS development and its improvement are also provided.

INTRODUCTION

The Federal Aviation Administration (FAA) 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, technical guidance on human factors methods and measures is needed to support the evolutionary system development process envisioned by the RTCA (RTCA, 1998). 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 (Deaton & Morrison, 1999).

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 is an air traffic management (ATM) measurement construct that 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 effectiveness of the functionality embodied in the equipment and its suitability for human use in the performance of tasks accomplished in the specified environment. Effectiveness and suitability are generally considered necessary but not sufficient conditions for operational acceptability. Acceptance is also influenced by less-easily-measured constructs such as impact on job satisfaction, the comfort level of the operator performing the prescribed duties, and the amount of required training. The construct further assumes that operational acceptability is highly correlated with the extent to which a DST will actually be used.

DEVELOPMENT AND APPLICATION OF THE CONTROLLER ACCEPTABILITY RATING SCALE (CARS)

pFAST

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 and is part of the FAA FFP1 deployment is the Passive Final Approach Spacing Tool (pFAST). Passive FAST is designed to provide advisory information to terminal-area radar controllers for the efficient runway balancing and sequencing of arrival traffic. More information regarding the development of pFAST can be found in Davis, Robinson, Isaacson, den Braven, Lee, & Sanford (1997) and Lee & Sanford (1998).

The development of pFAST led to the creation of the Controller Acceptance Rating Scale (CARS) for examining controller acceptance (Lee & Davis, 1996). As described earlier, the operational acceptance of a DST is dependent upon more than just the engineered performance of the DST. In the development of pFAST, during simulation testing, the goal was to determine when the tool was acceptable to begin an operational evaluation, and then during the operational evaluation, the goal was to determine when pFAST had demonstrated that it was acceptable as a daily-use operational DST. While controller comments and questions can provide indications about tool acceptability, more objective, quantitative 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. The Cooper-Harper Scale (CHS) was selected as a basis upon which to model a controller acceptability scale.

The CARS utilized many elements of the CHS, including the process of defining the attributes of the scale and the guidance that is required for its use (Harper & Cooper, 1986). The direction of the CHS rating scale was reversed for CARS, such that “1” was unacceptable, and “10” was completely acceptable. For presentation to the raters, the scale was also reoriented to move from top to bottom, rather than from bottom to top. The wording of the CHS was changed for CARS to reflect the controller evaluation of a software system, rather than the pilot evaluation of aircraft handling qualities. The key categories of evaluation for the CHS, retained in the CARS, were controllability, tolerability, satisfaction, and desirability (acceptability).

URET

Based on years of collaborative laboratory research to develop en route automation tools, the FAA and the MITRE Center for Advanced Aviation System Development have been conducting operational trials of an initial DST for the sector team, called the User Request Evaluation Tool (URET). A URET prototype continues in daily use at Indianapolis and Memphis Air Route Traffic Control Centers (ARTCCs) and is part of the FFP1 deployment. The 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 which 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 Brudnicki and McFarland (1997) and Kirk, Heagy, McFarland, and Yablonski (2000).

Like the pFAST development, the development of URET capabilities requires a measure of operational acceptability to routinely assess progress throughout the life cycle. The goal is to validate a measurement method that can be applied in the laboratory to determine when tool enhancements are ready for operational evaluation and in operational settings to determine when the DST is ready for daily use (Schick,1998). Because the initial experience using the CARS for pFAST evaluation was promising (Lee and Davis, 1996), it was selected for application to URET. We 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 to accommodate URET.

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 (Parasuraman & Riley, 1997), 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 in the FFP1 program. 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 (Campbell & Fiske, 1959). The first study was a field evaluation of pFAST designed to support a decision to proceed to daily use of pFAST. The second study was a field 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.

EVIDENCE FROM EMPIRICAL RESEARCH

pFAST

The pFAST field evaluation took place at DFW TRACON over the course of six months, during which pFAST advisories were presented to 5 to 7 controllers in 26 different traffic periods for controlling arrival traffic. CARS ratings as well as other questionnaire data were collected from each controller after each test period; a total of 166 cases were available for analysis. One of the questionnaires was based on the NASA Task Load Index (TLX). For the pFAST evaluation, the wording of the TLX scale was changed to reflect the evaluation of workload experienced by controllers in an ATC setting. (The Physical Effort rating from the NASA TLX was not included in this modified TLX scale, at the request of the controllers.) Overall, the mean CARS rating, averaged over all the traffic periods and all the controllers was 7.8 (SD = 1.1), which rounds to 8. This rating corresponds to the description, "Mildly unpleasant deficiencies. System is acceptable and minimum compensation is needed to meet desired performance."

CARS Validity

As reported in Lee & Sanford (1998), 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.

Although an analysis performed on the modified TLX data did not demonstrate any significant difference in the workload ratings between the two arrival specialties, the analysis of CARS ratings showed that the West side controllers rated the pFAST performance as significantly higher in acceptability than the East side controllers ($F[1,118] = 5.69$, $p < .02$).

Workload Correlations

Correlations between the CARS ratings and the modified TLX ratings were performed. In addition, a Fisher's r-to-z test was performed to examine the statistical significance of the correlation coefficients. As shown in Table 1, the modified TLX variables of satisfaction/frustration and overall effort were the most highly correlated with the CARS. The lower the frustration rating, the higher the CARS rating. The lower the overall effort rating, the higher the CARS rating.

Table 1.

CARS/TLX1 – mental demand	.397, $p < .0001$
CARS/TLX2 – temporal demand	.377, $p < .0001$

CARS/TLX3 – performance support	.184, $p < .05$
CARS/TLX4 – overall effort	.455, $p < .0001$
CARS/TLX5 – satisfaction vs. frustration	.515, $p < .0001$

Regression Analysis

A multiple regression was performed on the CARS data using the five modified TLX variables as predictors. The R^2 was .38; the adjusted R^2 was .36, suggesting that the modified TLX variables explain about 36% of the variance in the CARS' numerical ratings. Again, the Satisfaction/Frustration and Overall Effort scales were significant predictors of the CARS ratings.

The TLX is composed of three types of scales: task-related, behavior-related, and subject-related (Hart & Staveland, 1988). The pFAST results found that a behavior-related scale (frustration level) and a subject-related scale (overall effort) contributed to controller acceptance, more so than the task-related scales (such as mental demand, temporal demand). These results are consistent with the expected behavior of CARS in this study context. The TLX subject-related scales reflect the psychological impact on the operator of performing the tasks while behavior-related scales reflect the effort that the operator exerted to satisfy the task demands. The task-related scales reflect the objective demands imposed on the operator. Because this study did not manipulate the traffic demand experienced during the test, this may have reduced the variability in controller ratings of the task-related drivers of workload.

URET

A field 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 (Kerns, in press). 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. Combining these two independent variables resulted in six test conditions. Twelve participants were divided into six, Radar (R) and D, controller teams for the six test sessions. Dependent measures included acceptability, measured by the CARS, and workload, measured by the NASA TLX. The mean CARS rating averaged over the URET test conditions for all controllers was 8.33 (S.D.= 1.08), corresponding to a description, "Mildly unpleasant deficiencies. System is acceptable and minimum compensation is needed to meet desired performance." In general, TLX scores fell below 50 on the 100-point scale, indicating light to moderate workloads were experienced under all test conditions.

CARS Reliability

We used an intra-class correlation to assess the consistency of CARS ratings of the same conditions by different controllers (McGraw and Wong, 1996). Because URET was expected to affect R and D controller ratings differently, we analyzed the R and D

controller data separately. For the R and D controllers, the average measures of intraclass correlation were significant (r'= .78, p≤.01 and r'= .81, p≤.01, respectively). These correlations indicate that the CARS was effective in allowing controllers to consistently discriminate among the different levels of acceptability represented by the operational conditions.

CARS Validity

Workload

A multiple regression analysis was run using the six TLX subscales as predictors and the CARS as the criterion. Overall, there was a significant multiple correlation between the CARS and the TLX subscales (R=.45, p≤.01), indicating that the TLX accounted for about 20% of the variability in the CARS ratings. Three of the subscales accounted for a significant proportion of the variance in the CARS: a subject-related scale (frustration) and two task-related scales (mental demand and temporal demand). Significant negative relationships between CARS and frustration level and CARS and mental demand indicated that acceptability was higher when frustration and mental demand were lower. A significant positive relationship between CARS and temporal demand indicated that acceptability was higher when the 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.

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 1 shows the CARS scores for the D controllers.

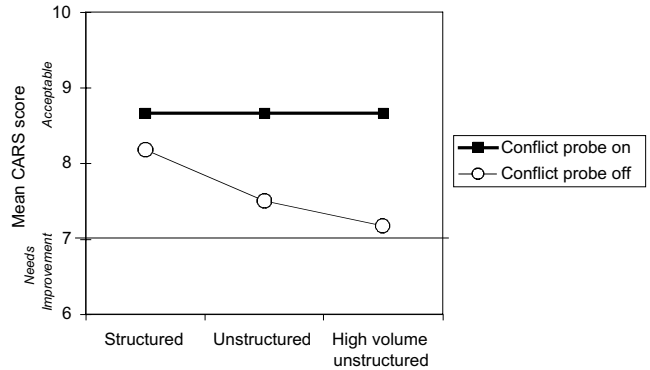

Figure 1. Mean D controller CARS scores as a function of conflict probe and traffic condition

The Friedman test indicated a significant difference in the test conditions, χ^2 (5)=10.69, p≤.05. 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.

DISCUSSION

The samples are small, and the evidence should be read as preliminary. But, taken overall, these preliminary 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 than a workload measure.

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 research contexts or limitations in the CARS. 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% 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.

Shortcomings in the implementation of the CARS may lie in the careful definition of the terminology 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 would have been valuable to have a "refresher" course in how to make CARS ratings mid-way through the testing period to address questions and concerns about how the scale was being used or interpreted.

PRELIMINARY CONCLUSIONS AND RECOMMENDATIONS

CARS was used to measure controller acceptance of two different controller tasks, for two different decision-support tools, pFAST and URET. It provided a simple

scalar measure indicating whether the experienced workload was compatible with operational use of the DST. In both cases, CARS was a simple measure to implement and use for data collection. However, before CARS can be used, it requires some significant investment on the part of the researcher to clearly define and train the users on how the CARS is structured, and how to interpret the information.

Further validation research should be conducted on the CARS to (1) examine the agreement among controller ratings with larger samples of controllers, (2) validate the relationship between CARS and the TLX scales over a wider range of operational conditions, designed to represent under-and over-load, and (3) validate the relationship between CARS and DST characteristics, such as unreliability.

REFERENCES

Brudnicki, D.J. and McFarland, A.L. (1997). User Request Evaluation Tool (URET) Conflict Probe Performance and Benefits Assessment, MP97W112, The MITRE Corporation, McLean, VA.

Campbell, D.T. and Fiske, D.W. (1959). Convergent and Discriminant Validation by the Multitrait-Multimethod Matrix. Psychological Bulletin, 56, 81-105.

Davis, T. J., Robinson, J.E. III, Isaacson, D. R., den Braven, Wim, 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.

Deaton, J.E. and 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.

Harper, R.P., and Cooper, G.E. (1986). Handling Qualities and Pilot Evaluation. Journal of Guidance, 9(5), 515-529.

Hart, S.G. and 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.

Kerns, K. (in press). An Experimental Approach to Measuring the Effects of a Controller Conflict Probe in a Free Routing Environment. IEEE Transactions on Intelligent Transportation Systems.

Kirk, D.B., Heagy, W.S., McFarland, A.L., and 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.

Lee, K.K., and Davis, T.J. (1996). Development of the Final Approach Spacing Tool (FAST): A Cooperative Controller-Engineer Design Approach. Control Engineering Practice, 4(8).

Lee, K.K., and Sanford, B.D. (1998). Human Factors Assessment: The Passive Final Approach Spacing Tool (FAST) Operational Evaluation. NASA Technical Memorandum 208750.

McGraw, K.O. and S.P. Wong. (1996). Forming Inferences about Some Intraclass Correlation Coefficients. Psychological Methods, 1(1), 30-46.

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.

Parasuraman, R. and Riley, V. (1997) Humans and Automation: Use, Misuse, Disuse, and Abuse. Human Factors, 39(2), 230-253.

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, FLA