Federal Aviation Administration

DOT/FAA/AM-21/29
Aviation Safety
Office of Aerospace Medicine
Washington, DC 20591

Exploring the Relationship between Flight Technical Error and NASA-TLX Subscale Ratings when Using HUD Localizer Takeoff Guidance in Lieu of Currently Required Infrastructure

Daniela Kratchounova [1], Inchul Choi[2], Theodore C. Mofle[2], Larry Miller[3], Scott Stevenson[3], and Mark Humphreys[2]

[1] FAA Civil Aerospace Medical Institute (CAMI)
6500 South MacArthur
Oklahoma City, OK 73169

[2] Cherokee Support, Services, & Solution, LLC
6500 S. MacArthur Blvd.
CAMI Building 13
P.O Box 25082
Oklahoma City, OK 73125

[3] FAA Flight Technologies & Procedures Division, Flight Research & Analysis Group
6500 South MacArthur
Oklahoma City, OK 73169

November 2021

NOTICE

Acknowledgment

This project was funded by the FAA NextGen Human Factors Division (ANG-C1) in support of the FAA Office of Aviation Safety, Low Visibility Operations Units, Flight Technologies & Procedures Division, Flight operations Group (AFS-410).

1. Report No.	2. Government Accession No.	3. Recipient's Catalog No.
DOT/FAA/AAM-21/29		

4. Title and Subtitle	5. Report Date
Exploring the Relationship between Flight Technical Error and NASA-TLX Subscale Ratings when Using HUD Localizer Takeoff Guidance in Lieu of Currently Required Infrastructure	October 26, 2021
	6. Performing Organization Code AAM-500

7. Author(s)	8. Performing Organization Report No.
Kratchounova, D[1]. , Choi, I.[2] , Mofle, T.[2] , Miller, L.[3] , Stevenson, S.[3] , Humphreys, M.[2]	

9. Performing Organization Name and Address	10. Work Unit No. (TRAIS)
1 FAA Civil Aerospace Medical Institute (CAMI) 6500 South MacArthur Oklahoma City, OK 73169 2 Cherokee Support, Services, & Solution, LLC 6500 S. MacArthur Blvd. CAMI Building 13 P.O Box 25082 Oklahoma City, OK 73125 3 FAA Flight Technologies & Procedures Division, Flight Research & Analysis Group 6500 South MacArthur Oklahoma City, OK 73169	11. Contract or Grant No.

12. Sponsoring Agency Name and Address	13. Type of Report and Period Covered
Office of Aerospace Medicine Federal Aviation Administration 800 Independence Ave., S.W. Washington, DC 20591	14. Sponsoring Agency Code

15. Supplementary Notes

16. Abstract

This was an exploratory in nature follow-on research. Our previous efforts were focused on the operational impact of using head-up display (HUD) with localizer guidance in lieu of Centerline Lights (CLL) for takeoff in low visibility conditions. Herein, the primary goal was to examine the relationship between the subjective NASA-TLX workload ratings and Flight Technical Error (FTE) as the objective measure of performance. Instead of the total weighted scores, we analyzed the raw, unweighted NASA-TLX subscale ratings. Based on the analyses conducted, we proposed methods to utilize FTE data in predicting individual pilot workload ratings and vice versa. The results indicated that the single best subjective predictor of FTE was the NASA-TLX Performance subscale in both normal and abnormal operations. Nonetheless, the most noteworthy finding was that when the abnormal condition included a trained failure (e.g., engine failure during takeoff), the ratings on the NASA-TLX Temporal Demand subscale had an inverse relationship with FTE. That is, during trained abnormal events, the increased time pressure was associated with improved pilot performance. The immediate automatic response was to prioritize maintaining aircraft directional control and to compartmentalize tasks by priority. In contrast, when the pilots were presented with an abnormal event that they had not been trained on, this automatic response was absent. The discussion also includes methodological limitations and future research.

17. Key Words	18. Distribution Statement
Head-up Display (HUD); Aviation Human Factors; Flight Technical Error; Crew Workload; NASA-TLX	Document is available to the public through the Internet: (http://www.faa.gov/go/oamtechreports/)

19. Security Classification (of this report)	20. Security Classification (of this page)	21. No. of Pages	22. Price
Unclassified	Unclassified	26	

Table of Contents

Table of Figures

List of Tables

List of Equations

List of Abbreviations

CLL	Center Line Lights
FTE	Flight Technical Error
HUD	Head-up Display
KDE	Kernel Density Estimation
NASA	National Aeronautics and Space Administration
RVR	Runway Visual Range
TLX	Task Load Index

Introduction

This was a follow-on study to our previous research efforts (Kratchounova et al., 2020a; Kratchounova et al., 2020b) that were focused primarily on the operational impact of using head-up display (HUD) with localizer guidance in lieu of Center Line Lights (CLL) for takeoff in low visibility conditions. In those studies, we identified the differential effects of guidance type, runway visual range (RVR), lighting conditions (day/night) and runway lighting infrastructure on crew workload and performance, as measured by the NASA Task Load Index (NASA-TLX) and Flight Technical Error (FTE), respectively. Only the total weighted NASA-TLX scores were used for those analyses.

Here, we focused on the relationship between the objective measure of performance (i.e., FTE) and the subjective assessment of workload (i.e., NASA-TLX). However, instead of the total weighted scores, we analyzed the raw, unweighted NASA-TLX subscale ratings. Based on these analyses, we propose methods to utilize FTE data in predicting individual pilot workload ratings and vice versa.

Background

In the last few decades, the evolution of workload research has advanced from trying to measure it, through trying to define it, to applying it to relevant real-world settings (Young et al., 2015). While the scientific definition of workload is still passionately debated, it is a commonly recognized notion in the literature that workload is a multidimensional construct defined by the task demands, the capacity of the operator performing the task, and the context in which the performance occurs (Hancock et al., 1995; McKendrick & Cherry, 2018; Young et al., 2015).

Hart (2006) underscored the concept of workload as "the human cost (e.g., fatigue, stress, illness, and accidents) of maintaining performance" (p. 904). When that cost is too high, the capacity of a human operator to perform a given task safely, efficiently and effectively, may be depleted. In that context, workload only becomes evident in the interaction between the operator and other components of a system during task performance. Therefore, examinations of workload attempt to measure this very interaction to determine where the operator is within a workload "envelope" (Lysaght et al., 1989).

At times, performing a difficult task well may be at a cost of an excessively high level of workload. Nonetheless, the cost of performing a less difficult task could likewise be very high

but this time - driven by boredom. Therefore, subjective workload ratings may or may not associate with measures of performance. Consequently, a concurrent examination of performance and workload adds an advantage in addressing the ongoing debate regarding the mapping between performance and workload in different task situations. Hancock (1996) pointed out that ‘‘if workload response always followed performance variation, then there would be little reason to collect such additional measures’’ (p. 1156). Yet, workload ratings and performance measures often exhibit a repeating pattern. That is, higher levels of subjective workload ratings are associated with poorer performance (Hart, 2006; Lysaght et al., 1989; Yeh & Wickens, 1988). Thus, it is our strong conviction that pairing subjective workload ratings (e.g., NASA-TLX) with objective measures of performance (e.g., FTE) is a prudent approach to use in the context of empirical research.

Due to its extensive usage, the NASA-TLX has become almost synonymous with the concept of workload. It was developed based on the assumption that a combination of six workload related factors - Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration - represents the workload experienced by most people performing most tasks. These dimensions of the larger construct of workload were operationalized as the six subscales of the NASA-TLX. Furthermore, they were selected after extensive analyses of factors that identify the subjective experience of workload for different people performing activities ranging from simple to complex tasks such as flying an aircraft (Hart, 1986; Hart 2006; Hart & Staveland, 1988). A series of frequently cited studies comparing subjective workload measures (Hill et al., 1992; Lysaght et al., 1989) reported that NASA-TLX ratings were more closely related to performance, had the highest inter-rater reliability, the highest overall workload factor validity, and the best user acceptance.

Although immensely popular and well established, especially for empirical research, the NASA-TLX has not been without criticism regarding its construct validity. Without engaging in this continuing debate, we thought it would be sensible to mention its critics’ claim. Specifically, instead of measuring perceived task workload, the NASA-TLX measures perceived task difficulty (Byers et al., 1989; de Winter, 2014; McKendrick, & Cherry, 2018).

Analyzing the subscale scores individually rather than a single overall workload score has been one of the two most common modifications of the NASA-TLX workload scale. According to Hart (2006), over 40 studies conducted subscale-rating analyses instead of generating a single

overall workload score and demonstrated the potency of the scale and the diagnostic value of the component subscales. The author concluded that the high reliability, sensitivity, and utility of the NASA-TLX component ratings allow for a very narrow identification of sources of a workload or performance issues. Hart (2006) also recognized one potential drawback of the NASA-TLX scale. Namely, because the subscale ratings are measuring different aspects of the same underlying construct, they may often be significantly correlated (Hart, 2006).

Method

Twenty-four pilot crews participated in this research: 12 airline crews and 12 business jet crews, who were deemed proficient in using a HUD. For normal operational scenarios, three independent variables were considered including three levels of RVR (300ft, 500ft, and 700ft), two levels of Lighting conditions (Day and Night), and five levels of Type of Guidance as follows:

- HUD; No Localizer guidance; Centerline markings only;
- HUD; No Localizer guidance; Centerline marking and lighting;
- No HUD; Centerline marking and lighting;
- HUD with Localizer guidance and centerline markings;
- HUD with Localizer guidance and no centerline markings or lighting.

The full research matrix for normal operation is shown in Table 1. For the six abnormal operations (failure conditions), three levels of RVR (300ft, 500ft, and 700ft), and two levels of Lighting conditions (Day and Night) were examined (Table 2). The study was conducted in the Federal Aviation Administration's Boeing 737-800NG Level D simulator, equipped with a Rockwell Collins Head-up Guidance System Model 6700. There were 60 normal takeoff scenarios and 36 abnormal takeoff scenarios per crew. In the normal operations scenarios, winds speeds ranged between 3kt (calm) and 22ktm from various directions. In the abnormal operations scenarios, winds ranged between 3kt (calm) and 15kt from various directions. All tailwinds were limited to 10kt (Boeing, 2017).

Table 1.

Research matrix for normal operations.

Operation Type / Condition	Lighting Conditions		
Baseline 1: HUD, no LOC guidance*, Centerline markings (RCLM) only	Day/ Night	Day/ Night	Day/ Night
Baseline 2: HUD, no LOC guidance, Centerline lighting (CLL)**	Day/ Night	Day/ Night	Day/ Night
Baseline 3: No HUD, with Centerline lighting	Day/ Night	Day/ Night	Day/ Night
Condition 1: HUD, with LOC guidance, RCLM only	Day/ Night	Day/ Night	Day/ Night
Condition 2: HUD with LOC guidance; no RCLM, no CLL	Day/ Night	Day/ Night	Day/ Night

* Localizer

** All CLL conditions assume existing RCLM

Table 2.

Research matrix for abnormal operations.

Operation Type / Condition	Lighting Conditions		
Failure Condition 1: Engine fail above V_1	Day/ Night	Day/ Night	Day/ Night
Failure Condition 2 : Engine fail below V_1	Day/ Night	Day/ Night	Day/ Night
Failure Condition 3: Engine fail below V_{mcg}	Day/ Night	Day/ Night	Day/ Night
Failure Condition 4: LOC fail	Day/ Night	Day/ Night	Day/ Night
Failure Condition 5: LOC Bend	Day/ Night	Day/ Night	Day/ Night
Failure Condition 6: Loss of HUD	Day/ Night	Day/ Night	Day/ Night

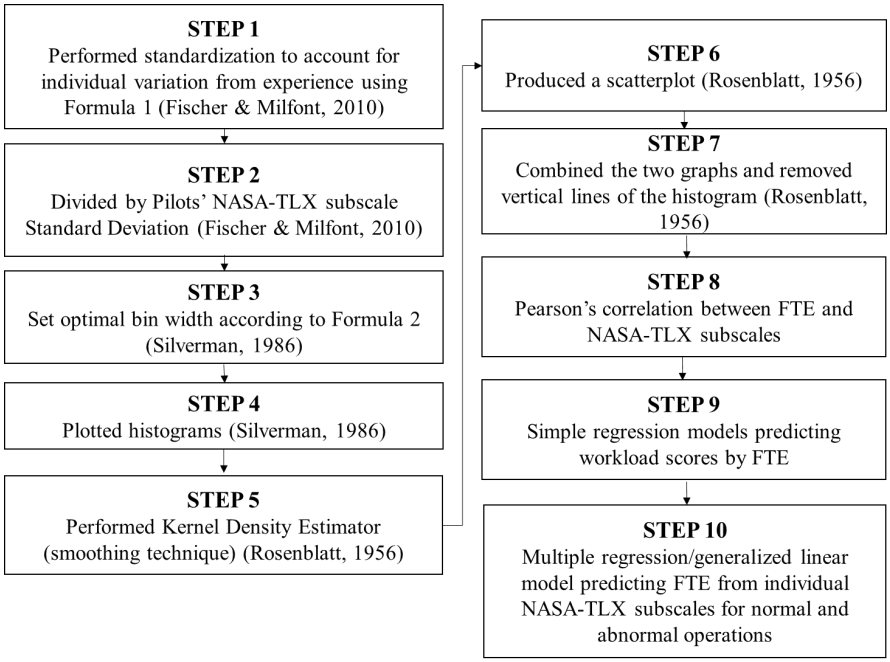
Notionally, NASA-TLX workload ratings may differ based on previous flight experiences. For example, some pilots may have previously experienced a situation (e.g., an emergency) that led to an extremely high workload level. Others may lack such experience. Consequently, reported workload levels may be higher for those pilots who never experienced a similar situation before

and relatively lower for those who did. To minimize the variance of workload ratings due to differences in experience between pilots, we applied a standardization process (Fischer & Milfont, 2010). The raw NASA-TLX subscale ratings were standardized for each individual pilot by subtracting the individual pilots' average NASA-TLX subscale score from the raw rating then dividing by the individual standard deviation as shown in (1). Figure 1 outlines all steps we followed analyzing the data for this study.

$$\text{Standardized NASA TLX score} = \frac{\text{TLX}_{\text{raw}} - Average\ TLX_{participant}}{TLX\ standard\ deviation_{participant}} \quad (1)$$

Figure 1.

Procedural steps for entire analytical methods.



Correlation Analysis, Kernel Density Estimation and Simple Regression Analysis (workload ~ FTE)

To ensure the raw FTE scores and the standardized NASA-TLX subscales were related, a correlation analysis using Pearson's R was conducted first. Before creating regression models, data density estimation was conducted of each variable's distribution. Building histograms was the starting point (Silverman, 2018). To examine the distributions for the raw FTE and the standardized NASA-TLX scores, two separate histogram graphs were constructed – one for

normal and one for abnormal operations. Setting the graphical origin and scale of the x-axis is essential for a meaningful graphical representation of the histogram and distribution. More specifically, dividing the x-axis into equal sections called bins, allows for an accurate graphical representation. For example, if the bins were set very narrow, the histogram would appear to have many artificial gaps. Conversely, if the bins are very wide, the distribution may not be visible and may even appear as a single density column. Therefore, examining the distributions created by the separate histograms helps with the selection of appropriate bin width (i.e., sub-intervals) (Silverman, 2018). Bin width selection dictates the amount of required smoothing for density estimation analysis. The goal of smoothing the data is to retain the original trends of the distributions while removing the noise around the bin width sectors. If the data is fundamentally Gaussian in nature (i.e., follows a normal distribution), then the width of these bins can be described by (2) (Silverman, 2018).

$$h = \left(\frac{4\sigma^5}{3n}\right)^{-\frac{1}{5}} \quad 2$$

Both the normal and abnormal operation distributions of the standardized NASA-TLX scores and the raw FTE data were set with 100 bins across the range of the distributions.
After standardizing the NASA-TLX subscale scores and the preliminary data exploration described above; Kernel density estimation (KDE) - a data smoothing technique - was implemented. This method of smoothing was proposed by Rosenblatt (1956) for estimating stochastic variables. To estimate standardized individual NASA-TLX subscale scores as a function of FTE, multiple simple regression analyses were conducted as described by (3).

$$Standardized\ NASA_{TLX}Subcale\ =\ \beta_0 + \beta_1 * FTE \quad 3$$

Multiple Regression (FTE ~ workload) for Normal Operations

For the purpose of methodological consistency, a model was fit independently to the distributions of normal and abnormal operations. The first model was an attempt to predict FTE scores as a function of the standardized individual NASA-TLX subscales during normal operations. After several failed attempts to construct a regression model with the raw FTE scores and standardized NASA-TLX subscales, it was determined that a transformation of the FTE scores was necessary for the model residuals to meet the parametric assumption of normality. Thus, a natural log transformation was applied to the raw FTE scores. We also established a

criterion stating that a predictor had to increase the explained variance by at least 5% to be included in the model. Otherwise, the variable was excluded.

Multiple Regression (FTE ~ workload) for Abnormal Operations

In the abnormal operational scenarios, each failure condition was associated with various system failures occurring during takeoff. Three conditions included an engine failure. Failure Condition 1 involved an engine failure above V_1[1] with the expectation for a continued takeoff. Failure Conditions 2 and 3 included engine failures below V_{mcg}[2] or below V_1 with the expectation of a rejected takeoff. Failure Conditions 4, 5 and 6 included issues with the HUD display, such as localizer bending, localizer failing, or a complete loss of HUD.

Based on the fundamentally different nature of these two groups of failure conditions, it was anticipated that the type of failure could affect the model and function of the predictor variables. The model for abnormal operations included the individual NASA-TLX subscales and a dummy variable indicating whether the failure situation involved an engine failure. To address the non-normality of residuals discussed earlier, a Generalized Linear Model was applied for abnormal operations.

Results

Correlation between FTE and Workload

The results from the Pearson's correlation indicated all NASA-TLX subscale score coefficients were statistically significant ($p < 0.001$). The coefficients for normal and abnormal operations are listed in Table 2 and Table 3, respectively. The subscales with the highest correlation were Performance and Physical Demand for both normal and abnormal operations.

[1] V_1 – The speed beyond which the takeoff should no longer be rejected.

[2] V_{mcg} – Velocity of Minimum Control on Ground is the speed at which the aircraft will remain controllable in the event of an engine failure on ground (occurs before V_1).

Table 3.

NASA-TLX subscale Pearson's R values for normal operations.

NASA TLX	Pearson's R
Mental Demand	0.259
Physical Demand	0.323
Temporal Demand	0.258
Performance	0.393
Effort	0.251
Frustration	0.256
Total Weighted	0.342

Table 4.

NASA-TLX subscale Pearson's R values for abnormal operations.

NASA TLX	Pearson's R
Mental Demand	0.256
Physical Demand	0.333
Temporal Demand	0.202
Performance	0.373
Effort	0.276
Frustration	0.289
Total Weighted	0.350

KDE for FTE and Workload and Regression Analysis (Workload ~ FTE)

Two histograms were constructed to identify the distributions of FTE and standardized NASA-TLX subscale scores. Figure 2 and Figure 3 show the distribution scores for the Performance subscale for normal and abnormal operations, respectively. The regression coefficient for FTE for normal operations was 0.13 ($p < 0.001$). For abnormal operations, the coefficient for FTE was 0.05 ($p < 0.001$), indicating that Performance would increase 0.05 standard deviations for each foot increase in FTE.

Multiple simple linear regression models enabled a better understanding of how well FTE might be able to predict each individual NASA-TLX subscale. FTE was the predictor variable, while the individual NASA-TLX subscale scores were set as a response variable in the various models.

The results indicated that all coefficients were statistically significant, including intercept and slope. Performance had the highest estimated coefficient (0.13) suggesting that increasing FTE by one foot would increase the Performance scores by 0.13 standard deviations. Temporal Demand had the lowest estimated coefficient (0.08). As shown in Table 4 for normal operations and Table 5 for abnormal operations, Temporal Demand would only increase with 0.08 standard deviations for a one-foot increase in FTE. The subscale with the largest Pearson's R^2 value was Performance for both the normal and abnormal operations.

Figure 2.

Standardized NASA-TLX Performance Subscale Scores with FTE distribution for normal operations: A total of 100 bins were set with equal spaced points from the range of the FTE scores.

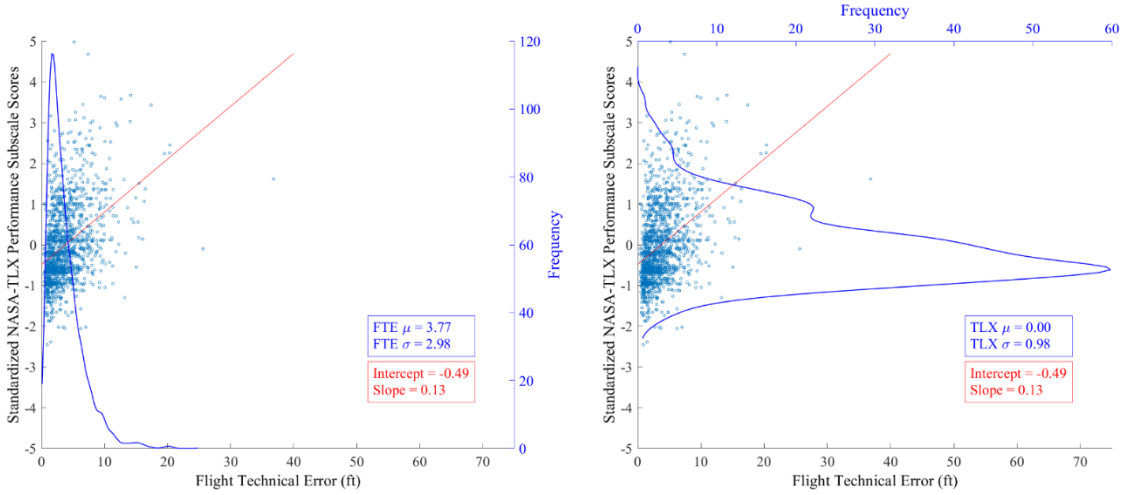


Figure 3.

Standardized NASA-TLX Performance Subscale Scores with TLX distribution for abnormal operations: A total of 100 bins were set with equal spaced points from the range of the TLX scores.

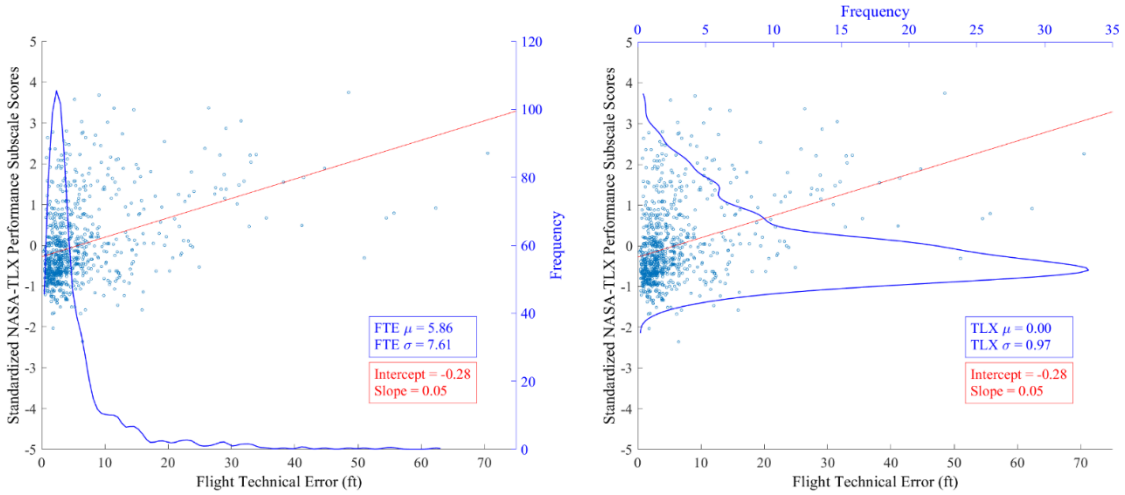


Table 5.

Linear regression results for each NASA-TLX subscale during normal operations.

Subscale	Parameter	Estimate	Stand. Error	*t*-stat.	*p*-value	Adj. R^2
Mental Demand	Intercept	-0.32	0.04	-7.99	<0.01	
	Coefficient	0.09	0.01	10.18	<0.01	0.07
Physical Demand	Intercept	-0.40	0.04	-10.15	<0.01	
	Coefficient	0.11	0.01	12.95	<0.01	0.10
Temporal Demand	Intercept	-0.32	0.04	-7.93	<0.01	
	Coefficient	0.08	0.01	10.11	<0.01	0.07
Performance	Intercept	-0.49	0.04	-12.72	<0.01	
	Coefficient	0.13	0.01	16.21	<0.01	0.15
Effort	Intercept	-0.31	0.04	-7.70	<0.01	
	Coefficient	0.08	0.01	9.82	<0.01	0.06
Frustration	Intercept	-0.32	0.04	-7.88	<0.01	
	Coefficient	0.08	0.01	10.05	<0.01	0.07
Total Weighted	Intercept	-0.43	0.04	-10.84	<0.01	
	Coefficient	0.11	0.01	13.82	<0.01	0.12

Table 6.

Linear regression results for each NASA-TLX subscale during abnormal operations.

Subscale	Parameter	Estimate	Stand. Error	*t*-stat.	*p*-value	Adj. R^2
Mental Demand	Intercept	-0.19	0.04	-4.74	<0.01	
	Coefficient	0.03	0.00	7.77	<0.01	0.06
Physical Demand	Intercept	-0.25	0.04	6.33	<0.01	
	Coefficient	0.04	0.00	10.38	<0.01	0.11
Temporal Demand	Intercept	-0.15	0.04	-3.70	<0.01	
	Coefficient	0.03	0.00	6.06	<0.01	0.04
Performance	Intercept	-0.28	0.04	-7.19	<0.01	
	Coefficient	0.05	0.00	11.72	<0.01	0.14
Effort	Intercept	-0.21	0.04	-5.14	<0.01	
	Coefficient	0.04	0.00	8.42	<0.01	0.08
Frustration	Intercept	-0.22	0.04	-5.41	<0.01	
	Coefficient	0.04	0.00	8.87	<0.01	0.08
Total Weighted	Intercept	-0.26	0.04	-6.69	<0.01	
	Coefficient	0.04	0.00	10.97	<0.01	0.12

Multiple Regression Analysis (FTE ~ Workload) for Normal Operations

To predict FTE as a function of individual NASA-TLX subscales, a stepwise multiple regression analysis was conducted. The residuals from the original model using raw FTE scores did not display attributes of normality and included signs of heteroscedasticity (i.e., unequal variance across the distribution). Therefore, a natural log transformation was applied to the raw FTE scores. The application of a natural log transformation corrected the positive skew and the model residuals met the assumptions of normality (Lilliefors normality test; $p = 0.13$).

The final model included the NASA-TLX Performance and Physical Demand subscales. The model was significant in predicting the transformed FTE scores ($F(2, 1437) = 184.05$, $p < 0.0001$) and explained about 20% of the observed variance, with a medium effect size (Adjusted $R^2 = 0.204$) (Cohen, 2013); (4) defines the final model for normal conditions and model parameters are outlined in Table 6.

$$FTE = 1.07 + .23(Performance) + .16\ (Physical\ Demand) \quad (4)$$

Table 7.

Linear regression model to predict natural log transformed FTE scores during normal operations.

Parameter	Unstandardized Coefficients	Standard Error	*t*-statistics	*p*-value
Intercept	1.07	.02	62.51	<0.001
Performance	0.23	.02	11.57	<0.001
Physical Demand	0.16	.02	8.25	<0.001

Multiple Regression Analysis (FTE ~ Workload) for Abnormal Operations

For abnormal operations, Generalized Linear Model was used to predict FTE as a function of individual NASA-TLX subscale scores and the type of failure condition. The final model included Mental Demand, Physical Demand, Performance, Temporal Demand, and a dichotomous variable - Engine Failure - indicating the type of abnormal condition (i.e., 1 = engine failure or 0 = HUD or other failure). An interaction effect was present between Temporal Demand and Engine Failure. As the Engine Failure variable was dummy coded, the interaction effect (Temporal Demand x Engine Failure) had a negative coefficient only in the engine failure conditions, otherwise the coefficient was nulled by a zero in the model, described by (5). The model was significant in predicting the transformed FTE scores ($F(5, 857) = 97.1$, $p < 0.0001$) and had a large effect size (Cohen, 2013), explaining about 40 percent of the observed variance (Adjusted $R^2 = 0.401$). Model parameters are provided in Table 7.

$$FTE = 0.40 + 0.04 * (Mental\ Demand) + 0.04 * (Physical\ Demand) + 0.10 * (Perfomrance) - 0.07 * (Temporal\ Demand \times Engine\ Failure) \quad (5)$$

Table 8.

Linear regression model to predict natural log transformed FTE scores during abnormal operations.

Parameter	Unstandardized Coefficients	Standard Error	*t*-statistics	*p*-value
Intercept	0.40	0.0.02	22.73	<0.001
Mental Demand	0.04	0.02	2.29	=0.022
Physical Demand	0.05	0.02	2.98	=0.003
Performance	0.10	0.01	8.46	<0.001
Temporal Demand x Engine Failure	-0.07	0.02	-3.19	=0.001

Note: The 'x' in the model and the table indicates an interaction effect

Discussion

The results of this research indicated that the single best subjective predictor of FTE is the NASA-TLX Performance subscale in both normal and abnormal operations. The Performance subscale's definition calls for the operator to answer the following questions: "How successful do you think you were in accomplishing the goals of the task set by the experimenter (or yourself)? How satisfied were you with your performance in accomplishing these goals?" (Hart, 1986). These questions could be précised as one single question: "How well do think you did?" Remarkably, the pilot evaluators' feedback during our original research (Kratchounova et al., 2020a; Kratchounova et al., 2020b) suggested that rating their own performance and its contribution to the overall workload levels was the easiest NASA-TLX subscale to conceptualize, observe, and assess. Furthermore, the theoretical underpinnings of the NASA-TLX Performance subscale are deeply rooted in the notion that lower subjective workload ratings accompanied better performance. As a whole, the results from the statistical analyses we conducted, exhibited that same familiar pattern between performance and workload. Namely, higher levels of subjective workload ratings were associated with inferior performance.

The Physical Demand subscale was also a significant predictor of FTE in both the multiple regression models and had the second highest correlation, next to the Performance subscale. This finding suggests that one potential method of optimizing pilots' workload profile under these conditions would be to conceive a way to reduce the task's physical demand by design.

Yet, the most noteworthy finding from this study was that when the abnormal condition included an engine failure, the ratings on the NASA-TLX Temporal Demand subscale had an inverse relationship with FTE. That is, during abnormal or emergency situations, the increased time pressure actually improved pilots' performance.

One plausible explanation for this finding would be that during a real abnormal or emergency situation in an aircraft, the mind seems to "accelerate" and time seems to slow down. Hancock and Weaver (2005) noted that under life-threatening stress, people often experience temporal distortion. To a pilot experiencing a serious emergency situation, what is actually 4 seconds may seem like 10 seconds. Furthermore, the authors also reported that although high-stress conditions consume part of the attentional resources, it is common for the remaining resources to be directed to specific task-related activities.

Moreover, the results of the research presented here, showed that the specific nature of the abnormal condition, and not merely the existence of an emergency, largely determined whether pilots' performance improved or declined under increased temporal demand. Specifically, when pilots were presented with an emergency that they have been highly trained on, such as an engine failure during takeoff, it was common for them to quickly recall and almost automatically perform the trained procedure. Such procedures are performed multiple times in a controlled environment (e.g., in a simulator or another training device), sometimes over the course of many years of a pilot's career. Therefore, when a trained emergency actually happens, the pilot goes into a methodical routine of dealing with the emergency in the calm and precise manner in which they were trained to respond. Similarly to the load-shedding during a partial electrical failure, a pilot load-sheds everything outside the specific emergency, essentially compartmentalizing certain tasks by priority. More, in the case of engine failure on takeoff, the trained immediate response was to prioritize maintaining aircraft directional control in alignment with the runway centerline, which happened to be the exact metric used to assess flight technical error in this research.

In contrast, when pilots were presented with an abnormal condition that they have *not* been trained on, that automatic response was absent. In this study, localizer failures and HUD failures represented this distinct group of abnormal conditions. Therefore, when a non-trained situation occurred, pilots frequently tried to focus most of their attention on analyzing the situation to a point where no attentional resources were left to deal with basic tasks such as "flying the

airplane". While such response is not a certainty, it has been observed in numerous aircraft accidents. The recent Boeing 737 MAX accidents may be one example of the overwhelming confusion occurring with a certain abnormal event that the pilots were not trained to respond to. Consequently, the pilots were unable to transfer their focus back to controlling the aircraft since the abnormal situation could not be resolved in the time leading up to the accident. Another example is the Air France 447 accident (Anuary et al., 2010; ECAA - Ethiopia Aircraft Accident Investigation Bureau, 2019).

The primary goal of the current study was exploratory-in-nature. Specifically, the focus was on the relationship between the subjective NASA-TLX workload subscale ratings and FTE as an objective measure of performance. The correlation analysis provided evidence that each NASA-TLX subscale had a significant relationship with FTE. This granted further support for the notion that the NASA-TLX is a proper workload measure in applied human factors and psychological research settings, regardless of the ongoing debate within the research community about the theoretical implication of the measure.

Limitations and Future Research

The first notable limitation for this study was that the simple regression models applied in the analyses used FTE to predict the *standardized* NASA-TLX subscale scores from the pilots' average and the standard deviation score of each NASA-TLX subscale. As a result, if this method were to be used to analyze new population sample datasets; both FTE and NASA-TLX measures would still need to be collected.

An additional limitation in the current research was applying a natural log transformation for the multiple regression models. While this transformation afforded a better fit to meet parametric assumptions, applying it to future datasets will produce predictions of FTE in the form of natural log. For meaningful results, the predicted scores will need to be transformed back to raw scores.

In our upcoming research efforts, we will continue utilizing FTE as an objective measure of performance and NASA-TLX as a subjective measure of crew workload. As more data are collected, we will continuously refine the models outlined in this paper for improved predictive power. Once these models are systematically validated through continued research, it may be possible to collect only one of the measures and predict the other with higher confidence.

Reference

Anuary, J., Russell, P., & Pardee, J. (2010). *Final Report On the Accident on 1st June 2009 to the Airbus A330-203 Registerd F-GZCP Operated by Air France Flight AF 447 Rio de Janeiro-Paris.* (Issue July).

Boeing. (2017). *Boeing 737-700 / 800 Flight Crew Operation Manual*. http://lukas1992.bplaced.net/737/FCOM-003-all-online.pdf

Byers, J. C., Bittner Jr, A. C., & Hill, S. G. (1989). Traditional and raw task load index (TLX) correlations: Are paired comparisons necessary. *Advances in Industrial Ergonomics and Safety*, *1*, 481–485.

Cohen, J. (2013). *Statistical power analysis for the behavioral sciences*. Academic press.

de Winter, J. C. (2014). Controversy in human factors constructs and the explosive use of the NASA-TLX: a measurement perspective. *Cognition, Technology & Work*, *16*(3), 289–297.

ECAA - Ethiopia Aircraft Accident Investigation Bureau. (2019). *Aircraft Accident Investigation Bureau Preliminary Report B737-8 (MAX) ET-AVJ* (Vol. 8).

Fischer, R., & L. Milfont, T. (2010). Standardization in psychological research. *International Journal of Psychological Research*, *3*(1), 88–96. https://doi.org/10.21500/20112084.852

Hancock, P. A., Williams, G., & Manning, C. M. (1995). Influence of task demand characteristics on workload and performance. *The International Journal of Aviation Psychology1*, *5*(1), 63–86.

Hancock, P. A. (1996). Effects of control order, augmented feedback, input device and practice on tracking performance and perceived workload. *Ergonomics*, *39*(9), 1146–1162. https://doi.org/10.1080/00140139608964535

Hancock, P. A., & Weaver, J. L. (2005). On time distortion under stress. *Theoretical Issues in Ergonomics Science*, *6*(2), 193–211. https://doi.org/10.1080/14639220512331325747

Hart, S. G. (1986). *NASA Task load Index (TLX). Volume 1.0; Paper and pencil package*. NASA Ames Research Center.

Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. *Proceedings of the Human Factors and Ergonomics Society*, 904–908. https://doi.org/10.1177/154193120605000909

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results

of empirical and theoretical research. *Advances in Psychology*, *52*, 139–183. https://doi.org/https://doi.org/10.1016/S0166-4115(08)62386-9

Hill, S. G., Iavecchia, H. P., Byers, J. C., Bittner, A. C., Zaklad, A. L., & Christ, R. E. (1992). Comparison of four subjective workload rating scales. *Human Factors*, *34*(4), 429–439. https://doi.org/10.1177/001872089203400405

Kratchounova, D., Humphreys, M., Miller, L., Mofle, T., Choi, I., & Nesmith, B. L. (2020). Crew workload considerations in using hud localizer takeoff guidance in lieu of currently required infrastructure. In J. Y. C. Chen & G. Fragomeni (Eds.), *Virtual, Augmented and Mixed Reality. Design and Interaction: Vol. 12190 LNCS* (pp. 507–521). Springer International Publishing. https://doi.org/10.1007/978-3-030-49695-1_34

Kratchounova, D., Miller, L., Choi, I., Mofle, T., Humphreys, M., & Nesmith, B. L. (2020). Flight technical error in using head-up display with localizer guidance in lieu of required infrastructure for takeoff. *AIAA/IEEE Digital Avionics Systems Conference - Proceedings*, *2020-Octob*, 1–6. https://doi.org/10.1109/DASC50938.2020.9256606

Lysaght, R. J., Hill, S. G., Dick, a O., Plamondon, B. D., Linton, P. M., Wierwille, W. W., Zaklad, a L., Bittner Jr, a C., & Wherry, R. J. (1989). Operator workload: Comprehensive review and evaluation of operator workload methodologies. In *United States Army Research Institute for the Behavioral Sciences, Technical Report* (Vol. 851).

McKendrick, R. D., & Cherry, E. (2018). A deeper look at the NASA TLX and where it falls short. *Human Factors and Ergonomics Society Annual Meeting*, 44–48.

Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *Annals of Mathematical Statistics*, *27*(3), 832–837.

Silverman, B. W. (2018). Density estimation: For statistics and data analysis. *Density Estimation: For Statistics and Data Analysis*, *1986*, 1–175. https://doi.org/10.1201/9781315140919

Yeh, Y.-Y., & Wickens, C. D. (1988). Dissociation of Performance and Subjective Measures of Workload. *Human Factors*, *30*(1), 111–120. https://doi.org/https://doi.org/10.1177/001872088803000110

Young, M. S., Brookhuis, K. A., Wickens, C. D., & Hancock, P. A. (2015). State of science: mental workload in ergonomics. *Ergonomics*, *58*(1), 1–17.