2007

Comparison of Three Subjective Workload Metrics for a Free Flight Enviroment

Jessica G. O'Connell

Shawn M. Doherty

Ian A. Wilson

COMPARISON OF THREE SUBJECTIVE WORKLOAD METRICS FOR A FREE FLIGHT ENVIROMENT

Jessica G O'Connell
Shawn M. Doherty
Ian A. Wilson
Embry-Riddle Aeronautical University
600 S. Clyde Morris Blvd
Daytona Beach, FL 32114

One recent change being debated in the aviation community is to implement techniques, similar to free-flight which would allow pilots rather than air traffic controllers to retain more control of their flight paths, speeds, and separate themselves from other flights. By implementing 'self-separation', one of the positions most significantly affected is that of the air traffic controller due to changes in their job tasking in a 'free flight' environment. As the safe capacity of the National Airspace System is limited by the controller workload, it is essential to measure the workload changes accurately and in a way that can be compared between experiments. This paper outlines a method to compare three indices used to measure controller workload. These measures include the Instantaneous Self Assessment method (ISA), the NASA Task Load Index (NASA-TLX) and the Subjective Workload Assessment Technique (SWAT). Workload measures from all three workload indices are measured and then compared by using the coefficient of variance. The workload scores are compared to assess similarities and discuss how these results can be used to create a common standard of comparison to other workload research using these metrics.

Introduction

Free Flight, now more often referred to as Airborne Separation Assurance, is a concept in which pilots are allowed to select their trajectory freely at real time, at the cost of acquiring responsibility for conflict prevention (PO-ASAS, 2001). It changes the responsibilities of air traffic management (ATM) in a fundamental way that it represents a paradigm shift for the national airspace. Instead of aircraft being controlled and deconflicted by ground control, decision making is distributed with pilots taking more or all responsibility for the separation of their aircraft. In a complete free-flight concept, responsibilities transfer from ground to air, air traffic control airspace structure and routes are removed and new technologies are brought in to assist the pilot in the new role. Free Flight can be characterized by the lack of a central control mechanism: conflicts between aircraft are not detected and solved by one dedicated controller. Instead, each individual aircrew has the responsibility to avoid conflicts; assisted by precise trajectory based navigation airborne surveillance systems, automation displaying conflict-solving trajectories and datalink negotiation between aircraft.

Real Time Simulation

Changes cannot be trialed live in the National Airspace System and even minor procedure changes are always the subject of real-time human-in-the-loop simulations before being used operationally. The impact of the change in roles that Airborne Separation Assurance could have is potentially extensive and therefore must be assessed as results are often non-intuitive. For example, simulations by Corker et al (1999) showed that controller workload could actually increase when aircraft self separate. Real time airspace simulations are complex and are often run separately by several civilian and military research agencies with nuanced changes in the concept of operations. The main metric from human-in-the-loop simulations is workload. The results of different simulations are compared to identify the best concepts and procedures. However, the metrics used may not lend themselves to this comparative approach.

Metrics

Mental Workload

In the field of human factors research, there appears to be many individual definitions of workload which at this time do not lead to a single common standard definition. Hart and Staveland (1988) proposed the assumption that workload is a hypothetical construct that represents the cost incurred by a human operator to achieve a particular level of performance. The importance of this assumption is that it is a heuristic placed on a multitude of characteristics. There are various measurements of mental workload which can be broken down into three main technique categories: subjective measurements such as self-report questionnaires such as the Instantaneous Self-Assessment (ISA) or NASA Task Load Index (NASA-TLX); performance measurements in which decrements in performance are recorded as workload volumes change; and physiological measurements

such as heart rate or eye blinks. Yet, only the first two are typically used for measuring mental workload; as mental workload efforts cannot be directly observed in declinations in performance in the same way as physical workload can be.

Workload Metric Validation

When a measurement technique is shown to accurately represent an individual's experienced or perceived level of overt or covert activity it has been validated. A cognitive measure is validated when a consensus is reached that the questions asked can accurately probe the person's perceptions of how busy they are. To achieve a consensus in the scientific community, results must be replicated by others in the human factors realm and those in the occupation being studied. A way to go about measuring a person's workload is to use both physiological and cognitive tests; yet, these methods generate disparate comparisons. For researchers looking at improving human performance and reevaluating levels of an individuals exerted mental workload, results that contradict each other and workload scores that rely on individual differences create a conundrum. The rift in workload score results has created a divide in workload theorists. Without human factors researchers having a commonly agreed metric for measuring workload, tests are created that seem to measure the same phenomenon associated with mental workload; yet, the results which do not have a common metric are not able to be statistically compared with confidence. Thus it is of immense importance to conduct research that would examine the commonly accepted and often used mental workload techniques; presenting statistical ways to compare them. The focus of this study is on three subjective measurement tools that are used in the aviation environment: The ISA, NASA-TLX, and SWAT techniques.

Subjective Workload Measurements

ISA. One measure of subjective workload is the Instantaneous Self Assessment (ISA) tool. The Instantaneous Self Assessment is a unidimensional subjective workload measure that can be administered during a simulation. This unidimensional measurement of the controller's workload is based on a single response scale to which numbers are assigned to statements inquiring on the person's belief about how much workload they are experiencing at the present time. The ISA was developed by Jordon (1992) as a tool that a researcher or ATM supervisor could use to estimate perceived workload during real-time simulations. The operator is prompted at regular intervals to give a rating of 1 to 5 of how busy they are (1 means under-utilized, 5 means excessively busy). This data can be used to compare operators' perceived workload, for example, with and without a particular tool, or between different systems. Two studies that have used the ISA in an ATM environment have been conducted by Eurocontrol (1997) and Whittaker (1995). It is also noted that ATM in Eurocontrol (1996) have used this technique in active work environments since the middle 1990's; which helps establish this as a European metric favored for ATC.

ISA, as a real-time measure, may enable the identification of events or tasks that may have contributed or caused the perceived high levels of workload. One method for administering this metric involves embedding the ISA into the ATM workstation to help maintain the flow of the scenario as much as possible. The user interface for the ISA is a box in the upper right hand corner of the computer screen, with a Likert scale of one thru five specifying perceived workload. The workload screen can be presented at any interval, such as every 90 seconds. The participant is asked to respond to the questionnaire as soon as possible answering with the present conditions in mind. This technique relies upon subjective evaluation of the workload situation by the controller. The researcher must consider factors such as experience, training, individual differences, group polarization, pride of the controller and other factors when evaluating the results. This measure has the advantage of assessing a person's workload periodically during the actual simulation compared to other workload measures, which can lead to a more accurate picture of the true perceived workload over time. There are a few disadvantages to this test; the biggest is connected to the overall use of a Likert Scale. Likert scales are created with subjective intervals, which force the respondent to choose an answer to what is closest, although not necessarily the most accurate.

NASA-TLX. The NASA-Task Load Index (NASA-TLX) was developed by NASA scientists Hart and Staveland (1988) to study human factors issues in workload. The NASA-TLX is a multi-dimensional rating scale for operators to report their mental workload. Two studies that have directly used the NASA-TLX in an ATC environment are: Brookings, Wilson & Swain (1986) and Hooijer & Hilburn (1996). The test is categorized as a multidimensional subjective workload questionnaire because it divides perceived workload into several factors. It uses six dimensions of workload to provide diagnostic information about the nature and relative contribution

of each dimension in influencing overall operator workload. Operators rate the contribution made by each of six dimensions of workload to identify the intensity of the perceived workload. It provides an overall workload score based on a weighted average of ratings on six sub-scales; 'Mental demands', 'Physical demands', 'Temporal demands', 'Performance', 'Effort' and 'Frustration'. The participant is asked to first give a score to each of the factors, and then a series of comparison questions are asked, responses are given on a continuum sectioned scale. This questionnaire is administered at the end of the tasks and the participant is asked to rate their overall workload experience. Although mental demand is one of the six workload scales and is described as mental and perceptual activity, thinking, deciding, calculating, remembering, looking, searching and task complexity, the technique does not consider the activities separately which are a large representative sum of ATC tasks. The NASA-TLX is not presented during a simulation but after a trial run, which can lead to several kinds of human time based interpretational errors on past activities. The NASA-TLX is however, used extensively in the aviation industry allowing a researcher access for comparison to many similar experiments.

SWAT. SWAT is a subjective scale of workload developed by Reid and Nygren (1988) that can be administered easily in operational situations. It is a multi-dimensional tool incorporating factors of temporal load, mental effort and psychological stress, similar to the NASA-TLX; although the SWAT has scaled the results into three areas rather than the six used in the NASA-TLX. The SWAT has two stages: the participant ranks the levels of the three workload scales in order from the lowest to highest workload prior to the trial and rates each of the scales after the trial. It was originally designed to assess aircraft cockpit and other crew-station environments to assess the workload associated with the operators' activities.

This measurement tool is more time consuming to administer than either the ISA or the NASA-TLX. The participant is asked to sort thru and rank 27 cards with various levels of workload statements before the experiment to adjust for individual differences and familiarize them with the descriptors that are used to describe mental workload. Following the simulation the participant is asked a series of questions with the focus of their overall workload experience in three areas: time load (the total amount of time available to accomplish a task as well as overlap of tasks), mental effort load (the amount of attention or concentration needed to perform a task), and psychological stress load (the presence of confusion, frustration, or anxiety during task performance). It has been noted by Nygren (1991) that the test does not capture low or vigilance workload task performance possibly due to the limited number of assessment levels. Yet, adding more levels to the SWAT would increase the time and complexity in administering of the test, for this reason Luximon and Goonetilleke, (2001) have presented an adaptation to the original SWAT process to shorten the time it takes to administer the test and to capture low-end workload performance. The SWAT, similar to the NASA-TLX is not presented during a simulation but after a trail run, which can lead to several kinds of human time based interpretational errors on past activities.

Method

There are multiple points of overlap between the three metric tools that can help identify their ability to measure workload. For example, one important difference between the three proposed metrics is that the ISA is used to gather real-time data and the other two collect post experimental data. Kirwan et al.(1997) has stated that information on a controller's mental workload collected during the task presents a more accurate picture of mental workload due to the ISA's application in real-time ATC simulation. The above reasoning suggests that the ISA will reflect greater changes to controller workload than the NASA-TLX or SWAT techniques.

A second set of common factors relates to the difference in sensitivity of the three workload techniques in measuring ranges of mental workload on a scale between low vigilance and high overload states. The ISA would be the most sensitive to mental effort changes by the participant Hart and Staveland (1988) propose that the NASA-TLX can measure a wider range of mental workload levels than the SWAT, which they conclude does not capture vigilance states accurately. Yet, the researchers that created the NASA-TLX believe that the two tests, the NASA-TLX and SWAT are fairly equal in measuring overall mental workload in simulation experiments (Hart & Staveland, 1988). Therefore it is expected that the NASA-TLX and SWAT techniques will result in identification of similar levels of workload from both tools. Brookhuis and de Waard (2001) in their review of measurement tools they found the SWAT and the NASA-TLX were the most commonly used self report indices of mental workload. In a complex work environment such as in air traffic control, many tasks are completed to accomplish the main goal. If as a researcher one is trying to relieve some percentage of an individual's mental workload

examination of the overall goal is too broad to extract the necessary information to make the small changes. Perhaps a more precise measurement such as the ISA which can help to narrow which sets of time, and tasks appears to be of more value. The implied narrowing ability of the ISA versus the NASA-TLX or SWAT may have led the researchers in earlier ATC experiments to choose it as the study metric.

Coefficient of Variance

The conundrum created by these multiple metrics, ISA, NASA-TLX, SWAT, is that none of them have the same measurement scale and there is not solid research showing how to statistically equate these different metrics. The NASA-TLX and SWAT analyses return a single numerical value, although that value is not directly equitable due to different interval scales. The ISA returns values represented with time and thus does not have a final numerical value at the end of the analysis.

One statistical measure that can compare all three human factors metrics is the coefficient of variance. The measure of relative variation is defined by Ostle (1975) as CV=s/x; where s is the sample standard deviation and x is the sample mean, further in percentage form, 100CV=100(s/x) percentage. As noted by Freund (1988) that one disadvantage of using standard deviation alone as a measure of variation is that it depends on the units of measurement being similar; therefore a measure of relative variation needs to be used such as the coefficient of variation. Ostle (1975) concurs, stating to afford a valid comparison of the variation among large values and the variation among small values, the coefficient of variation is an ideal device for comparing the variation in two series of data that are measured in two different units. This metric reduces the variability found in all three measures to a common comparison, providing a means to compare the relative sensitivity of each to capture workload against the others.

Discussion

A current study revisits the two prior experiments, Corker, Fleming, and Lane (1999) and Wilson and Fleming (2002) to further investigate the results of the subjective human factors measures in which to understand the sensitivity of the three measures for measuring workload in controllers. In these past studies the ISA measure was used. This measure provides useful workload information because the workload question is asked in time sequence within the simulation. The ISA score is however not used very often in human factors experiments and is therefore not widely accepted. The planned experiment will be using the ISA and comparing workload scores to the better known measures of the NASA-TLX and SWAT. Both the NASA-TLX and SWAT are military developed and used in the aviation industry at large, unlike the ISA. In the development for the future framework of the NAS many military and some nonmilitary researchers will be investigating ideas and a common metric or set of metrics that would validate the collection of their works in the air traffic control environment, specifically free flight. Thus with the high level of military investments the NASA-TLX and SWAT will be used more often and referred to more often as a valid metric, and yet, this study will investigate if these are the best metrics for the challenging environment.

References

Brookhuis, K.A., & de Waard, D. (2001). Assessment of driver's workload: Performance and subjective and physiological indexes. In P.A. Hancock, & P.A. Desmond (Eds.), *Stress, workload, and fatigue*. Mahwah, NJ: L. Erlbaum.

Brookings, J.B., Wilson, G.F., & Swain, C.R. (1986). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology, 42*(3), 361–377.

Corker, K., Fleming, K., and Lane, J., "Measuring Controller Reactions to Free Flight in a Complex Transition Sector," *Journal of ATC*, pp. 9-16, Oct-Dec., 1999.

Eurocontrol. (1996). *ERGO V2. for Instantaneous self assessment of workload in a real-time ATC simulation environment.* France

Eurocontrol. (1997). *PD/1 final report annex A: Experimental design and methods.* (Report PHARE/NATS/PD1-10.2/SSR;1.1). Brussels, Belgium: Author.

Freund, J. E. (1988). *Modern Elementary Statistics.* P. 78. Englewood Cliffs, New Jersey. Prentice-Hall.

Hart, S.G., Staveland, L.E. (1988). Development of a NASA TLX (NASA Task Load Index): Results of empirical and theoretical research; In P.A. Hancock and N. Meshkati (eds), Human Mental Workload, Amsterdam.

Hooijer, J.S., & Hilburn, B.G. (1996). *Evaluation of a label oriented HMI for tactical datalink communication in ATC.* (NLR Tech. Pub. TP 96676 L). Amsterdam: National Aerospace Laboratory NLR.

Jordan, C.S. (1992**)** Experimental study of the effects of an instantaneous self assessment workload recorder on task performance, Defense Evaluation &

Research Agency, Report No. DRA/TM(CAD5)/92011.

Kirwan, B., Evans, A., Donohoe, L., Kilner, A., Lamoureux, Atkinson, T., and MacKendrick, H. (1997). 'Human Factors in the ATM System Design Life Cycle', FAA/Eurocontrol ATM R&D Seminar, Paris, France. Retrieved on September 19, 2006 from http://atm-seminar-97.eurocontrol.fr/.

Luximon, A., Goonetilleke, R.S. (2001). Simplified Subjective Workload Assessment Technique. *Journal of Ergonomics*, 2001, Vol. 44, No. 3, pp. 229-243. Taylor & Francis, Ltd.

Nygren, T.E. 1991. Psychometric properties of subjective workload measurement techniques: Implications for their use in the assessment of perceived mental workload, *Human Factors*, 33, 17-33.

Ostle, B. (1975). *Statistics in Research*. P. 67-68. Ames, Iowa. Iowa State University Press.

Principles of Operations for the Use of ASAS, Action Plan 1 FAA/Eurocontrol Cooperative R&D, Version 7.1, June 15, 2001.Paris, France. Retrieved on August 20, 2006 from http://adsb.tc.faa.gov/RFG/po-asas71.pdf

Reid G.B. and Nygren, T.E. (1988). The subjective workload assessment technique: a scaling procedure for measuring mental workload, in: P.S. Hancock and N. Meshkati (Eds), Human Mental Workload.

Whittaker, R.A. (1995). *Computer assistance for en-route (CAER) trials programme: Future system 1 summary report*. (CS Rep. 9529). London: Civil Aviation Authority National Air Traffic Services.

Wilson, I. A., & Fleming, K. (2002). Controller Reactions to Free Flight in A Complex Transition Sector Re-Visited using ADS-B+. Daytona Beach, Florida: Embry-Riddle Aeronautical University, Center For Applied Air Traffic Control.