USAARL-TECH-TR--2023-04

UNITED STATES ARMY AEROMEDICAL RESEARCH LABORATORY

A Literature Review of Applied Cognitive Workload Assessment in the Aviation Domain

Jon Vogl, Cristina Delgado-Howard, Hillary Plummer, Aaron McAtee, Amanda Hayes, Chris Aura, & Paul St. Onge

Notice

Qualified Requesters

Qualified requesters may obtain copies from the Defense Technical Information Center (DTIC), Fort Belvoir, Virginia 22060. Orders will be expedited if placed through the librarian or other person designated to request documents from DTIC.

Change of Address

Organizations receiving reports from the U.S. Army Aeromedical Research Laboratory on automatic mailing lists should confirm correct address when corresponding about laboratory reports.

Disposition

Destroy this document when it is no longer needed. Do not return it to the originator.

Disclaimer

The views, opinions, and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy, or decision, unless so designated by other official documentation. Citation of trade names in this report does not constitute an official Department of the Army endorsement or approval of the use of such commercial items.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
11-10-2022	Technical Report	MAR 2020 - OCT 2022

4. TITLE AND SUBTITLE
A Literature Review of Applied Cognitive Workload Assessment in the Aviation Domain

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Vogl, J.[1,2], Delgado-Howard, C.[1,3], Plummer, H.[1,3], McAtee, A.[1,2]
Hayes, A.[1], Aura, C.[1], & St. Onge, P.[1]

5d. PROJECT NUMBER
23460

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
U.S. Army Aeromedical Research Laboratory
P.O. Box 620577
Fort Rucker, AL 36362

8. PERFORMING ORGANIZATION REPORT NUMBER
USAARL-TECH-TR--2023-04

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Development Command
Military Operational Medicine Research Program
504 Scott Street
Fort Detrick, MD 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)
USAMRDC MOMRP

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release; distribution unlimited.

13. SUPPLEMENTARY NOTES
[1]U.S. Army Aeromedical Research Laboratory; [2]Goldbelt Frontier, LLC; [3]Oak Ridge Institute for Science and Education

14. ABSTRACT
Advancements in technology have pushed the standard unit of work from the joule to the byte, as operators are tasked with increasing cognitively demanding information processing in domains such as transportation, warfighting, and medicine. Due to these advancements, interest in the field of cognitive workload and its assessment in applied environments has grown exponentially over the last two decades. Operator state monitoring systems promise to use performance, physiological, and subjective cognitive workload assessment metrics to predict when operators are approaching or experiencing cognitive overload and the system will take remedial action. A systematic literature review was conducted to survey the last decade (2010-2020) of cognitive workload assessment literature in the aviation domain. The objective of the literature review was to identify cognitive workload assessment techniques that have seen success in the aviation domain and examine the usability of composite cognitive workload metrics in an operational use case. Articles were obtained from three databases using keywords that surveyed cognitive workload terminology, measures, and domains.

15. SUBJECT TERMS
cognitive workload, mental workload, performance, subjective, physiological, workload assessment, aviation

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Loraine St. Onge, PhD
UNCLAS	UNCLAS	UNCLAS	SAR	35	19b. TELEPHONE NUMBER *(Include area code)* 334-255-6906

REPORT DOCUMENTATION PAGE (SF298)
(Continuation Sheet)

14. Abstract (continued)
Articles were filtered down to the aviation domain, organized by level of subject experience in flight hours (novice, intermediate, and expert), and detailed by the types of cognitive workload assessment techniques (performance, physiology, and subjective) utilized in the study. Analysis was then conducted to determine associations, insensitivities, and dissociations of cognitive workload assessment techniques to assess the practicality of composite cognitive workload assessment metrics. Findings in the literature across the recent decade and the outlook of operator state monitoring system-aided adaptive automation in the aviation domain are discussed.

Summary

Operators in the domains of transportation, war fighting, and medicine are increasingly being tasked with more cognitively demanding information to process due to advancements in technology enabling additional operator functionality. Due to these advancements, over the past two decades, interest in the field of cognitive workload and its assessment in applied environments has grown exponentially. Operator state monitoring systems promise to use performance, physiological, and subjective cognitive workload assessment metrics to predict when operators are approaching or experiencing cognitive overload and the system on which they are working will take remedial action to avoid operator overload and mitigate potential mistakes or mishaps resulting from the operator's cognitive overload. A systematic literature review was conducted to survey the last decade (2010-2020) of cognitive workload assessment literature in the aviation domain. The objective of the literature review was to identify cognitive workload assessment techniques that have seen success in the aviation domain and examine the usability of composite cognitive workload metrics in an operational use case. Articles were obtained from three databases using keywords that surveyed cognitive workload terminology, measures, and domains. Articles were filtered down to the aviation domain, organized by level of subject experience in flight hours (novice, intermediate, and expert), and detailed by the types of cognitive workload assessment techniques (performance, physiological, and subjective) utilized in the study. Analysis was then conducted to determine associations, insensitivities, and dissociations of cognitive workload assessment techniques to assess the practicality of composite cognitive workload assessment metrics. Findings in the literature across the recent decade and the outlook of operator state monitoring system-aided adaptive automation in the aviation domain are discussed.

This page is intentionally blank.

Acknowledgements

This research was supported in part by an appointment to the Oak Ridge Institute for Science and Education Research Participation Program at the U.S. Army Aeromedical Research Laboratory administered by the Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the U.S. Army Medical Research and Development Command.

This page is intentionally blank.

Table of Contents

List of Figures

Table of Contents (continued)

List of Tables

Introduction

The next fleet of U.S. Army rotary-wing aircraft will allow for longer periods in combat in addition to hosting a suite of advanced technologies and weaponry. These aircraft will likely be the most advanced and complex system of systems the Army has in its arsenal. This means that these aircraft are likely to require pilots to multi-task on a level far above current helicopters. Because of the increasing demands placed on the pilots while flying these vehicles, the need to monitor pilot cognitive workload, health, and well-being in real-time has become integral to mission accomplishment. With real-time physiological monitoring, it may be possible to track and understand the degree of task cognitive demand and associated cognitive workload (CWL) placed upon the pilots throughout the various phases of the multi-domain operation (MDO) mission sets. These data may then inform leadership and team members, as well as provide critical feedback to the individual operators. These data will also inform key decision points for the cockpit layouts specific to human system interaction. However, much work remains, as unknowns exist regarding which measures are most effective at capturing and quantifying CWL, how best to deploy those sensors within the cockpit, and how to quantify the data such that the results can easily be interpreted in real-time to aid decision-making.

In order to support the expanded future vertical lift (FVL) mission, research is ongoing at the U.S. Army Aeromedical Research Laboratory (USAARL). The end-state goal is the implementation of physiological measures as a means to assess CWL in an operator state monitoring (OSM) driven adaptive automation environment. This report presents a systematic review of recent CWL literature to identify which CWL assessment techniques are seeing the most use and success within the aviation domain, with a specific focus on rotary-wing aviation. First, a formal definition of CWL is provided alongside evidence of growing interest in the construct of CWL. A brief summary of different CWL metrics follows along with considerations for the usage of multiple, i.e., composite metrics to assess CWL.

Defining Cognitive Workload

A consistently used formal definition of CWL has yet to be generally accepted by the research community at large. As such, it is often found that different definitions are used across researchers (Cain, 2007). For consistency, we adopted a resource-demand framework with a definition put forth by Van Acker et al.'s (2018) concept analysis:

"Mental workload is a subjectively experienced physiological processing state, revealing the interplay between one's limited and multidimensional cognitive resources and the cognitive work demands being exposed to."

To eliminate any points of confusion, notice that Van Acker et al. (2018) uses the term "mental workload" (MWL) versus our use of the term "cognitive workload" in this review. The body of literature concerning the assessment of cognitive resource expenditure due to cognitive work demands has utilized both terms (i.e., mental and cognitive) interchangeably (even sometimes used interchangeably within the same paper). Figure 1 details the use of each respective term over decades of research; the term "mental workload" appears earlier in the literature (Westbrook et al., 1966) and is used more often than "cognitive workload." The term "cognitive workload" has been adopted for the work being conducted at the USAARL.

The Van Acker et al. (2018) definition includes three key components (for a more extensive discussion on these points, refer to Vogl et al., 2020). First, CWL occurs due to the interaction of a specific human and a specific task/environment (or a task + environment combination). This interaction of applying cognitive resources to meet task demands leads to the perception of CWL. This sets the foundation for the resource-demand framework that has evolved since its initial presentation by Kahneman (1973) in his book *Attention and Effort*. Second, as cognitive resources are utilized for a task, it becomes evident to the effort-exerting human that their resources are limited, and should a task demand resources beyond the available limit, the human's performance will falter. To the introspective human, it is also observed that multiple task demands can be met more efficiently in some cases more so than others. Wickens (2008) clarified this perception through Multiple Resource Theory, which states that rather than having a single limited pool of resources available to address task demands, multi-tasking experiences are better explained by a model of multiple pools of resources. Third, Van Acker et al. (2018) state that CWL is a subjectively experienced physiological processing state; that is, humans understand and are able to communicate that they are experiencing CWL and their physiology changes as a function of CWL. As such, it is possible to assess CWL not only through performance measures on task(s) themselves, but through self-report measures (i.e., subjective measures) and by monitoring changes in physiological signals (i.e., physiological measures). Of final note, the Van Acker et al. (2018) definition serves well for brief presentations of the concept, but it is desired to have a more all-encompassing definition that distinctly highlights other facets of human experience (e.g., individual differences, situational factors, attention, etc.) and the dynamic of the CWL and performance relationship. For a more in-depth definition and reanalysis of the CWL concept, see Longo et al. (2022).

CWL has become an increasingly popular area of study since its first formal references in the 1960s. Over the last decade, CWL research has seen a surge of publications as indexed by the Google Scholar search engine (Figure 1). Both exact phrase matches throughout articles and in titles only follow the same accelerated growth pattern over the last ten years. This accelerating interest speaks to investigations of more advanced and efficient physiological metrics, modeling techniques, and general emphasis on improving performance in safety-critical domains such as aviation and driving. In 2015, Young et al. (2015) examined the CWL literature and identified prominent domains of study across the decades. In the 1980s, among the continuing development of major theoretical advances in CWL, domains such as software engineering/computer-aided design (CAD), and adaptive interfaces (i.e., automation that responds to the CWL of an operator) were of primary interest. Examinations of CWL continued most frequently in the aviation and driving domains in the 1990s. Ultimately, the field of driving would pull far ahead of any other domain of interest throughout the first decade of the 2000s, while research in the railway domain became of increasing interest and aviation and air traffic control (ATC) remained steady (Figure 5). Considering the domains of interest across decades, it is clear that CWL evaluation is a vital component to safety-critical domains, especially in the realm of transportation.

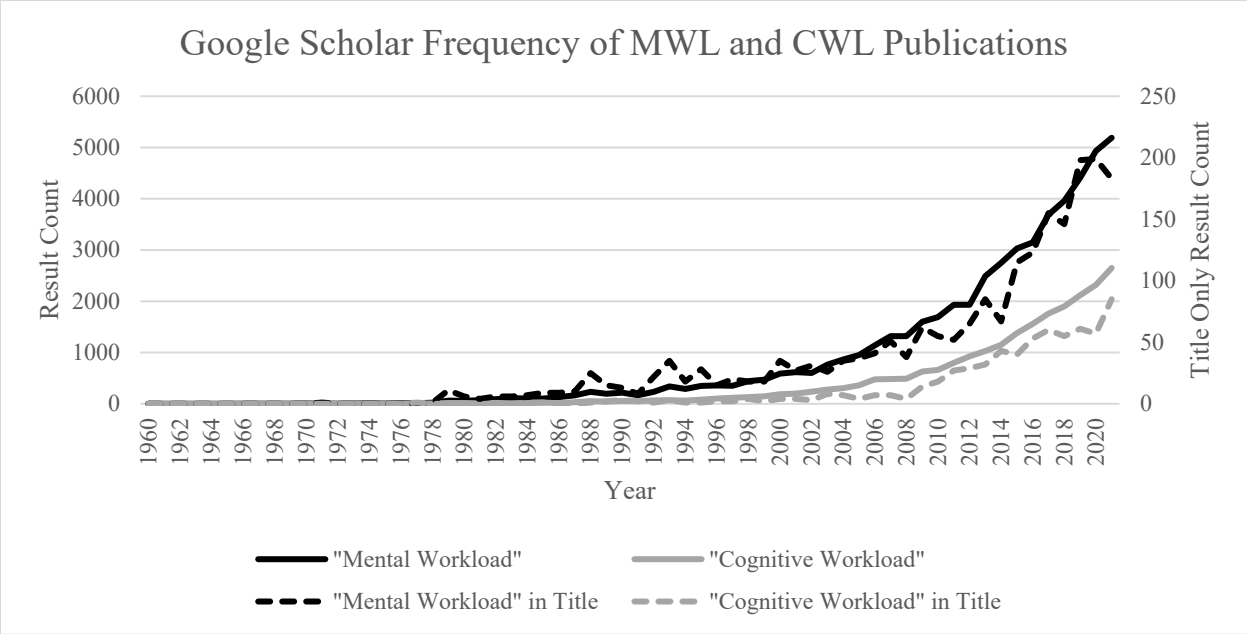

Figure 1. Frequency of mental workload and cognitive workload publications over 60 years. Frequency data obtained from the Google Scholar search engine.

Composite Cognitive Workload Assessment

As outlined in the presented definition of CWL, the concept of CWL is operationally quantifiable using measures of performance, physiology, and subjective appraisal. These categories of measurement have been persistently used throughout the CWL literature, and each category provides a few trade-offs across different evaluation criteria (O'Donnell & Eggemeier, 1986). A quick literature search shows over 20,000 examinations have been conducted on these measurement techniques (for reviews see Cain, 2007; Heard et al., 2018; Tao et al., 2019; Charles & Nixon, 2019; Vogl et al., 2020).

Performance and CWL are coupled in an inverse manner, especially when under optimal levels of task demand, but the relationship does not simply boil down to one rising and the other falling. Instead, through the voluntary recruitment of cognitive resources (i.e., effort leading to increased CWL), performance can remain at high levels while workload is increased. That is to say, humans can exert more effort, marshal more resources, or 'try harder' on a task as the demands increase to maintain their performance. Only at a certain point, traditionally referred to as the 'red line,' does performance begin to falter, resulting in an inverse relationship with high levels of CWL. Figure 2 details the performance-workload relationship as a function of increasing task demand (adapted from De Waard, 1996 and Young et al., 2015). This modified region model illustrates how performance and workload have a consistent inverse relationship in regions D, A2, and C, while more dynamic changes occur in regions A1, A2, and B. Using this model as a framework, it is easy to see that primary task performance measures may suffer from a lack of **sensitivity** unless looking within regions D or B. Additional shortcomings in performance measures are seen in **diagnosticity**, or in their ability to diagnose which set of cognitive resources are being affected by the task. As performance metrics are generally global surveys of task efficiency, they rarely provide a diagnostic picture of cognitive loading across resources. While these types of performance metrics may function at a relatively coarse scale,

they are minimally **intrusive** on overall task performance as the data is often readily available. In aviation, measures of criterion deviation, input activity, and instructor pilot ratings have seen extensive use in differentiating between high and low levels of CWL.

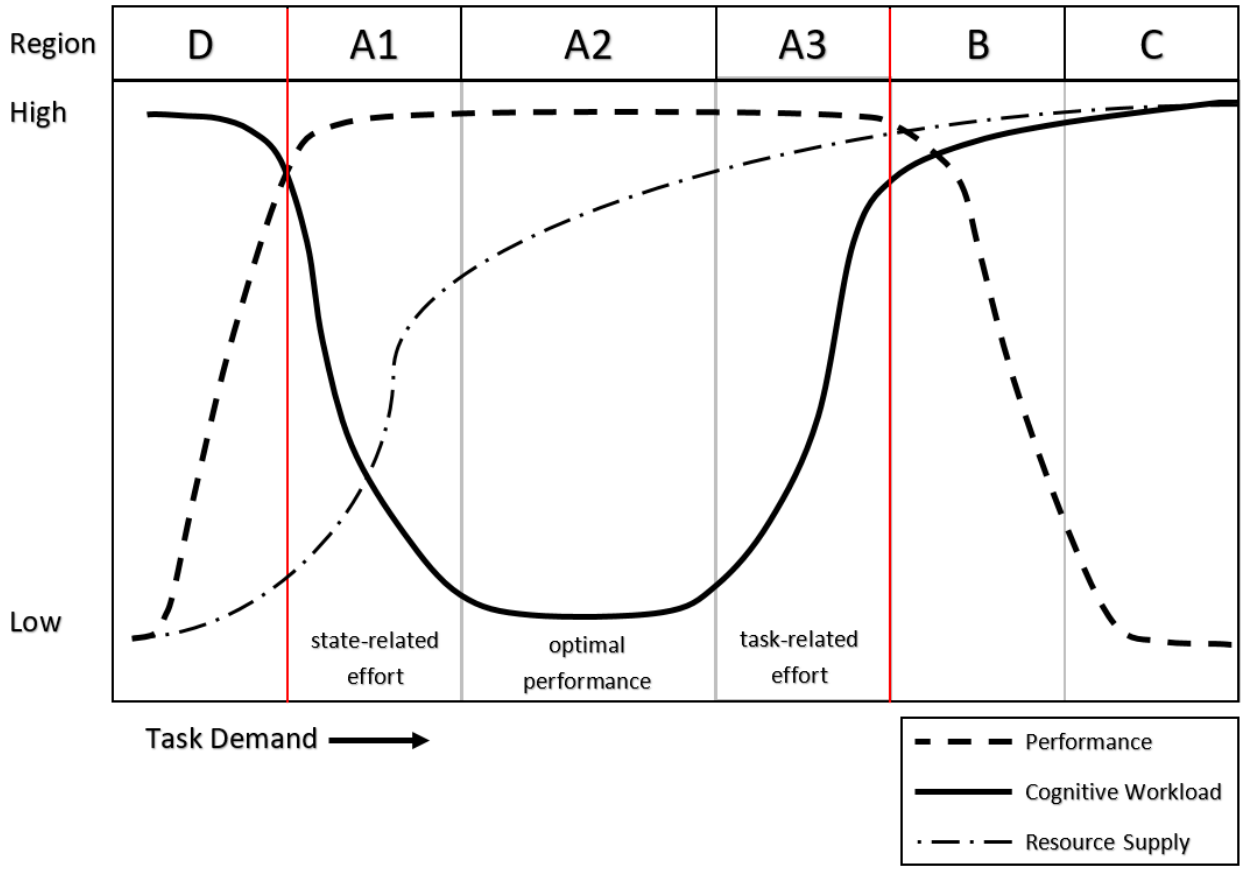

Figure 2. Depiction of the performance-cognitive workload relationship (adapted from De Waard, 1997 and Young et al., 2015).

Identifiable physiological signals have been observed to change under different levels of experienced CWL and some physiological metrics have found success as operational measures of CWL. Metrics such as heart rate, heart rate variability, pupil diameter, electroencephalogram (EEG) signal bands, cerebral oxygenation as measured through functional near infrared spectroscopy (fNIRS), and many other metrics have a body of literature to support their use as a proxy measure for CWL. Unlike performance metrics, physiological metrics allow researchers to tap in to CWL changes that occur in regions where performance remains stable (i.e., regions A1 and A2) whereas CWL is actively changing. In a way, physiological measures allow researchers to see what is going on 'under the hood' as task demands increase. This ability to observe changes in CWL during the approach to a redline speaks to the generally high **sensitivity** of the measurement category. Additionally, it offers researchers in applied domains a means to predict the onset of performance failure and remediate task demands before performance ever begins to suffer. However, other physiological phenomena, such as fatigue, anxiety, or physical movement, can highly interfere with the overall sensitivity of these metrics. Physiological metrics can vary significantly in terms of their **diagnosticity**. Some physiological indices provide a more global scale of experienced CWL, such as pupil diameter or heart rate variability.

Others show higher CWL resource diagnostic ability by pinpointing areas of activation within the brain, such as with EEG or fNIRS metrics, or by being driven by specific task demands (i.e., blink dynamics). Work is ongoing to limit the overall **intrusiveness** of physiological sensors to move towards real-world applications. Some sensors are minimally intrusive (e.g., electrocardiography [ECG], remote eye trackers, etc.) while others can cause higher levels of intrusion (e.g., head-mounted eye trackers, fNIRS, electrodermal activity, etc.). In an aviation context, heart rate and heart rate variability metrics are one of the most widely used physiological metrics due to having a fair balance between sensitivity, diagnosticity, and intrusiveness for the operational environment (Backs, 1995). However, recent research has been moving closer to optimizing metrics such as pupil diameter, fNIRS, and EEG as additional physiological measures to be used within aviation.

CWL is a unique experience that humans can identify and describe through introspection. As such, the self-reportable experience can be captured through the use of structured validated questions that take the form of subjective measures. Over the years, many different subjective scales of CWL have been tested across domains of study and have shown that humans can reliably indicate the level of CWL that they experienced during a specific task. Overall, subjective metrics have shown a great deal of **sensitivity** to changes in CWL. Subjective ratings allow researchers to sample CWL across all regions depicted in Figure 2. Subjective metrics can also range from low to high levels of **diagnositicity** through the use of unidimensional and multidimensional measures, respectively. Unidimensional subjective metrics ask the operator to evaluate a single aspect of their experienced CWL, such as by rating effort expenditure, resource capacity, or general CWL itself. Multidimensional subjective metrics are more diagnostic as multiple questions or subscales tap into many similar, yet distinct, elements of the CWL experience. Unfortunately, subjective metrics are generally very high in **intrusiveness** if they were to be completed during task performance. As such, most subjective measures are completed post-task performance by asking the operator to reflect on their previous CWL experience when answering the questions. Of course, some unidimensional subjective metrics seek to circumvent this limitation by prompting subjects to indicate their subjective CWL during task performance, thus creating a tradeoff with diagnositicity. Overall, subjective metrics have been used a means to validate systems and other CWL metrics. The most ubiquitous measure to stem from this area of research is the National Aeronautics and Space Administration Task Load Index (NASA-TLX), and it still sees widespread use today. The NASA-TLX is often used as a multidimensional CWL subjective metric in the aviation domain, but some measures such as the Bedford Workload Scale and Modified-Cooper Harper Handling Qualities Rating Scale were designed specifically for the aviation domain and are also commonly used as unidimensional measures today.

With each type of cognitive workload assessment technique having their own strengths and drawbacks, it seems natural to combine performance, physiological, and subjective measures into a composite measure of cognitive workload. The logic follows that with each of these cognitive workload responses being measured from the same individual, that the responses would correlate with each other, and should one response fail, the others can serve as a redundant backup. When a composite measure shows increasing cognitive workload in each individual measure, one could be confident that the studied operators were experiencing higher levels of cognitive workload or vise-versa, with matching decreasing cognitive workload responses. In the event that different responses come from each cognitive workload assessment metric, the

experience of the studied operator becomes less clear and more puzzling given that the responses are collected from a single individual. For example, an operator can report low levels of cognitive workload in subjective measures but their physiological measures indicate increasing levels of workload while their performance metrics remain stable. Likewise, the same inconsistencies can be modeled across the different cognitive workload measures, each with responses that indicate high, low, or stable cognitive workload. Hancock and Matthews (2019) explored the concept of associations, insensitivities, and dissociations (AIDs) of cognitive workload assessment to create a framework with which to understand the possible states of composite workload assessment metrics. The three-dimensional matrix defining these possible composite cognitive workload assessment states can be seen in Figure 3.

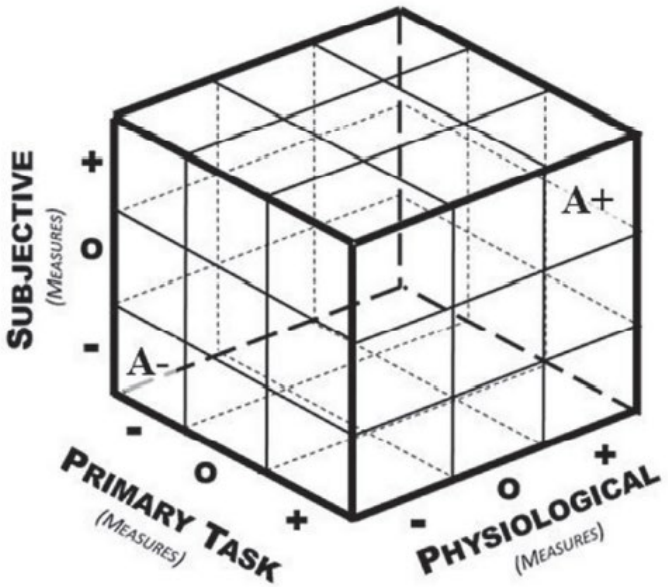

Figure 3. The matrix of Association, Insensitivity, and Dissociation (AIDs) framework of cognitive workload assessment techniques by Hancock and Matthews (2019). Each measure can indicate an increasing (+), decreasing (-), or stable (o) cognitive workload response. With each state represented by a cube within the matrix, 27 combinations of outcomes across performance (primary task), physiological, and subjective measures are possible. When measures agree with each other (i.e., all measures are either showing a decrease or increase in cognitive workload), a double association (denoted by A- for decreasing and A+ for increasing measures) occurs.

The Hancock and Matthews (2019) AIDs taxonomy for composite cognitive workload assessment states presents the three primary methods of cognitive workload assessment along the axes of a cube-shaped matrix. Each method allows for one of three responses: Increasing cognitive workload response (+), decreasing cognitive workload response (-), and stable (i.e., insensitive) cognitive workload response (o). Combining each individual measure's response outcome yields a three-dimensional matrix defining 27 unique states of a composite cognitive workload measure. When responses from different types of workload measures match one another (e.g., increasing workload as indicated by both physiological and subjective measures), an association occurs between the two measures. In the event that two measures' responses disagree with each other, a dissociation occurs. A double association (as denoted by the A+ and A- states in Figure 1) occurs in the event that all three measures report the same responses (i.e.,

all measures show a matched response of increasing, stable, or decreasing cognitive workload). Likewise, double dissociations occur when then all measures disagree with each other. While double associations simplify the problem of interpretation of cognitive workload data, recognizing factors that affect the convergence of measures can aid in understanding why dissociations occur. Hancock and Mathews (2019) elaborate on these issues of convergence between measures and identify common problems that can influence mismatching responses between measures. Factors such as the granularity between measures, the timing of cognitive workload responses across measures, self-regulating strategies, and workload history can all lead to inconsistencies between different measurement techniques. Ultimately, these problems are still unsolved, but recognizing their presence can aid in the interpretation of even the most dissociated data sets.

Goals of This Report

This report examines the CWL literature over the last decade (2010s) to expand on trends reported by Young et al. (2015). To guide the development of CWL research being conducted at the USAARL, a focused search on composite CWL assessment literature in the aviation domain was conducted. Both rotary-wing and fixed-wing aviation platforms were included in the search. From these aviation articles, the frequency of use and success of different CWL metrics are reported. Data regarding differences in CWL assessment as a function of individual differences (i.e., flight experience) and research platform (i.e., simulator or aircraft) are also examined. Lastly, composite CWL metric results are examined through the lens of the AIDs model of CWL assessment.

Methods

A literature review was conducted to explore the literature associated with CWL measurement techniques used within the aviation domain. Initial searches were conducted in three databases, PsycINFO, ProQuest, and Web of Science. Each search included a Boolean list of terms to refer to the use of trained, professional populations with a special emphasis on military and aviation populations used in the papers retrieved.

Boolean list of trained population criteria included in searches:

trained OR trained subjects OR trained operator OR homogeneous population OR air traffic control OR pilots OR aviation OR helicopter pilots OR unmanned aircraft systems operator OR UAS pilots OR UMS pilots OR military OR ATC OR UAS OR driver OR driving OR trucker OR truck driver OR EMS OR EMT OR ambulance OR police OR first responder OR power plant OR nuclear plant OR water plant OR gas plant OR oil rig OR supervisory control and data acquisition OR FPV drone OR drone

This space is intentionally blank.

Each search conducted included a combination of independent variables and dependent variables. The independent variables were different variations of the term CWL or terms that could be used interchangeably in the literature of CWL. Independent variables included:

workload, task load, effort, and load

Dependent variables consisted of individual CWL measurement techniques, or categories of CWL techniques. Dependent variables were divided into three categories, subjective, physiological, and performance-based measurement techniques. Techniques categorized as subjective included:

Instantaneous Self Assessment of Workload, the Malvern Capacity Estimate, The Modified Cooper Harper, the NASA TLX, the Subjective Workload Assessment Technique, Workload Profile, and the Rating Scale Mental Effort

Dependent variables categorized as physiological included:

catecholamines, hormonal response, electrodermal activity, electromyography, electrooculography, pupillometry OR pupil diameter, Index of Cognitive Activity OR ICA, fixat, saccad*, blink duration, blink rate, blink interval, gaze, blood pressure, electrocardiography, electroencephalography, fNIRS, heart rate, heart rate variability, and respiration rate*

Dependent variables categorized as performance-based included:

accuracy, reaction time, secondary task, subsidiary task, loading task, dual task, multiple task, peripheral task, multiple resource theory OR MRT, task overload, and redline

Additionally, searches were filtered to restrict all articles to a 10-year period (2010-2020), be written in the English language, utilize human subjects, and be peer-reviewed. Searches were conducted in the month of March 2020.

The resulting article information found for each search in each database was downloaded and reviewed against exclusion criteria. Duplicates and irrelevant listings (non-assessment, non-peer-reviewed, literature reviews, books, etc.) were removed through automated and manual inspection. Articles were then manually categorized by domain of study. Articles examining the aviation domain using air-rated pilots with reported flight experience were identified for final analysis. Figure 4 details the method and article count results. Study details, utilized CWL metrics and responses, and success rates for a metric to discriminate between CWL conditions were extracted from these aviation articles into a database for analysis.

Results

Using the combinations of independent and dependent variable keywords along with the Boolean list of trained population keywords, a total of 17,423 articles were found across all three search databases. The frequency results of the searches for each CWL measurement technique are reported in Appendix B. Following an automated procedure to remove duplicate article titles (reducing the list to 4130 articles), individual researchers read through each study title and determined those that were irrelevant to this review; the studies deemed irrelevant either were not related to CWL at all, did not use healthy adult populations, or were not experimental studies. Following the initial review of all study titles, a peer-review was conducted to ensure that at least two individual researchers deemed each article irrelevant prior to its removal. Any articles that the two researchers were unable to agree upon were discussed in a group meeting, where a majority vote determined whether the study title was to be deemed irrelevant.

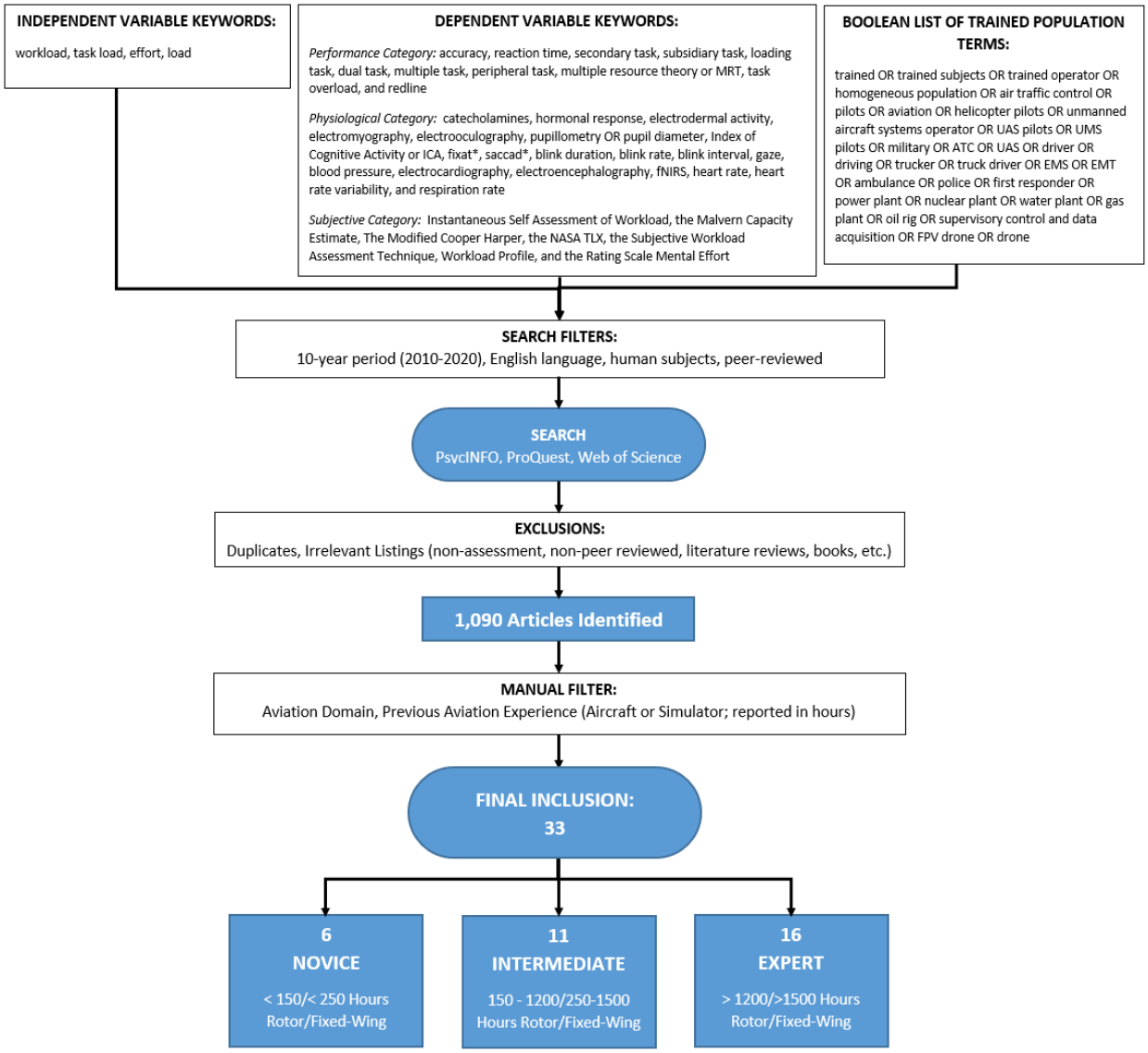

Figure 4. Search terms and filtering process.

Following the exclusion phase, a total of 1090 articles remained for more extensive review. Researchers reviewed the remaining article titles, abstracts, and available text to determine the field of study (i.e., domain) featured in the article. After examining an article, the researcher assigned it a label to classify its domain. The top five most frequently occurring domains of study, making up 72.8% of the surveyed articles, throughout the 2010s included (in descending rank order): laboratory tasks (i.e., simplified variable demand tasks performed in a laboratory setting), driving, aviation, medicine, and human-computer interaction. The frequencies were further broken down by year of publication to visualize the trends over time. The results are depicted in Figure 5.

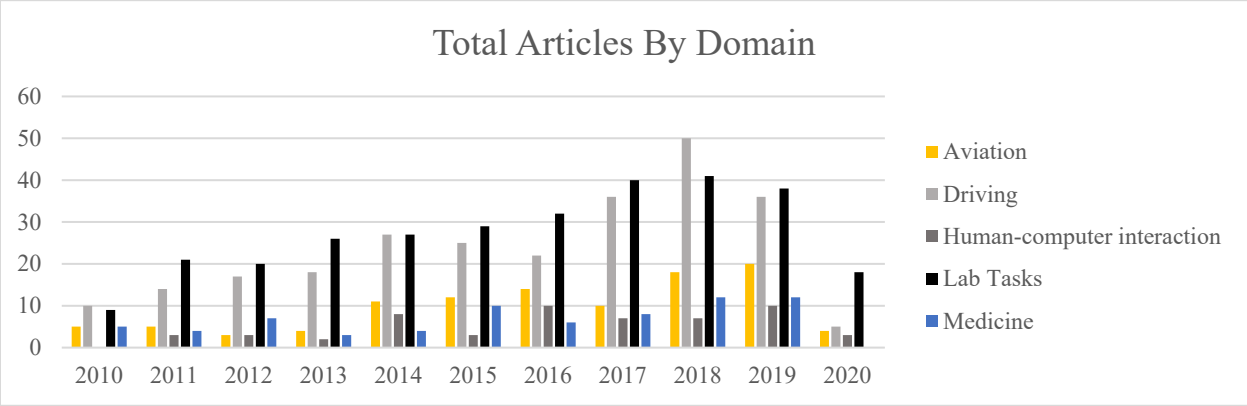

Figure 5. Total articles by domain across the decade. Only the five most frequent domains of study are depicted. Note, only three months of the year 2020 were surveyed at the time the searches were conducted.

Additionally, all articles were classified as to whether they employed a machine learning approach in the classification of CWL metrics between different levels of task demand. This data was extracted from article titles and abstracts to determine the degree of CWL state classification research available. Sixty-one articles took a focused machine learning approach across all domains, with laboratory tasks and human-computer interaction studies being the dominant domains employing machine learning techniques. Over the last decade, there has been a trend of increasing interest in machine learning approaches to CWL assessment and prediction (Figure 6).

This space is intentionally blank.

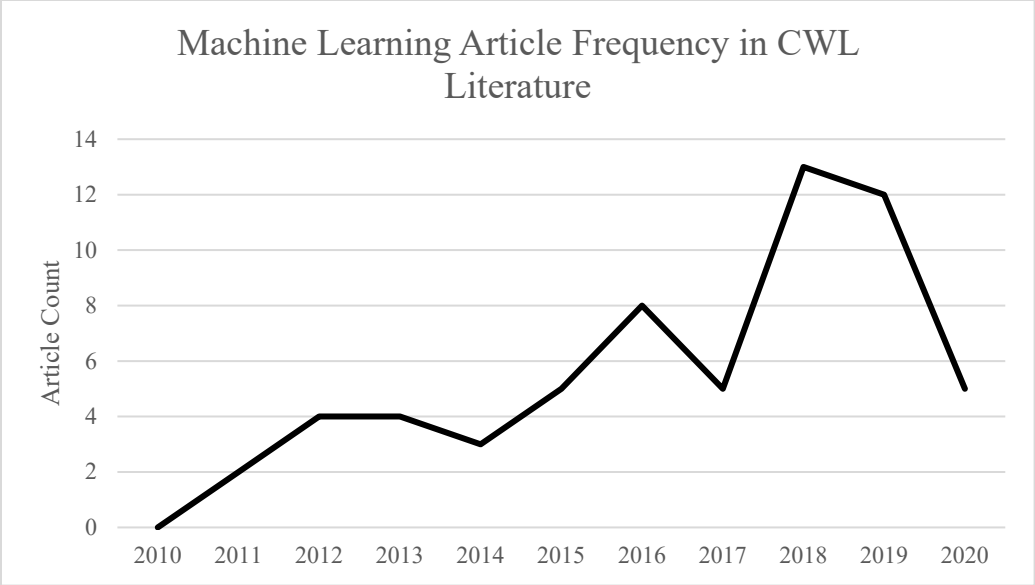

Figure 6. Frequency of machine learning focused articles in the field of CWL over the last decade.
Note. Only the first three months of 2020 were surveyed as that was when the searches were performed.

To further develop the USAARL's understanding of CWL assessment in the aviation domain, the remaining articles were reviewed to isolate aviation articles that utilized trained pilots in a flight simulator or aircraft environment. The criterion used for determining if trained pilots were utilized in a study was whether flight experience in total flight hours was reported. A total of 33 articles met the aviation domain and experience reporting criteria. A total of 7 articles examined rotary-wing aviation and the remaining 26 articles examined fixed-wing aviation. All aviation articles examined the regions of cognitive overload (Regions A3, B, and C in Figure 2).

Vogl et al. (2021) provides an annotated bibliography of the aviation articles included in this review. Each article was analyzed to obtain a list of the utilized CWL metrics and a determination (i.e., yes or no) of whether the metric was able to discriminate between the different CWL levels assessed in the study. Each unique metric was classified in the appropriate category (i.e., performance, physiological, or subjective) and grouped by the method of metric assessment (i.e., accuracy, eye tracking, NASA-TLX, etc.) This served to provide a list of 239 CWL assessment metrics utilized across the 33 aviation articles. First, a general survey of the use of CWL assessment methods was conducted across all aviation articles to determine the most frequently utilized methods used to assess CWL. While multiple metrics were often derived from each method, each method was only counted once for each study. The frequency results are depicted in Figure 7.

This space is intentionally blank.

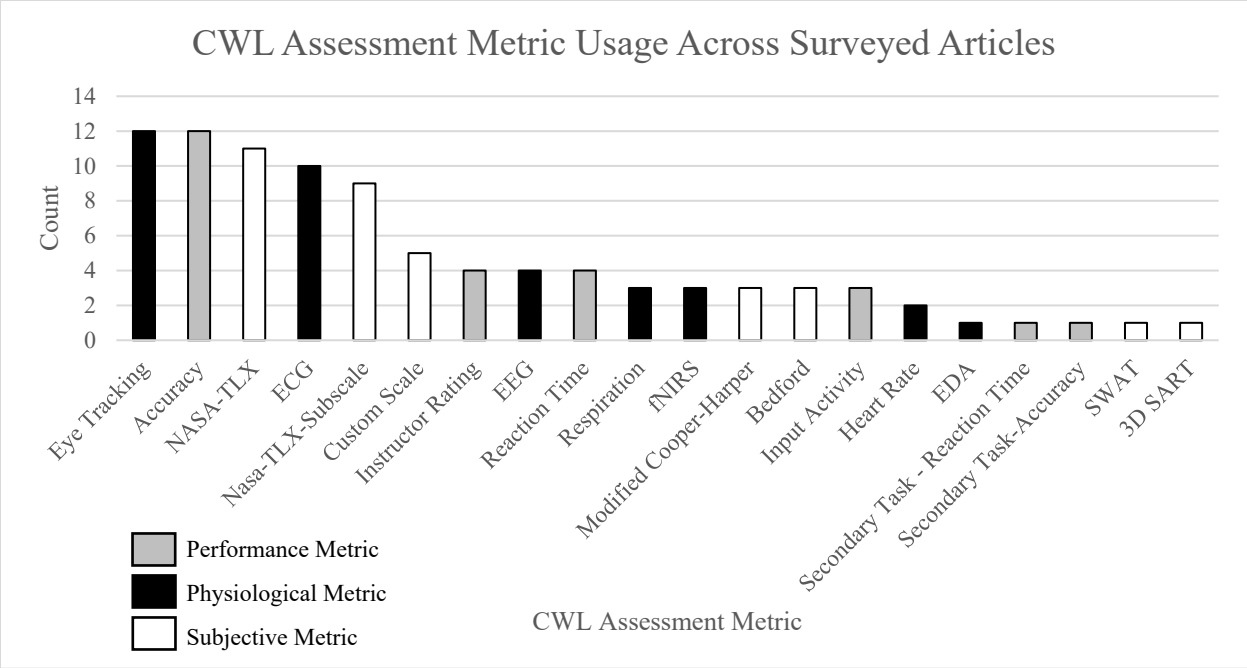

Figure 7. CWL assessment methods usage across surveyed aviation articles. One count per use in a study. ECG = electrocardiogram; EDA = electrodermal activity; SWAT = Subjective Workload Assessment Technique; 3D SART = 3-Dimensional Situation Awareness Rating Technique.

Frequency of use and success rates of CWL level discrimination were derived for specific CWL assessment metrics used within the surveyed studies. A metric was defined as successful in its implementation if it was able to accurately demonstrate a significant difference between CWL levels within a study. Success metrics were assigned a value of one if the metric was successful and a value of zero if it failed to discriminate between CWL conditions. Specific performance metrics were categorized based on the aspects of performance being assessed. Relatively few reaction time, input activity, instructor rating, or secondary task performance assessments were conducted in the surveyed articles; however, several accuracy metrics, defined as deviations from a criterion value for the task, were utilized in the aviation domain. As such, accuracy metrics were split across specific metrics in which deviation from criteria was assessed. For example, path, heading, altitude, and speed deviation from criteria were split from the general accuracy assessment method to provide a more descriptive picture of the metrics utilized. Results are depicted in Figure 8.

This space is intentionally blank.

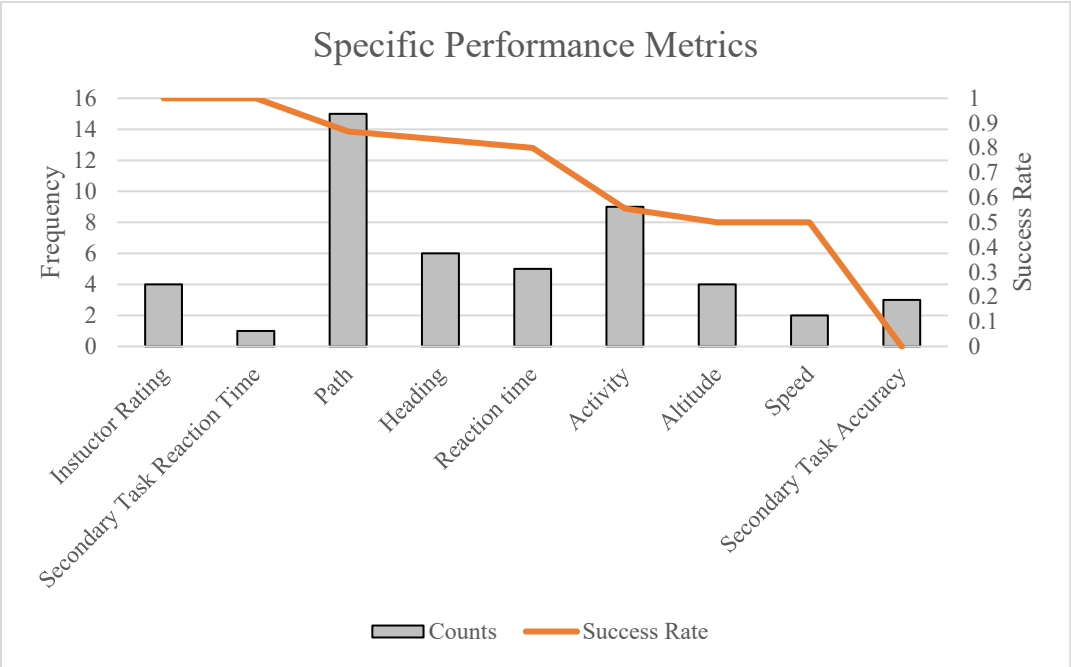

Figure 8. Frequency and success rates of specific performance metrics utilized in the surveyed aviation articles. Multiple uses of the same metric within an article are included in the data.

The same approach was used to derive the frequency of use and success rates for specific physiological metrics. A total of 37 unique physiological metrics were utilized, and the larger graph with each specific metric can be found in Appendix C, Figure C1. General results of physiological metric groups utilized in the surveyed aviation articles (Figure 9). Frequency of use and success rates were also derived for all subjective metrics and subscales used across the surveyed articles (Figure 10).

This space is intentionally blank.

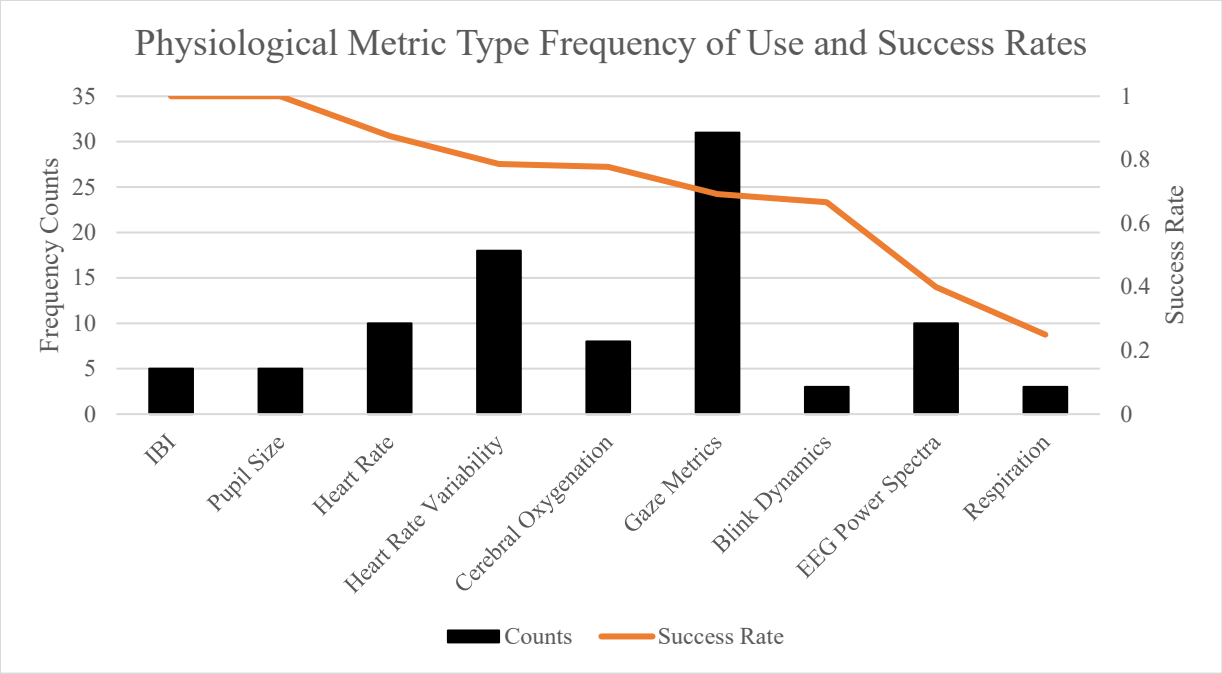

Figure 9. Frequency and success rates of specific physiological metric categories utilized in the surveyed aviation articles. Multiple uses of the same metric within an article are included in the data. IBI = inter-beat interval.

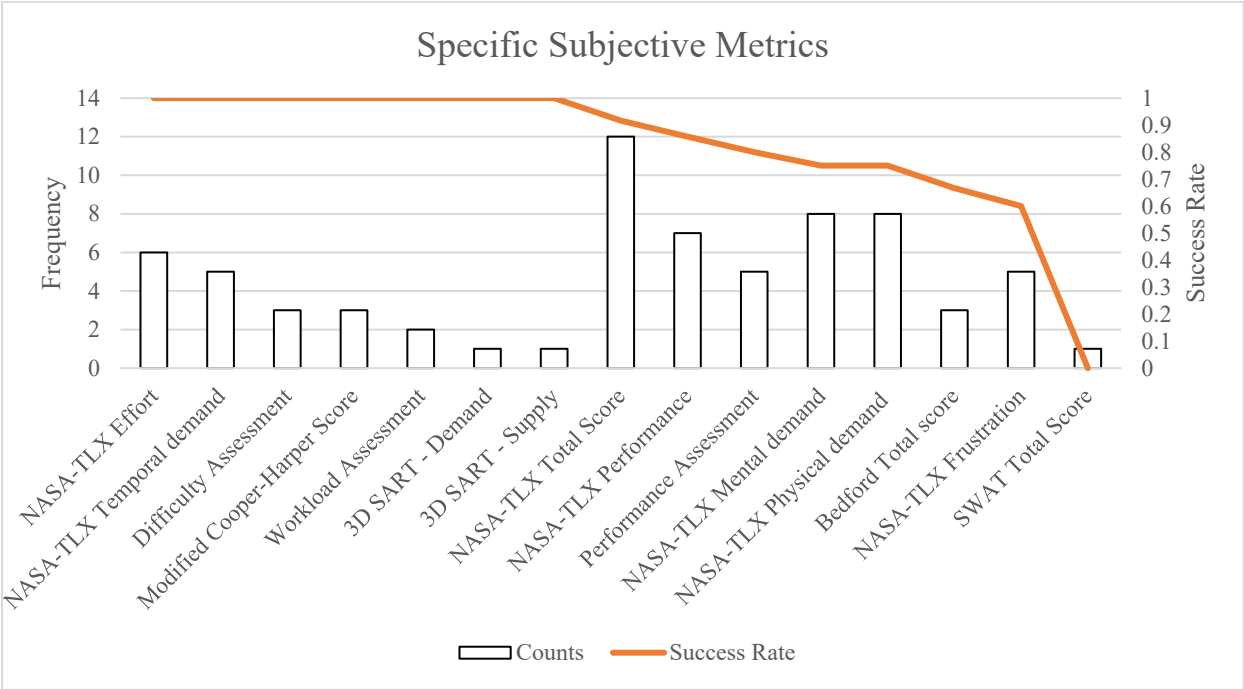

Figure 10. Frequency and success rates of specific subjective metrics used in the surveyed aviation articles. Multiple uses of the same metric within an article are included in the data.

The flight hour reports were pulled from each aviation article and compared against licensure requirements set forth by the Code of Federal Regulations (CFR) for determining experience categories. The novice experience category included those with less than 150 reported flight hours in rotary-wing aircraft and 250 reported flight hours in fixed-wing aircraft. The ceiling of the novice threshold marks the transition from private pilot status (requiring 40 flight hours, as described in CFR Title 14, Chapter I, Subchapter D, Part 61, Subpart E, Section 61.109 to commercial pilot status. The intermediate experience category consisted of reported flight hours between 150 and 1200 for rotary-wing and 250 and 1500 for fixed-wing aircraft (CFR Title 14, Chapter I, Subchapter D, Part 61, Subpart F, Section 61.129). Lastly, the expert group consisted of any pilot with more than 1200 flight hours with rotary-wing or 1500 hours of fixed-wing aircraft, qualifying for airline transport pilot status as per CRF Title 14, Chapter 1, Subchapter D, Part 61, Subpart G, Sections 61.159 and 61.161. This categorization yielded a total of 6 novice experience studies, 11 intermediate experience studies, and 16 expert experience studies, as depicted in Figure 4. CWL assessment categories were categorized as a function of aviator experience (Figure 11).

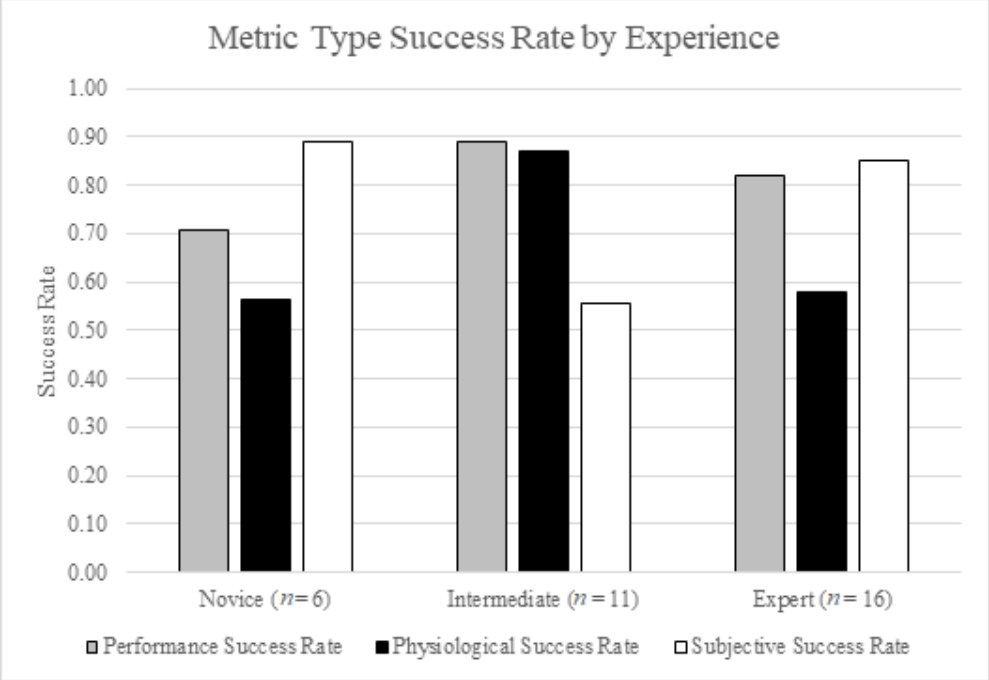

Figure 11. Success rates of CWL assessment categories by reported aviator experience.

In addition to factors intrinsic to the study population (i.e., flight experience in hours) CWL metrics can also be affected by extrinsic variables, such as environmental factors. To determine if experimental platforms affected the success of CWL assessment metrics, the success rates of each CWL assessment category were plotted as categorized into flight simulations or real aircraft platforms in Figure 12. Note that no performance metrics were collected in studies utilizing real aircraft.

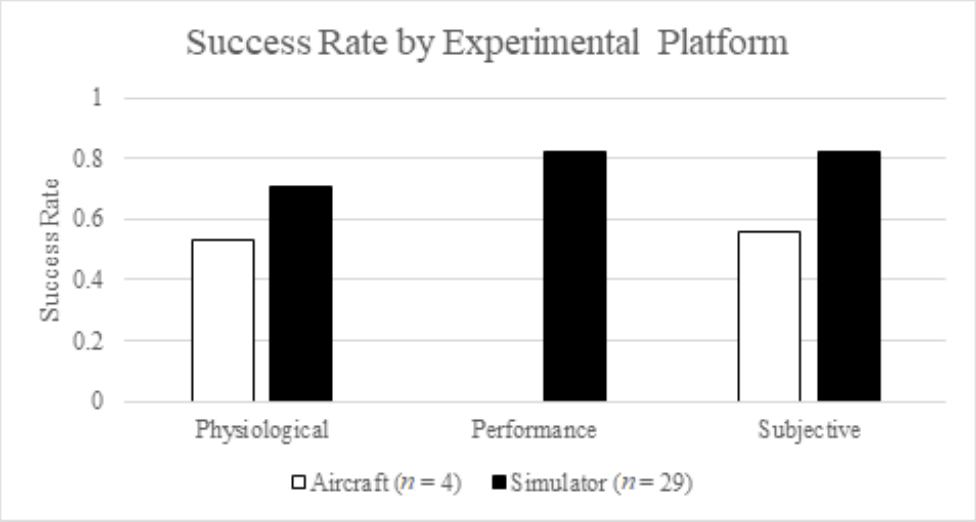

Figure 12. Success rates of CWL assessment metric categories based on experiment platform (flight simulator vs. real aircraft).

Lastly, the surveyed aviation articles were assessed through the lens of the AIDs model of workload. Success rates of the CWL metrics for each study were examined as to whether the responses were typical for increasing, insensitive to, or decreasing CWL signals relative to the demand level experienced. Results are depicted in Figure 13.

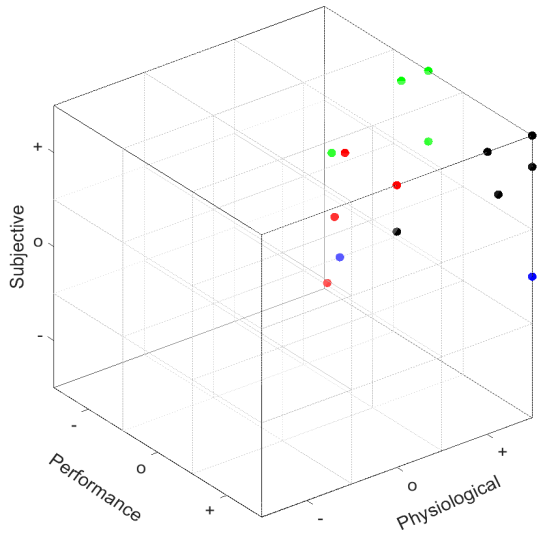

Figure 13. Depiction of where each study with three CWL metrics (black dots) falls within the states defined by the AIDs model of CWL assessment. Studies with two measures were also included for reference; however, are not typical use cases for the AIDs model. Performance + physiological metric studies are in blue, performance + subjective metric studies are in red, physiological + subjective studies are in green.

Discussion

The assessment of CWL has continuously been developed over the last 60 years and has become an increasingly popular field of study in the recent decade (see Figure 1). The general focus of applied CWL research circles around safety-critical domains such as transportation, medicine, and military contexts. Young et al. (2015) detailed the trends in research area-focus over the decades. In the 1980s, CWL studies featured manipulations of adaptive interfaces as computer memory and speeds necessitated the development of human system interfaces for this rapidly developing technology. Moving through the 1990s and early 2000s, driving and aviation emerged as the most frequent independent variables of interest. The Army has interest in adaptive automation in the design and development of future rotary-wing cockpit design within the aviation domain. With the advancement of sensors and computer processing power, in the near future, it may be possible to monitor aviators CWL in near real-time and have the machine adapt the level of automation to support the crew. In other words, adaptive automation allows for dynamic regulation of tasks by means of monitoring the pilot to determine which cognitive resources are being overtaxed. By determining the overloaded resources, adaptive automation may be able to take control of specific tasks or offload tasks to other cognitive resources that are less taxed (as outlined by Wickens' [2008] work with multiple resource theory [MRT]).

The results of this literature review show that CWL research in the aviation domain was surpassed by the driving and general laboratory-based task domains throughout the 2010s. A total of 106 aviation articles were identified in the database search process, but studies focused on active pilots (with reported flight hours) interacting with actual aviation tasks (flight simulator or aircraft) reduced the relevant articles to 33 for this review. This roughly 31% yield indicates a need for researchers in the aviation domain to continue to use more flight-rated pilots in real or simulated flight settings in their research. Additionally, the proportion of rotary-wing (7 articles) to fixed-wing (26 articles) articles analyzed in this review is staggeringly in favor of fixed-wing platforms. The open literature would benefit from additional research in the rotary-wing domain.

Within the aviation articles reviewed, all three categories of metrics were similarly represented. Twenty-five used performance, 33 used subjective, and 35 used physiological measures (Figure 7, recall that some studies used more than one type of measure within each category). These high frequencies (relative to the total number of 33 articles) indicate that a composite or redundant CWL assessment approach is consistently used within the aviation literature. Counting uses of each CWL category, 20 out of 33 of the surveyed aviation articles included two or more types of CWL assessment categories. Many articles (29 out of 33) reported multiple CWL assessment methods within the same category, as a form of redundancy (i.e., recording multiple physiological signals such as eye tracking and ECG).

For performance metrics, accuracy-based metrics (defined as deviation from criteria) were far more widely used than reaction time, input activity, or secondary task performance metrics. To further analyze how accuracy metrics have been measured in the aviation domain, subcategories of accuracy metrics were identified and plotted with other performance metrics and the reported success rates in Figure 8. Path deviation (i.e., horizontal and lateral deviation from a predefined path or landing location) was the most frequently reported performance metric across the articles, and demonstrated a high success rate (86.7%) in discriminating between different levels of task demand. Instructor ratings of the studied pilots offered a reliable method

(100% success rate) to assess performance between different task demand conditions. Only one study, performed by Hildalgo-Munoz (2018), added a secondary task CWL assessment, while flying was the primary. It is not surprising to see little use of secondary task techniques in an aviation setting due to the high intrusiveness of the assessment techniques. However, many of the studies evaluate each subject one at a time and did not include all responsibilities of crew coordination. Aviators perform multiple tasks simultaneously in flight, which lends toward researchers being able to easily add secondary, or more, embedded tasks (i.e., secondary tasks that are performed naturally in the aviation setting and are part of crew coordination). Creating a research paradigm where secondary CWL tasks are embedded into the primary task could provide a more diagnostic look at performance metrics in the future. Notably, no performance metrics were reported in the four studies that used aircraft; all performance measures were collected during simulated flight conditions.

Two primary physiological-based methodologies were employed in the aviation CWL articles, ECG and eye tracking. It is no surprise to see ECG as a highly used CWL assessment technique given its extensive use in the aviation domain. However, the heavy use of eye tracking metrics is more surprising. Eye tracking technology has seen significant advancements over the decades, advancements which not only improve the robustness of the technology in applied environments, but also reduce the overall burden of using it as a research tool, both as a financial investment in the technology and in ease of use. Increased robustness and ease of use may explain why eye tracking has become such a commonly used tool in the assessment of CWL in the recent decade. As with most physiological CWL assessments, there are numerous methods to analyze the resultant data to derive specific metrics that demonstrate sensitivity to changes in experienced CWL. The number of different metrics within eye tracking and ECG measurements offers an appealing exploratory testbed for the applied researcher. Figure C1 in Appendix C breaks down the types of metrics into more specific physiological parameters. Both CWL assessments derived from cardiac activity and ocular behavior have shown great success in discriminating between the tested CWL conditions within the aviation domain. The most frequently calculated physiological metrics were heart rate and eye gaze fixation duration. Both were reported in eight articles; however, heart rate (75%) was much more successful at discriminating CWL conditions than fixation duration (25%), which highlights that more research may need to be conducted with eye tracking CWL metrics. Of course, there may be an artificial inflation of success rates given the lack of consistent usage of each metric across studies. Future research can aid in CWL assessment comparisons by adopting a composite metric approach by utilizing multiple physiological CWL assessment techniques within a single study.

It may be that measures of cardiac and ocular responses to CWL in aviation benefit from little movement of the pilots who sit in a relatively fixed posture. This required element of aviation reduces movement artifacts in the data, which improves the quality and quantity of useable information available from each data collection session. Stationary subjects, with limited movement, and limited fields-of-view are ideally suited for both heart and eye dynamics. Both of these measures emerged from ideal, well-controlled clinical and laboratory settings. As these measurement technologies have been adapted in operational domains, the clinical and laboratory precision has been reduced due to the less controlled conditions of operational settings. The aviation and driving domains are much more controlled operational environments than other physiological research areas that require varying levels of CWL, such as free-living people; Service Members; and professional athletes engaged in sports like football, soccer, baseball, auto

racing, sailing, golf, etc. Another setting that also limits physical movement while providing varying levels of CWL is in the surgical suite. It was not surprising to see that the medical domain was also frequently investigated in the articles we found in our review of the initial articles returned from the literature search.

Performance and physiological measures and metrics derived from these assessment techniques are used to provide objective evaluations of CWL to avoid the inherent biases that complicate subjective measures of CWL. Subjective ratings of CWL requires study participants to accurately assess the condition and provide a value based on established criteria. An individual's experience, skill at the task, current cognitive state, and trait tendencies influence their perception of CWL, which may cause increased variability in their responses. However, one metric that was specifically designed to capture subjective CWL in the aerospace domain is the NASA-TLX. Among subjective metrics reported, it was no surprise to see the NASA-TLX total score as the most frequently reported subjective workload metric in the aviation literature. In fact, the NASA-TLX and its six subscales made up half of the subjective metrics reported in the articles reviewed. The NASA-TLX Total Score demonstrated a total score success rate of 92%, while subscales were found to have success rates of 100% for effort, 100% for temporal demand, 86% for performance, 75% for mental demand, 75% for physical demand, and 60% for frustration.

Another subjective measure designed for the aerospace domain is the Modified Cooper-Harper (MCH). The MCH handling qualities rating scale showed a 100% success rate at discriminating among conditions of CWL in the three studies that reported its use. Other subjective scales, such as the Bedford Workload Scale (67%) and Subjective Workload Assessment Technique (SWAT) (0%), were not used as frequently and had lower success rates. In the aviation articles reported here, 10 studies specifically utilized subjective metrics paired with either performance or physiological metrics. This approach allows for the further development and validation of objective scales of CWL based on performance and physiological methods of CWL assessment, which show more variability in success rates.

As mentioned previously, other factors impact the experience of CWL and CWL impact on performance, physiology, and subjective ratings. One of these factors is experience at the task being performed. To this end we were able to separate the subjects based on flight hour to separate three groups of increasing experience. Success rates of the three categories of CWL evaluation across pilots who reported low (novice), intermediate, and high (expert) flight hours were analyzed in this review. These three groups were determined by the Code of Federal Regulations definition of experience. Unfortunately, there were a limited number of articles that provided the flight experience of the subject pools used; however, it is interesting to see that the experience of the pilot can affect the general outcome of a CWL assessment. The data presented here, by visual inspection alone, suggests that subjective metrics are less discriminating when used with pilots with intermediate experience compared to novice or expert pilots. Conversely, physiological metrics employed with pilots with intermediate experience found the most success while the success rate dropped substantially for novice and expert pilots. Across experience groups, performance metrics remained steadier, with CWL discrimination success rates above 70%.

In addition to exploring the impact of experience in aviation CWL, we also looked at the impact of the type of research platform utilized in the study. Differences in success rates based on the type of experiment platform (i.e., flight simulator vs. real aircraft) reveal that a metric's ability to successfully discriminate between task demand conditions is generally poorer in real aircraft environments. This is of little surprise given the more controlled settings of flight simulators. Interestingly, no performance metrics were recorded for studies involving real aircraft platforms. While most flight simulators offer performance readouts, it may be more difficult to track performance in a real aircraft if it is not instrumented for research purposes. Only four aircraft studies met our strict inclusion criteria, and future efforts should involve aircraft flight and provided adequate information about the subjects to help tease apart the CWL imposed by the flying tasks and the other factors named above that can also impact performance, physiology during flying, and subjective ratings. More research to examine the differences between CWL assessment methods between simulated and real flight environments is highly desired.

The surveyed aviation articles were examined through the lens of the AIDs model of CWL assessment. Only seven out of the 33 aviation articles surveyed in this review reported using all three CWL assessment categories. Double associations occurred for six out of the seven articles, with only physiological metrics showing insensitivity in one outlier study. This three-dimensional association shows a strong relationship among the three dependent variables. Similarly, the two-dimensional relationships also tend to demonstrate a positive association among the measures reported in each study. Of course, this steps out of the use-case for the AIDs model, as a third dimension is missing in these cases. However, it does allow for demonstration that CWL assessment results generally agree with one another, as sampled through the surveyed aviation articles. The low number of samples provided from this literature review, and tendency of publication bias to select for studies with only significant results, may skew the results of this graphic. Conversely, using highly trained populations, such as flight-rated pilots, may also lead to a more consistent tendency of CWL metric results. It would be worthwhile for even insignificant findings and unpublished results to be plotted in a future report based on the method used to generate Figure 13. This would allow researchers to start determining where CWL assessment techniques truly associate and dissociate for a specific task domain. This approach may also serve as a starting point to clearly define what other cognitive influences (e.g., fatigue, fear, frustration, or other operational stressors) affect the utilized CWL metrics (i.e., to determine the metric's specificity for CWL). Given specific task paradigms, dissociations are more likely to occur when other cognitive processes influence the output of the CWL metrics being captured.

Overall, CWL assessment continues to be a popular and growing area of research as the price of technology continues to decrease, computational power increases, and the field shares software toolchains for processing the data into meaningful results quickly and accurately. There is reinvigorated interest in the original 1980s ideas of adaptive system development as driven through CWL assessment techniques to meet the needs of future Army aviation. Research is focused on using physiological CWL metrics as a means of real-time "operator state monitoring" to drive adaptive automation at the USAARL for the emerging FVL platforms and future block updates to the enduring fleet of Black Hawks, Chinooks, and Apache airframes. The results of this review detail commonly used performance, physiological, and subjective metrics that have seen success in the aviation domain and which may serve as cornerstone metrics upon which future research can build. Path deviation, heart rate, and NASA-TLX total score metrics made up

the most frequent and successful discriminating CWL metrics among the articles reviewed here. Promising and popular metrics of ocular activity have come to the forefront in the surveyed aviation literature and offer a supporting physiological metric that has potential for success in the aviation domain after solving a few more technological challenges, such as accounting for bright sunlight and changes in pupil diameter caused by luminance changes in the cockpit. As a final word, the results of this review demonstrate successes in the area of CWL measurement in aviation with trained pilots.

Limitations

This review has some limitations. This review surveyed the open literature using three databases (PsycInfo, ProQuest, and Web of Science). There may be many studies regarding CWL assessment that were not accessible via these databases and not included in this analysis. For example, these databases may not tap into more applied military research in the aviation domain, such as that found in the Defense Technical Information Center database. Another limitation is that the success rate metric used to evaluate if a metric was able to differentiate between different CWL conditions is heavily influenced by the number of reported metrics across multiple articles. As such, each metric has a different denominator in its success rate calculation, leading to potential artificial inflation of metric success. Lastly, additional advancement in CWL assessment metrics has undoubtedly occurred between the date the initial database searches were conducted and the publication of this review.

Conclusions

This review was conducted to explore CWL research and assessment techniques utilized in the aviation domain over the last decade. The following information was learned:

1. Aviation is an active and growing domain of CWL research; however, it is dwarfed by research in the driving domain.
2. There is generally equal representation of each category (performance, subjective, physiological) of CWL assessment techniques over the past decade in the aviation domain. Path deviation, heart rate metrics, and the NASA-TLX total score made up the most frequent and successfully discriminating CWL metrics among the surveyed articles.
3. The most frequently used physiological CWL assessment methods were ocular and cardiac activity. This suggests that the limitations in aviation environment (such as restrained movement and fields-of-view) may be well-suited for these types of metrics and are worthy of continued research focus.
4. Authors should report how much time the subject population has had performing the specific study task at the time of data collection to help establish the subjects' level of expertise, which is useful in teasing out CWL from other factors that impact CWL measures to afford comparisons across articles.
5. Authors using multiple CWL assessment techniques should examine the composite CWL signature alongside individual metrics. Determining which measures show associations, insensitivities, and dissociations within specific contexts and populations will help identify which measures excel in the studied environment.

Additionally, identification of dissociations of metrics can aid in future development of measures to improve their specificity, diagnosticity, and reliability.

References

Babu, M. D., JeevithaShree, D. V., Prabhakar, G., Saluja, K., Paskilkar, A., & Biswas, P. (2019). Estimating pilots' cognitive load from ocular parameters through simulation and in-flight studies. *Journal of Eye Movement Research, 12*(3).

Backs, R. W. (1995). "Going beyond heart rate: autonomic space and cardiovascular assessment of mental workload." *The International Journal of Aviation Psychology*, *5*(1), 25-48.

Bezerra, F. G., & Ribeiro, S. L. (2012). Preliminary study of the pilot's workload during emergency procedures in helicopter air operations. *Work, 41*(1), 225–231.

Cain, B. (2007). *A review of the mental workload literature.* Defence Research and Development Canada.

Çakır, M. P., Vural, M., Özgür Koç, S. & Toktaş, A. (2016). *Real-time monitoring of cognitive workload of airline pilots in a flight simulator with fNIR optical brain imaging technology*. Springer International Publishing.

Causse, M., Chua, Z., Peysakhovich, V., Del Campo, N., & Matton, N. (2017). Mental workload and neural efficiency quantified in the prefrontal cortex using fNIRS. *Scientific Reports, 7*(1), 5222.

Charles, R. L., & Nixon, J. (2019). Measuring mental workload using physiological measures: A systematic review. *Applied ergonomics*, *74*, 221–232.

Cheung, B., Craig, G., Steels, B., Sceviour, R., Cosman, V., Jennings, S., & Holst, P. (2015). In-flight study of helmet-mounted symbology system concepts in degraded visual environments. *Aerospace Medicine and Human Performance, 86*(8), 714–722.

Cheung, B., McKinley, R., Steels, B., Sceviour, R., Cosman, V., & Holst, P. (2015). Simulator study of helmet-mounted symbology system concepts in degraded visual environments. *Aerospace Medicine and Human Performance, 86*(7), 588–598.

Dahlstrom, N., Nahlinder, S., Wilson, G., & Svensson, E. (2011). Recording of psychophysiological data during aerobatic training. *The International Journal of Aviation Psychology, 21*(2), 105–122.

Dalecki, M., Bock, O., & Guardiera, S. (2010). Simulated flight path control of fighter pilots and novice subjects at +3 Gz in a human centrifuge. *Aviation, Space, and Environmental Medicine, 81*(5), 484–488.

De Waard, D. (1996). *The measurement of drivers' mental workload* [Doctoral Dissertation; University of Groningen].

Di Stasi, L. L., Diaz-Piedra, C., Suarez, J., McCamy, M. B., Martinez-Conde, S., Roca-Dorda, J. & Catena, A. (2015). Task complexity modulates pilot electroencephalographic activity during real flights. *Psychophysiology, 52*(7), 951–956.

Diaz-Piedra, C., Rieiro, H., Cherino, A., Fuentes, L. J., Catena, A., & Di Stasi, L. (2019). The effects of flight complexity on gaze entropy: An experimental study with fighter pilots. *Applied Ergonomics, 77*, 92–99.

Eichinger, A. & Kellerer, J. (2014). Between laboratory and simulator: A cognitive approach to evaluating cockpit interfaces by manipulating informatory context. *Cognition, Technology & Work, 16*(3), 417–427.

Gateau, T., Durantin, G., Lancelot, F., Scannella, S., & Dehais, F. (2015). Real-time state estimation in a flight simulator using fNIRS. *PLoS One, 10*(3), e0121279.

Han, S., Kim, J., & Lee, S. (2019). *Recognition of pilot's cognitive states based on combination of physiological signals.* 2019 7th International Winter Conference on Brain-Computer Interface (BCI).

Hancock, P. A. (2013). Task partitioning effects in semi-automated human-machine system performance. *Ergonomics, 56*(9), 1387–1399.

Hancock, P. A. & Matthews, G. (2019). Workload and performance: Associations, insensitivities, and dissociations. *Human Factors, 61*(3), 374–392.

Heard, J., Harriott, C. E., & Adams, J. A. (2018). A survey of workload assessment algorithms. *IEEE Transactions on Human-Machine Systems*, *48*(5), 434–451.

Hidalgo-Muñoz, A. R., Mouratille, D., Matton, N., Causse, M.,. Rouillard, Y., & El-Yagoubi, R. (2018). Cardiovascular correlates of emotional state, cognitive workload and time-on-task effect during a realistic flight simulation. *International Journal of Psychophysiology, 128*, 62–69.

Hsu, C.-K., Lin, S., & Li, W. (2015). *Visual Movement and Mental-Workload for Pilot Performance Assessment*. Springer International Publishing.

Kahneman, D. (1973). *Attention and effort* (Vol. 1063, pp. 218-226). Prentice-Hall.

Kallus, K. W., Tropper, K., & Boucsein, W. (2011). The importance of motion cues in spatial disorientation training for VFR-pilots. *The International Journal of Aviation Psychology, 21*(2), 135–152.

Kim, S. H., Prinzel, L., Kaber, D., Alexander, A., Stelzer, E., Kaufmann, K., & Veil, T. (2011). Multidimensional measure of display clutter and pilot performance for advanced head-up display. *Aviation, Space, and Environmental Medicine, 82*(11), 1013–1022.

Li, W.-C., Horn, A., Sun, Z., Zhang, J., & Braithwaite, G. (2020). Augmented visualization cues on primary flight display facilitating pilot's monitoring performance." *International Journal of Human-Computer Studies, 135*, 102377.

Longo, L., Wickens, C. D., Hancock, G., & Hancock, P. A. (2022). Human mental workload: A survey and a novel inclusive definition. *Frontiers in psychology*, *13*.

Mansikka, H., Simola, P., Virtanen, K., Harris, D., & Oksama, L. (2016). Fighter pilots' heart rate, heart rate variation and performance during instrument approaches. *Ergonomics, 59*(10), 1344–1352.

Mansikka, H., Virtanen, K., Harris, D., & Simola, P. (2016). Fighter pilots' heart rate, heart rate variation and performance during an instrument flight rules proficiency test. *Applied Ergonomics, 56*, 213–219.

Mansikka, H., Virtanen, K., & Harris, D. (2019a). Comparison of NASA-TLX scale, Modified Cooper-Harper Scale and mean inter-beat interval as measures of pilot mental workload during simulated flight tasks. *Ergonomics, 62*(2), 246–254.

Mansikka, H., Virtanen, K., & Harris, D. (2019b). Dissociation between mental workload, performance, and task awareness in pilots of high performance aircraft. *IEEE Transactions on Human-Machine Systems, 49*(1), 1–9.

Martins, A. P., Holscher, S., & Daqutermann, T. (2020). Evaluation of a tunnel-in-the-sky head-up display design for curved approaches using eye-tracking. *Aviation Psychology and Applied Human Factors,* 10(1), 3–12.

Müller, S., Schreiter, K., Luckner, R., & Manzey, D. (2017). Manual flying and energy awareness. *Aviation Psychology and Applied Human Factors,* 7(1), 18–27.

O'Donnell, R. D. and F. T. Eggemeier (1986). Workload Assessment Methodology. In K. R. Boff, & J. P. Thomas (Eds.), *Handbook of Perception and Human Performance. Volume II. Cognitive Processes and Performance* (pp. 41-49). John Wiley and Sons.

Scannella, S., Peysakhovich, V., Ehring, F., Lepron, E., & Dehais, F. (2018). Assessment of ocular and physiological metrics to discriminate flight phases in real light aircraft. *Human Factors, 60*(7), 922–935.

Stanton, N. A., Plant, K., Roberts, A., Allison, C., & Harvey, C. (2018). The virtual landing pad: Facilitating rotary-wing landing operations in degraded visual environments. *Cognition, Technology & Work, 20*(2), 219–232.

Tao, D., Tan, H., Wang, H., Zhang, X., Qu, X., & Zhang, T. (2019). A systematic review of physiological measures of mental workload. *International journal of environmental research and public health*, *16*(15), 2716.

Van Acker, B. B., Parmentier, D., Vlerick, P., & Saldien, J. (2018). Understanding mental workload: From a clarifying concept analysis toward an implementable framework. *Cognition, Technology & Work,* 20(3), 351–365.

Vogl, J. Hayes, A., Aura, C., Milam, L., Temme, L. & St. Onge, P. (2020). *A case for a measurement of incremental workload* (USAARL-TECH-TR--2020-38). U.S. Army Aeromedical Research Laboratory.

Wang, Z., Zheng, L., Lu, Y., & Fu, S. (2016). Physiological indices of pilots' abilities under varying task demands. *Aerospace Medicine and Human Performance, 87*(4), 375–381.

Westbrook, C. B., Anderson, R. O., & Pietrzak, P. E. (1966). *Handling qualities and pilot work load* (AFFDL-FDCC-TM-66-5). AirForce Flight Dynamics Laboratory.

Wickens, C. D. (1984). Processing resources in attention. In *Varieties of attention*. Academic Press.

Wickens, C. D. (2008). Multiple resources and mental workload. *Human Factors, 50*(3), 449–455.

Yang, J. H., Kennedy, Q., Sullivan, J., & Fricker, R. (2013). Pilot performance: Assessing how scan patterns & navigational assessments vary by flight expertise. *Aviation, Space, and Environmental Medicine, 84*(2), 116–124.

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.

Yu, C. S., Yang, E., Li, W., & Braithwaite, G. (2014). Pilots' visual scan patterns and situation awareness in flight operations. *Aviation, Space, and Environmental Medicine, 85*(7), 708–714.

Zhang, X., Xue, H., Qu, X., & Li, T. (2017). *Can fixation frequency be used to assess pilots' mental workload during taxiing?* Springer International Publishing.

Zheng, Y., Lu, Y., Wang, Z., Huang, D., & Fu, S. (2015). Developing a measurement for task complexity in flight. *Aerospace Medicine and Human Performance, 86*(8), 698–704.

Zheng, Y., Lu, Y., Jie, Y., & Fu, S. (2019). Predicting workload experienced in a flight test by measuring workload in a flight simulator. *Aerospace, Medicine and Human Performance, 90*(7), 618–623.

Appendix A. Acronyms and Abbreviations

ATC	Air Traffic Control
CFR	Code of Federal Regulations
CWL	Cognitive Workload
ECG	Electrocardiogram
EEG	Electroencephalogram
EMS	Emergency Medical Services
EMT	Emergency Medical Technician
FAA	Federal Aviation Administration
FARA	Future Attack Reconnaissance Aircraft
FLRAA	Future Long-Range Assault Aircraft
fNIRS	Functional Near Infrared Spectroscopy
FPV	First Person View
FVL	Future Vertical Lift
ICA	Index of Cognitive Activity
MCH	Modified Cooper Harper
MDO	Multi-Domain Operations
MRT	Multiple Resource Theory
NASA-TLX	National Aeronautics and Space Administration Task Load Index
OSM	Operator State Monitoring
SWAT	Subjective Workload Assessment Technique
UAS	Unmanned Aircraft Systems
USAARL	U.S. Army Aeromedical Research Laboratory

Appendix B. Summary of Initial Search Results

Table B1. Summary of Initial Search Results, Number of Duplicate Articles, Number of Irrelevant Titles, and Remaining Articles for Subjective CWL Measurement Techniques

CWL Measurement Technique	Initial search Number of Hits	Number of Duplicate Article Titles	Number of Irrelevant Article Titles	Number of Articles Remaining
Instantaneous Self Assessment of Workload (ISA)	100	39	13	48
Malvern Capacity Estimate (MACE)	88	7	7	74
MCH	20	7	1	12
SWAT	341	210	29	102
NASA-TLX	1406	973	218	215
Workload Profile	236	140	17	79
Rating Scale Mental Effort (RMSE)	100	48	26	22

Table B2. Summary of Initial Search Results, Number of Duplicate Articles, Number of Irrelevant Titles, and Remaining Articles for Physiological CWL Measurement Techniques

CWL Measurement Technique	**Initial search Number of Hits**	**Number of Duplicate Article Titles**	**Number of Irrelevant Article Titles**	**Number of Articles Remaining**
Catecholamines	338	9	162	167
Hormonal response	515	15	313	186
Electrodermal Activity	369	15	165	188
Electromyography (EMG)	497	7	232	240
ECG	199	115	38	46
EEG	881	286	397	198
fNIRS	152	54	62	36
Blood Pressure	139	29	89	21
Heart rate	859	270	439	150
Heart Rate Variability	320	71	155	94
Respiration Rate	142	31	69	42
Electrooculography (EOG)	102	51	29	22
Pupillometry or pupil diameter	507	264	134	112
Index of Cognitive Activity (ICA)	384	138	194	52
Index of Pupillary Activity (IPA)	54	24	22	8
Blink Duration	104	51	28	26
Blink rate	215	119	48	50
Blink Interval or Interblink Interval	9	1	5	4
Gaze	1588	561	886	143
Fixat*	1374	526	721	136
Saccad*	957	324	527	113

Table B3. Summary of Initial Search Results, Number of Duplicate Articles, Number of Irrelevant Titles, and Remaining Articles for Performance Based CWL Measurement Techniques

CWL Measurement Technique	**Initial Search Number of Hits**	**Number of Duplicate Article Titles**	**Number of Irrelevant Article Titles**	**Number of Articles Remaining**
Accuracy	993	133	N/A	N/A
Reaction Time	747	92	N/A	N/A
Secondary Task	562	79	226	257
Subsidiary Task	345	24	212	109
Loading Task	352	26	193	159
Dual Task	571	68	266	237
Multiple Task	571	68	223	280
Peripheral Task	322	7	190	125
MRT	332	9	193	130
Task Overload	320	6	199	115
Redline	322	7	183	132

Appendix C. Frequency and Success Rates of Specific Metrics

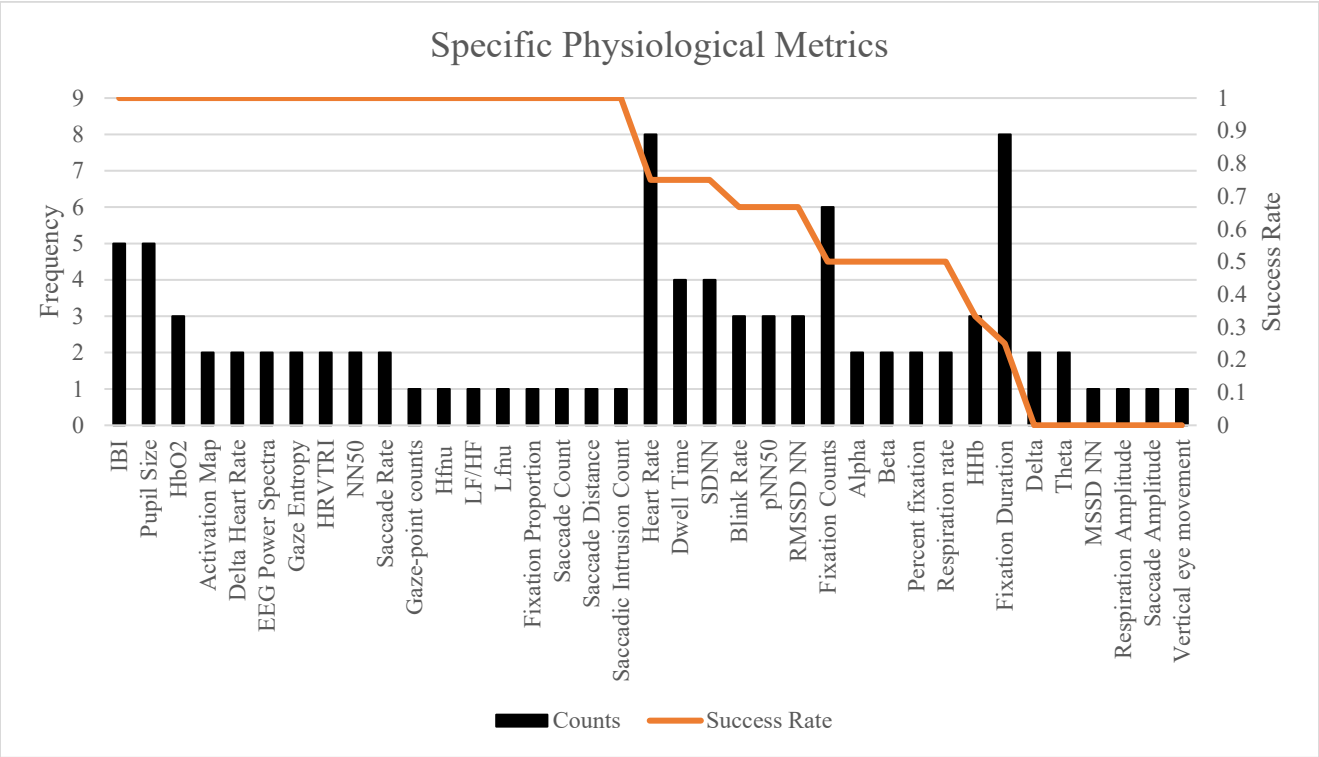

Figure C1. Frequency and success rates of specific physiological metrics utilized in the surveyed aviation articles.

Appendix D. Metrics Used in Aviation Articles

Table D1. Performance Metrics Used Within the Reviewed Aviation Articles

Article	Instructor Rating	Secondary Task Reaction Time	Path Deviation	Heading Deviation	Reaction Time	Activity Rating	Input Activity	Altitude Deviation	Speed Deviation	Secondary Task Accuracy
Dalecki et al. (2010)					X		X	X		
Dahlstrom et al. (2011)										
Kallus et al. (2011)	X				X					
Kim et al. (2011)			X							
Bezerra & Ribeiro (2012)										
Yang et al. (2013)			X							
Hancock (2013)			X		X	X				
Eichinger & Kellerer (2014)					X					
Yu et al. (2014)										
Di Stasi et al. (2015)										
Gateau et al. (2015)										
Zheng et al. (2015)										
Hsu et al. (2015)										
Cheung et al. (2015a)			X	X			X			
Cheung et al. (2015b)			X	X				X	X	
Çakır et al. (2016)										
Mansikka et al. (2016)										
Mansikka et al. (2016)	X									
Wang et al. (2016)										
Causse et al. (2017)			X							
Zhang et al. (2017)										
Muller et al. (2017)			X				X	X		
Hidalgo-Munoz et al. (2018)		X		X		X		X	X	X
Scannella et al. (2018)										
Stanton et al. (2018)										
Diaz-Piedra et al. (2019)	X									
Han et al. (2019)										
Mansikka et al. (2019a)	X									
Mansikka et al. (2019b)						X				
Zheng et al. (2019)										
Babu et al. (2019)										
Martins et al. (2019)			X							
Li et al. (2020)						X				

Table D2. Physiological Metric Methods Used Within the Reviewed Aviation Articles

Article	Interbeat Interval	Pupil Size	Heart Rate	Heart Rate Variability	Blink Dynamics	Gaze Metrics	Cerebral Oxygenation	EEG Power Spectra	Respiration
Dalecki et al. (2010)									
Dahlstrom et al. (2011)			X		X	X		X	
Kallus et al. (2011)			X	X					
Kim et al. (2011)									
Bezerra et al. (2012)									
Yang et al. (2013)						X			
Hancock (2013)									
Eichinger & Kellerer (2014)									
Yu et al. (2014)		X				X			
Di Stasi et al. (2015)								X	
Gateau et al. (2015)							X		
Zheng et al. (2015)			X						
Hsu et al. (2015)		X				X			
Cheung et al. (2015a)									
Cheung et al. (2015b)									
Çakır et al. (2016)							X		
Mansikka, Simola et al. (2016)	X		X	X					
Mansikka, Virtanen et al. (2016)	X		X	X					
Wang et al. (2016)			X		X	X			X
Causse et al. (2017)							X		
Zhang et al. (2017)						X			
Muller et al. (2017)									
Hidalgo-Munoz et al. (2018)			X	X					
Scannella et al. (2018)			X	X		X			
Stanton et al. (2018)									
Diaz-Piedra et al. (2019)						X		X	
Han et al. (2019)								X	
Mansikka et al. (2019a)	X								
Mansikka et al. (2019b)	X		X						
Zheng et al. (2019)			X		X				X
Babu et al. (2019)		X				X			
Martins et al. (2019)						X			
Li et al. (2020)		X				X			

Table D3. Subjective Metrics Used Within the Reviewed Aviation Articles

Article	NASA TLX Total	NASA TLX Subscale	Modified Cooper-Harper Score	Bedford Total Score	SWAT Total Score	3D SART	Custom Difficulty Assessment	Custom Workload Assessment	Custom Performance Assessment
Dalecki et al. (2010)									
Dahlstrom et al. (2011)							X	X	X
Kallus et al. (2011)		X						X	
Kim et al. (2011)	X								
Bezerra et al. (2012)	X	X							
Yang et al. (2013)							X		
Hancock (2013)					X				
Eichinger & Kellerer (2014)	X								
Yu et al. (2014)									
Di Stasi et al. (2015)									
Gateau et al. (2015)									
Zheng et al. (2015)				X					
Hsu et al. (2015)									
Cheung et al. (2015a)	X	X	X						
Cheung et al. (2015b)	X	X	X						X
Çakır et al. (2016)									
Mansikka et al. (2016)									
Mansikka et al. (2016)									
Wang et al. (2016)									
Causse et al. (2017)							X		
Zhang et al. (2017)									
Muller et al. (2017)	X	X							
Hidalgo-Munoz et al. (2018)									
Scannella et al. (2018)	X								
Stanton et al. (2018)	X	X		X					
Diaz-Piedra et al. (2019)	X								
Han et al. (2019)									
Mansikka (2019a)	X	X	X						
Mansikka (2019b)									
Zheng et al. (2019)	X			X					
Babu et al. (2019)									
Martins et al. (2019)	X	X				X			
Li et al. (2020)		X							

This page is intentionally blank.

U.S. Army Aeromedical Research Laboratory
Fort Rucker, Alabama

All of USAARL's science and technical
information documents are available for
download from the
Defense Technical Information Center.

https://discover.dtic.mil/results/?q=USAARL

Army Futures Command
U.S. Army Medical Research and Development Command