2016

Analyzing Cognitive Workload through Eye-Related Measurements: A Meta-Analysis

Melissa Patricia Coral
Wright State University

ANALYZING COGNITIVE WORKLOAD THROUGH EYE-RELATED MEASUREMENTS: A META-ANALYSIS

A thesis submitted in partial fulfillment

of the requirements for the degree of

Master of Science in Industrial and Human Factors Engineering

By

Melissa Patricia Coral

B.S., Wright State University, 2011

2016

Wright State University

WRIGHT STATE UNIVERSITY

GRADUATE SCHOOL

May 5, 2016

I HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER MY SUPERVISION BY **Melissa Patricia Coral** ENTITLED **Analyzing Cognitive Workload Through Eye-related Measurements: A Meta-Analysis** BE ACCEPTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF **Master of Science in Industrial and Human Factors Engineering**

____
Mary E. Fendley, Ph.D.
Thesis Director

____
Jaime E. Ramirez-Vick, Ph.D.
Chair, Department of Biomedical,
Industrial and Human Factors Engineering

Committee on Final Examination:

____
Mary E. Fendley, Ph.D.

____
Frank W. Ciarallo, Ph.D.

____
Trevor J. Bihl, Ph.D.

____
Robert E. W. Fyffe, Ph.D.
Vice President for Research and
Dean of the Graduate School

ABSTRACT

Coral, Melissa Patricia. M.S.I.H.E., Department of Biomedical, Industrial and Human Factors Engineering, Wright State University, 2016. Analyzing Cognitive Workload Through Eye-Related Measurements: A Meta-Analysis.

Understanding cognitive workload has become a vital topic for researchers in developing future systems. Existing research has investigated the use of physiological measurements of the eye with cognitive workload, though a quantitative synthesis has yet to be performed. A meta-analysis was conducted to examine the effects of cognitive workload on eye-related measurements. The objective of this meta-analysis is not to determine a difference between the levels of workload, but to identify reliable measurements. Measurements through blinks, saccades, pupils, and fixations were examined. Twenty-two studies, contributing to a total of sixty entries, met the appropriate inclusion criteria for the meta-analysis. Findings conclude the use of specific eye-related measurements as a reliable assessment of cognitive workload. Similar results obtained for moderator variables of task type and eye-tracking system did not indicate significant influences. Further research should be conducted in this domain to identify causal influences and provide an understanding for the results.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

Table | Page

ACKNOWLEDGMENTS

Foremost, I would like to express my sincere gratitude to my advisor, Dr. Mary Fendley, for her continued guidance and support throughout this research. Her passion, motivation, and patience were pivotal all the way through the completion of this study.

In addition to my advisor, I would like to show appreciation for my family; especially to my husband, Ceir M. Coral II, for his continued reassurance and encouragement, making this all possible.

I. INTRODUCTION

Cognitive workload, or the interaction between systems and tasks with the capabilities, motivation, and state of the human operator, has become an important research aspect to understand when designing and developing the systems of the future (Kramer, 1990). Understanding the state of a human operator has become a fundamental aspect commonly studied in the human-computer interaction domain. To be able to assess and predict cognitive workload relies on the availability of a known measurement linked to measuring cognitive workload. Understanding this topic is critical to the successful redesign and development of systems incorporating human operators.

Much of the cognitive workload research has investigated the use of physiological measurements, such as eye-related measurements, as a significant factor for assessing the state of the operator; however, a quantitative synthesis examining this relationship has yet to be performed. With the sufficient interest created around understanding cognitive workload in systems, it would be beneficial to perform a meta-analysis to combine those studies examining eye-related measurements with cognitive workload. A meta-analysis will allow an examination of the scope of the research domain and will provide a single estimate of the reliability and magnitude for the use of eye-related measures. This meta-analysis is intended to evaluate the effect of cognitive workload on eye-related measurements.

With the growing research involving cognitive workload and the lack of a prior synthesis already being performed exclusively on eye-related measurements, completing a meta-analysis on this topic provides both a cumulative summary of the research and a conclusive response to support its continued reliability and use in research. By including multiple eye-related measurements, this meta-analysis will attempt to further differentiate or identify those measures that have a significant link to cognitive workload. In addition to determining whether one measure is a sufficient variable for measuring cognitive workload, this analysis attempts to identify multiple useful measures to aid in future research where one measure may be more obtainable or measureable than another.

The second chapter of this thesis begins with addressing and defining the topic of cognitive workload based on its relevance and importance to the research community. By being able to discover significant measurements that are linked to understanding and evaluating an operator's state, many researchers have attempted to study the effect of cognitive workload through a variety of measurements.

These measurements can be described under three types: performance, subjective, and physiological. Research observing each of these measurement types has identified both their advantages and disadvantages; however, using physiological measurements can allow for a more objective measure of cognitive workload which can exceed most disadvantages (Endsley & Garland, 2000). One physiological system with promising connections to evaluating and predicting cognitive workload are measurements from the human eye. The use of measurements from this system are not without criticism, since it has been argued that real life situations, outside of a laboratory setting, can show diminishing values of certain measurements due to the presence of factors acting as noise like body movement and varying light conditions (Hogervorst, Brouwer, & van Erp, 2014; Wickens, Hollands, Bandury, & Parasuraman, 2013). Current research findings from studies observing different eye-related measurements are discussed in the second chapter with implications for continued use in evaluating cognitive workload. For example, EEG workload and feature saliency has been an observed topic (Laine, Bauer, Lanning, Russell, & Wilson, 2002; Noel, Bauer, & Lanning, 2005; East, Bauer, & Lanning, 2002). To aid in future research efforts and to determine and identify the most reliable eye-related measurements for measuring cognitive workload, a meta-analysis is performed.

The purpose of performing a meta-analysis for this topic is to provide a single estimate of the reliability and magnitude for a measurement of cognitive workload based on the combined results of multiple studies that observed these measurements individually. The third chapter of the thesis further discusses the background and motivation for performing a meta-analysis for this topic.

A discussion of the meta-analytic approach utilized in this thesis is also discussed in the third chapter. First, a comprehensive search of the literature was performed with keywords and search terms including workload, processing load, cognitive workload, mental workload, physiological measurements, eye, pupil, blink, fixation, pupillary response, pupillometry, and eye movement. From this literature review, a total of 57 references were considered for inclusion in the meta-analysis. This total number of references was reduced further through the evaluation of inclusion criteria; 1) a proper quantification of the independent variable of workload and the dependent variable of a measurement of eye movement, 2) a publishing date within the past 25 years, dating back to 1990, 3) sufficient statistical information to determine effect size estimates, and 4) findings presented in terms of a single eye-related measurement with cognitive workload.

After all studies were examined based on the defined inclusion criteria, a total of 22 studies, contributing to a total of 60 entries, remained and contributed to the meta-analysis.

Additional evaluations and examinations are performed on those entries included in the meta-analysis. Identifying moderator variables could be important for modeling the effect of cognitive workload on eye-related measurements. As a result, the moderator variables of individual eye-related measurements, type of task being performed to examine cognitive workload, and the system being utilized for the collection of the eye-related measurements are examined further. Finally, the third chapter concludes with a description of the specific steps performed for this meta-analysis, including the conversion procedures for calculating the effect sizes of each study included in the meta-analysis.

The fourth chapter of this thesis presents the results of the meta-analysis, referencing back to the usability of eye-related measurements to assess and predict cognitive workload. From the large, significant effect size of 0.668 achieved for the examination of the studies collectively, these results indicate that eye-related measurements would provide a reliable measurement of cognitive workload. In particular, the measurements of blink duration, rate, interval and frequency, saccade extent and peak velocity, pupil size and dilation and horizontal fixation were identified as those specific significant and reliable eye-related measurements for assessing cognitive workload. The results for the different eye-related measurements are also presented as a forest plot, providing a visual depiction of the results. The conclusions and future implications of this analysis are further discussed in the fifth and final chapter of this thesis.

II. BACKGROUND

2.1 COGNITIVE WORKLOAD

Cognitive workload involves the demands of specific tasks and the mental resources available for one to meet those demands (Wickens, 2008). It can also be identified as observed delays in information processing capabilities when a considerable amount of mental effort is exerted by an individual (Rozado & Dunser, 2015). As seen in the previous statements, many descriptions of cognitive workload exist since there is no universally accepted definition for the term. Nevertheless, through general consensus, workload can be summarized as the interaction between the structure of systems and tasks with the capabilities, motivation, and state of the human operator (Kramer, 1990). Similar to the way in which physical workload characterizes the energy demand put upon muscles, cognitive workload describes the demands of tasks, either cognitive or physical, that require the limited information processing capability of the brain (Wickens et al., 2013).

Cognitive workload has become a commonly studied concept in human-computer interaction, especially as an integral part of understanding operator state. The concept of cognitive workload is useful in explaining human performance errors in terms of overload, or when the required capacity of the information-processing system exceeds the available capacity (De Rivecourt, Kuperus, Post, & Mulder, 2008). According to De Rivecourt et al. (2008), when the operator is overloaded, a decrease in performance will be experienced. As a result, understanding cognitive workload allows for a direct comparison to the ability of an operator to sustain or reach desired performance levels (Xie & Salvendy, 2000). Additionally, understanding the cognitive state of an operator can be important for identifying instances when additional information can be presented, avoiding overloading the operator. In fact, according to Haapalainen, Kim, Forlizzi, and Dey (2010), "presenting information at the wrong time can drastically increase one's cognitive demands, can have negative impacts on task performance and emotional state, and in extreme cases, even be life threatening." Identifying cognitive workload, instances of overload, and changes in performance are pivotal for many system designs; however, specific measurements to recognize these are still being scrutinized. Given its usefulness, many efforts have been made to discover and identify those measurements of workload (Recarte, Perez, Conchillo, & Nunes, 2008).

The ability to measure workload can be pivotal in detecting and preventing situations where operator performance would be negatively affected. Recarte et al. (2008) share this importance, agreeing that having the knowledge and predictability for human information processing errors has become crucial to improve human interaction with systems involving risk. In fact, Wang, He, and Chen (2014) state that "the main purpose of workload measurement is to identify conditions for overload so that they can be avoided by design." For example, data overload is a significant problem in many systems. Many systems can require a high intake of information from an operator; however, the volume and changing rate of data quickly surpass an operator's ability to gather and understand the data (Endsley, 2012). By understanding overload conditions, these systems can be redesigned to reduce overload. Thus, having the ability to assess and predict workload has become an important topic to consider for designing new systems, modifying existing systems, and through task reallocation or adaptive automation by avoiding task overload (Van Orden, Limbert, & Makeig, 2001).

2.2 COGNITIVE WORKLOAD MEASUREMENTS

Research performed to date has shown that cognitive workload can be assessed under three measurement types: performance, subjective, and physiological. Subjective measurements are based on judgments of the operators in terms of the workload associated with the performance of a task or a system function. Performance measurements assess workload through the ability of an operator to perform tasks or functions of a system. Physiological measurements evaluate the physiological responses of the operator with the system or task demands (Wierwille & Eggemeier, 1993). Physiological measurements are used to evaluate cognitive workload based on the assumption that with an increase in task demands, noticeable changes in various physiological systems can be observed (Stanton, Salmon, Walker, Baber, & Jenkins, 2005).

The selection for use of one or more of these measurement types can depend on several factors, one being the use in a particular application. For instance, there are certain properties that are recommended for use in test and evaluation applications. These properties include sensitivity, intrusion, diagnosticity, global sensitivity, transferability, and implementation requirements (Wierwille & Eggemeier, 1993). Similar to test and evaluation applications, any application with the need for cognitive workload evaluation and measurement will expect to utilize those measurement techniques that have proven a relationship to cognitive workload.

Research using different measurements within these three categories have identified many to be significantly linked to measuring cognitive workload. Using subjective measurement techniques has the advantages of ease of use and low cost for the researchers (Endsley, 1995). They can also be multidimensional and have the capability to permit some predictive assessment of the workload connected to proposed systems and designs (Wierwille & Eggemeier, 1993). Performance measures hold the advantages of being objective and usually nonintrusive (Endsley, 1995; Stanton et al., 2005). However, these measurement types also have some disadvantages. For instance, De Rivecourt et al. (2008) found that with subjective measures, "participants are having difficulties distinguishing task demands from invested effort." Also, there is the opportunity to experience critical information loss when there is a long delay between the operator's subjective ratings and completion of the task (Wierwille & Eggemeier, 1993).

Only providing indirect insights about cognitive workload are additional disadvantages for both performance and subjective measurements. Furthermore, performance techniques with primary task measures have limitations in regards to the varying levels of workload, by sometimes being insensitive to distinctions at low and moderate levels of demand. This occurs due to the operator's ability to expend extra processing resources to meet the increased demands at these levels of workload (Wierwille & Eggemeier, 1993). For insights about cognitive workload, physiological measurements provide direct measurements over time, identifying these measurements as potentially being more practical and unbiased compared to performance and subjective measurements (Rozado & Dunser, 2015). Previous associations between cognitive workload, or more specifically high cognitive workload, and physiological measurements have been identified through such aspects as increased cognitive processing, increased arousal and increased energy demand (Hogervorst et al., 2014).

Physiological measures are not without their share of disadvantages. Techniques to record these measurements are substantially more expensive than those for performance and subjective measures and a larger problem exists for discriminating between signal and noise for these measurements compared to performance and subjective measures (Kramer, 1990). Still, the strengths of physiological measures can far outweigh the disadvantages. Such strengths include the ability to record a measurement in the absence of behavior, and to provide measures that respond quickly to shifts in workload. As well, these measures are relatively unobtrusive and are multidimensional (Kramer, 1990). Using physiological measurements can allow for a more objective measure of cognitive workload (Endsley & Garland, 2000). Also, by using physiological measurements, systems can address the need for in-the-moment, automatic assessments of cognitive workload; this includes being able to evaluate workload even when no change in task performance can be detected (Haapalainen et al., 2010). In other words, these types of measurements are often more attractive as an assessment approach since they can be obtained without an intervention by a subjective response or through a transformation of a performance response (Marquart, Cabrall, & de Winter, 2015). The advantages and disadvantages of these three measurement types are summarized in Table 1.

Table 1. *Advantages and Disadvantages of Cognitive Workload Measurements*

Measurement Type	Advantages	Disadvantages
Performance	Objective	Indirect insights about workload
	Nonintrusive	Insensitive sometimes to variations in workload at low to moderate levels of demand
Subjective	Low cost	Difficulties distinguishing task demands from invested effort
	Ease of use	Indirect insights about workload
	Multidimensional	
	Predictive assessment capabilities	Loss of critical rating information with extended delays
Physiological	Direct measurements over time	Difficulty discriminating between signal and noise.
	Recorded in the absence of behavior	
	Unobtrusive	
	Responds quickly to shifts in workload	Expensive recording techniques
	Multidimensional	
	Unbiased	

Even while considering the advantages and disadvantages of the different techniques of evaluating and measuring cognitive workload, it is not atypical for studies investigating this to incorporate the use of more than one measurement within two or more techniques (Brookings, Wilson, & Swain, 1996; Recarte et al., 2008; Di Stasi, Antoli, Gea, & Canas, 2011; Engstrom, Johansson, & Ostlund, 2005; Bommer & Fendley, 2015). According to Cegarra and Chevalier (2008), there are no methods that can evaluate and measure cognitive workload alone. Instead, with the inclusion of measurements under different techniques, the validity and reliability for identifying cognitive workload would amplify. Before the decision on what those measurements to be utilized should include, it is important to justify its relationship to cognitive workload.

2.3 EYE-RELATED MEASUREMENTS OF COGNITIVE WORKLOAD

Physiological measurements encompass those obtained through the different systems of the human body. One discipline under these physiological measurements is that of ophthalmic physiology. The study of eye movements and eye tracking research actually pre-dates the use of computers but it did not begin to thrive until the 1970s due to advances in technology for eye tracking and the development of a physiological theory linking eye tracking data to cognitive processes. This research only continued to evolve with technological advances and became a means of human-computer interaction (Jacob & Karn, 2003). Most recently, this technology involves the use of video recordings of the eye in real time from high speed cameras placed either on a headband or a computer monitor. Through these means, data can be collected in any environment without interfering with an operator's task performance (Marshall, 2007).

By observing an operator's eye and head movements, researchers have the use of a non-intrusive tool to understand how the mind acquires and processes visual information (Yang, McDonald, & Zheng, 2012; Holmquist et al., 2011; Poole & Ball, 2005). Past research also provides an argument that cognitive processes such as reading, visual search, and problem solving can be studied based on the relationship between the behavior of the operator's eyes and cognition (Maier, Baltsen, Christofersen, & Storrle, 2014; Kahneman, Beatty, & Pollack, 1967). With the use of eye-tracking technology, researchers have a more objective measurement of a user's cognitive workload through eye movements and pupillary responses (Buettner, 2013).

Compared with other physiological measurements, there are many benefits associated with the use of eye movements in adaptive systems, and thus in identifying cognitive workload. Benefits identified include, insensitivity to limb movements, including being adjusted for head movements, and the equipment required for observing and recording eye movements does not require extensive amounts of training to setup and the calibration procedure can be completed rather quickly (Di Nocera, Camilli, & Terenzi, 2007). In addition, Kahneman (1973) states that "a useful physiological measurement for mental effort should be sensitive to both between-tasks and within-task variations." Eye-related measurements meet these criteria.

Many studies have previously researched the relationship between cognitive workload and eye-related measurements, with some measurements studied more frequently than others. Some of the eye-related measurements are related to eye blinks; these include blink rate, blink duration, and blink latency. Other measurements are characterizations of eye fixations including the number of fixations, fixation duration, saccadic duration, saccadic peak velocity, and gaze distribution. One of the most commonly studied measurement is of the pupil diameter, also referred to as pupillometry (Marquart et al., 2015).

The diameter, or size, of the pupil has often been observed and evaluated. According to Hess and Polt (1964), the "pupil response not only indicates mental activity in itself but shows that mental activity is closely correlated with problem difficulty, and that the size of the pupil increases with the difficulty of the problem." Changes in pupil diameter have previously been interpreted as indicators of second-to-second variation in the amount of workload imposed by a task (Kahneman et al., 1967). Measurements involving the pupil, however, are not without the most criticism as the largest changes in the pupil can occur in response to other factors than cognitive workload. Some of the main functions of the pupil occur outside of the amount of mental stimulation, such as in changes in the amount of light that enters the eye or a shift in the fixation from a far to a near object (Kramer, 1990).

Research focusing on pupil dilations has shown that they occur at short latencies following the onset of a task and subside quickly once the task is completed. More importantly, the magnitude of the pupillary dilation appears to be a function of processing load, or the mental effort required to perform the cognitive task (Iqbal, Zheng, & Bailey, 2004; Beatty & Kahneman, 1966; Beatty, 1982). In addition to cognitive tasks, pupillary changes have also been found to be sensitive to perceptual and response related demand tasks (Kramer, 1990). A specific pupil reaction known as the task-evoked pupillary response has been repeatedly associated with a variety of cognitive processes that are linked to cognitive load (Klinger, Kumar, & Hanrahan, 2008). One constraint discovered for pupillary dilation involves the limits of information-processing capacity of the operator. Once these limits are exceeded, any additional increases involved with task demands no longer yield an increase in pupillary dilation (Beatty, 1982).

Blinking has also being linked to certain cognitive processes. Holland and Tarlow (1975) found the rate of blinking to be significantly reduced during processing of information in memory. Indications of the relationship between blink rate and cognitive processes even dates back to Telford and Thompson (1933), who showed that blink rate was reduced during tasks that required concentration and intense mental stimulation. Blink rate is more sensitive to cognitive workload through task difficulty than other eye-related measurements in demanding visual tasks (Brookings, Wilson, & Swain, 1996). The relationship between blink rate and task demands is often attributed to an operator's attempts to minimize the possibility of missing important information (Fogarty & Stern, 1989). Other studies have also led to discoveries such as a decrease in blink rate with increases in cognitive demand. Similarly, blink duration shows a tendency to decrease while experiencing increases in visual demand (Wierwille & Eggemeier, 1993).

Continued research has provided other measurements from the eye with a relationship to cognitive workload. For instance, dramatically different results can be obtained when even minor changes are made in the parameters defining a fixation (Jacob & Karn, 2003); however, dwell time and fixation duration are generally believed to increase with an increase in cognitive workload (Marquart et al., 2015). Recent research has also indicated that the size of the functional visual field decreases with increasing task difficulty (Young & Hulleman, 2013).

As previously discussed, different eye-related measurements have been observed to either increase or decrease as cognitive workload increases. A summary of those eye-related measurements observed in this meta-analysis and their currently understood indication of increased cognitive workload are summarized in Table 2.

Table 2. *Summary of Eye-related Measurements and their Relationship to Increased Cognitive Workload*

Indicator of Increased Cognitive Workload	
↑	**Blink Duration**
↑	**Blink Interval**
↑	**Blink Frequency**
↑	**Saccade Rate**
↑	**Saccade Peak Velocity**
↑	**Saccade Amplitude**
↑	**Pupil Size**
↑	**Pupil Dilation**
↑	**Fixation Frequency**
↑	**Fixation Duration**
↑	**Horizontal Fixation**
↑	**Vertical Fixation**
↑	**Mean Dwell Time**
↓	**Saccade Extent**
↓	**Blink Rate**
↓	**Area of Visual Field**

Being able to understand the relationship between cognitive workload and eye-related measurements could aid in attaining greater reliability in detecting operator cognitive states; which, according to Rozado and Dunser (2015), would "lead the way to better and more robust systems for direct, real-time measurement of cognitive workload, supporting better human-computer interaction and achieving greater user satisfaction." Given this importance and the abundance of research performed individually, it would be beneficial to the research community to combine and summarize these individual studies through the technique of meta-analysis.

III. META-ANALYSIS

3.1 META-ANALYSIS METHODOLOGY

Meta-analysis is a technique that provides a single estimate of the reliability and magnitude of an effect, either supporting or refuting a given hypothesis, based on the combined results of multiple studies that observed a given hypothesis (Horrey & Wickens, 2004; Cooper, 2010; Rosenthal & DiMatteo, 2001; Hall & Rosenthal, 1995; Borenstein, Hedges, Higgins, & Rothstein, 2009). The methodology for a meta-analysis requires an extremely thorough search for relevant studies and performing a careful review and analysis; thus, preventing reliance on the results of a single study or review when attempting to understand a specific phenomenon (Rosenthal & DiMatteo, 2001). This past reliance was in part due to the statistical significance of a finding being the only information reported in the literature.

It has been this focus that has often misled researchers and is why a meta-analysis typically focuses on effect sizes. In fact, Rosenthal and DiMatteo (2001) state "meta-analysis prevents our reliance on the significance test of any one finding as a measure of its value and helps us realize that repeated results in the same direction across several studies, even if not one is significant, are much more powerful evidence than a single significant result." Methods of reporting statistical results in any analysis are facing scrutiny, with many issues that exist when relying on p-values, especially for comparisons (Bihl, Bauer, Temple, & Ramsey, 2015; Halsey, Curran-Everett, Vowler, & Drummond, 2015). For example, there is currently disagreement in statistics literature on the appropriateness for using *p*-values due to issues involving its incorrect application (Bihl, 2015). The American Statistical Association has even released a statement addressing the misconceptions and misuse of the *p*-value (Wasserstein & Lazar, 2016). Using effect sizes, or the measures of the magnitude of an effect, allow researchers to determine estimations of differences between groups or the strength of associations between different variables (Nakagawa & Cuthill, 2007). This means that the effect size refers to the degree to which the phenomenon is present in the population, with a larger effect size meaning a greater degree of manifestation (Cooper, 2010).

Cooper (2010) summarizes the meta-analysis process into seven steps; 1) formulating the problem, 2) searching the literature, 3) gathering information from studies, 4) evaluating the quality of studies, 5) analyzing and integrating the outcomes of studies, 6) interpreting the evidence, and 7) presenting the results. Herein, we apply this step-by-step methodology to conduct a meta-analysis of the use of eye-related measurements for cognitive workload. The step-by-step methodology for this meta-analysis is summarized in Figure 1.

According to Rosenthal and DiMatteo (2001), a meta-analysis has many advantages including being able to see the scope of a research domain, keeping statistical significance in perspective, minimizing wasted data, and asking focused research questions. Cooper (2010) states that "a topic is probably not suitable for research synthesis unless it already has created sufficient interest within a discipline or disciplines to have inspired enough research to merit an effort at bringing it all together" (p. 23). The meta-analysis technique is utilized more routinely today when there is any size of research literature addressing a common hypothesis (Hall & Rosenthal, 1995). It allows for both formulating potential causal influences and trying to understand why the various results occurred (Rosenthal & DiMatteo, 2001). Although the idea of performing a quantitative synthesis of research seems immense, Hall and Rosenthal (1995) reiterate that using simple and straightforward techniques that are easily executed, described, and understood can deal with the research questions raised in the meta-analysis.

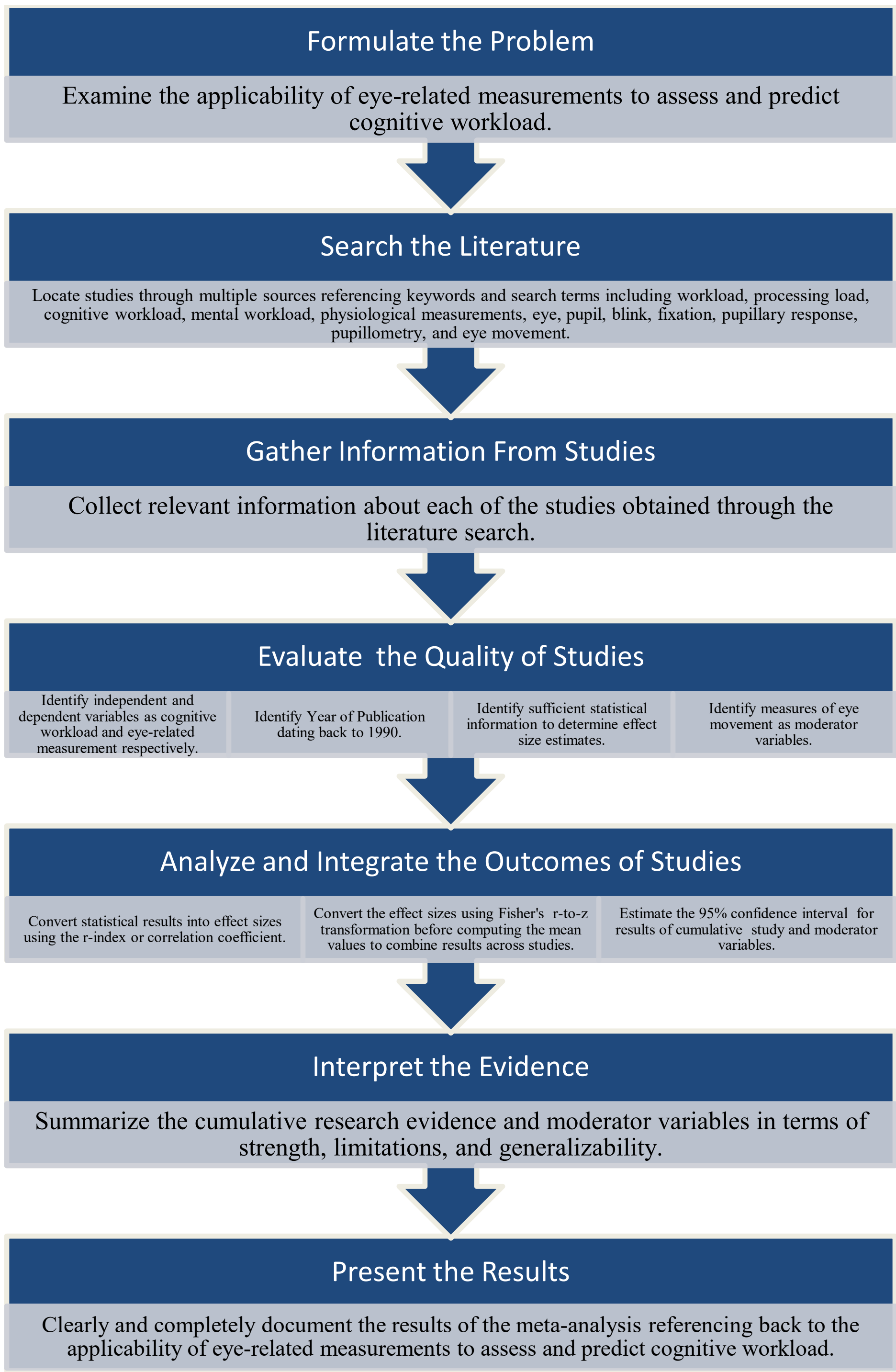

Figure 1. Step-By-Step Meta-Analysis Methodology

3.2 META-ANALYTIC APPROACH

A comprehensive search of the literature was performed first to discover those relevant studies to be included in the analysis. Keywords and search terms included workload, processing load, cognitive workload, mental workload, physiological measurements, eye, pupil, blink, fixation, pupillary response, pupillometry, and eye movement. The literature search focused on sources including library databases, such as PsycINFO and IEEE Xplore, and the web-based search engines of Google and Google Scholar. Together, these different sources allowed a comprehensive search for any references with potential relevance to be included in the analysis. These included journal articles and conference proceedings. Once the primary search was complete, backwards referencing, or a review of the reference lists for all obtained studies, was performed to determine whether any related studies could be included. The result of this stepped literature review comprised of a total of 57 references considered for inclusion in the meta-analysis.

Once the comprehensive search of the literature was completed, each study was examined based on different criteria predetermined for study inclusion. The inclusion criteria included a proper quantification of the independent variable of workload and the dependent variable of a measurement of eye movement. With the advancements in technology and understanding for recording eye movement measurements, this meta-analysis focused on studies published within the past 25 years, dating back to 1990. Eye measurements collected before this time involved not only a large effort with data collection, but even more so with data analysis, where spending days processing data that only took minutes to collect was not uncommon (Jacob & Karn, 2003). Additionally, any prospective study must also have included sufficient information to determine effect size estimates; from Rosenthal and DiMatteo (2001), certain effect sizes cannot be computed from kappa, percent agreement, relative risk, risk difference, or the odds ratio unless the raw data is available. Since main effects are most often the focus of a meta-analysis (Cooper, 2010), the findings must be presented in terms of a single eye-related measurement with cognitive workload and not a fusion of two or more measurements. This allowed direct interpretation for singular eye measurements.

Since prior meta-analyses have been performed in different disciplines that included multiple studies using different factor levels (Uttal et al., 2013; Glioma Meta-analysis Trialists (GMT) Group, 2002), those studies which analyzed outcomes using different factor levels were included in this meta-analysis. Furthermore, since the objective of this meta-analysis is not to determine a difference between the varying levels of cognitive workload, but to identify reliable measurements capable of predicting cognitive workload, it is not imperative that any ordinal aspect of data is captured; thus, including estimates of effect sizes based on the F ratios for multiple conditions from the studies was allowed. Although this concept may differ from many meta-analyses previously performed, there is no particular statistical method defined for this "analysis of analyses" (Onnasch, Wickens, Li, & Manzey, 2014). Additionally, it was not necessary to identify the data type of the variable examined in the studies for inclusion since a meta-analysis can be performed using dichotomous, continuous, and ordinal variables (Higgins & Green, 2008). In the end, studies that did not meet the established inclusion criteria were removed from the analysis. As mentioned by Schaefer, Chen, Szalma, and Hancock (2016), the process of rejecting studies for inclusion in the meta-analysis is both common and necessary. This procedure ensures that meaningful results are achieved when combining effect sizes from multiple studies.

From the original set of research papers, a total of 22 studies remained and contributed to the meta-analysis. Those remaining studies and their attributes are summarized in Table 3. From the studies included, we can identify a selection of 72 unique authors with research published in 16 unique journals contributing to the research on cognitive workload with eye-related measurements. Some of these studies analyzed several relevant eye-related measurements as dependent variables, leading to multiple effect sizes. With multiple effect sizes, these can be used individually in an analysis of subgroups or in examination of moderating variables (Rosenthal & DiMatteo, 2001). Because of this the overall number of entries for the meta-analysis increased to 60.

Table 3. *List of Studies and their Attributes Included in the Meta-Analysis.*

Reference	Eye Measure(s) Studied	Participants Studied / Task Type	Levels of Workload	Real World vs. Laboratory Setting	System Used
Backs, R. W., & Walrath, L. C. (1992)	Pupil Dilation	8 participants, control vs. search task	High vs. no cognitive load	Laboratory	Applied Science Laboratories Eye View Monitor and TV Pupillometer System model 1994-S
	Blink Duration				
Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A., & Montanari, R. (2011)	Blink Duration	15 drivers, single- and dual-tasks, primary (Lane Change Test) and secondary (IVIS) tasks	Baseline, dual-task, control	Laboratory	SMI iView HED head-mounted monocular eye-tracker
	Blink Rate				
	Average Pupil Size				
Brookings, J. B., Wilson, G. F., & Swain, C. R. (1996)	Blink rate	8 subjects, simulated air traffic control tasks	High, medium, low	Computer-based air traffic control simulation	Psycho-physiological Assessment Test System
	Saccade Rate				
	Saccade Amplitude				

Reference	Eye Measure(s) Studied	Participants Studied / Task Type	Levels of Workload	Real World vs. Laboratory Setting	System Used
De Rivecourt, M., Kuperus, M., Post, W., & Mulder, L. (2008)	Mean dwell time	19 pilots, instrument flight task	4 levels	ALSIM AL 100 Flight Trainer	Jazz Synchronic system Version RS-232
	Fixation Duration				
Di Nocera, F., Camilli, M., & Terenzi, M. (2007)	Eye Fixations	10 Pilots, flight simulation	High workload (departure and landing) vs. low to moderate workload (climb, cruise, and descent)	Microsoft Flight Simulator 2004	Tobii ET17 eye-tracking system
May, J. G., Kennedy, R. S., Williams, M. C., Dunlap, W. P., & Brannan, J. R. (1990)	Saccadic eye movement	5 subjects, tone-counting tasks	High, medium, low	Laboratory	Infrared eye-tracking instrument (Eye Trac, Model 106)
	Saccadic extent	10 subjects, levels of a visual counting task			
	Saccadic extent	10 subjects, tone counting task			
Niezgoda, M., Tarnowski, A., Kruszewski, M., & Kamiński, T. (2015)	Blink Rate	46 drivers, primary task - driving in traffic, secondary task - delayed digit recall task	n-back test, three levels (0-back, 1-back, 2-back)	AutoSim AS 1200-6 driving simulator	Mobile eye-tracking device, SMI glasses
	Pupil Size Diameter				
	Fixation Durations				
	Fixation location on the vertical axis				
	Fixation location on the horizontal axis				
Pomplun, M., & Sunkara, S. (2003)	Pupil Size	10 participants, recall tasks	High, medium, low	Laboratory	EyeLink-II System
Rantanen, E., & Goldberg, J. (1999)	Area of Visual Field	13 subjects, tone counting tasks	high, moderate, none	Laboratory	Goldmann perimeter
Recarte, M. Á., Pérez, E., Conchillo, Á., & Nunes, L. M. (2008)	Pupil dilation	29 participants, cognitive tasks (listening, talking , or calculating) with visual detection or no visual detection.	Single- and dual-task (cognitive and visual task)	Laboratory	ASL 5000 eye-tracking system
	Blink Rate				

Reference	Eye Measure(s) Studied	Participants Studied / Task Type	Levels of Workload	Real World vs. Laboratory Setting	System Used
Reyes, M. L., & Lee, J. D. (2008)	Fixation Duration Horizontal Fixation Position Vertical Fixation Position Gaze Concentration (dwell) Saccade Duration Saccade Speed Saccade Distance	12 participants, In-vehicle information system (IVIS) task	High, low, baseline	DriveSafety Research Simulator	Seeing Machines' faceLAB eye tracking system (version 4.1)
Rosenfield, M., Jahan, S., Nunez, K., & Chan, K. (2015)	Blink rate	16 subjects, Reading text from different methods (tablet or hard copy)	Low vs. high workload	Laboratory	Videotaped using Kodak EasyShare M853 zoom digital camera
Ryu, K., & Myung, R. (2005)	Blink interval	10 subjects, tracking tasks	Low, medium, high speed	Laboratory	EOG was recorded using sternal leads.
Savage, S. W., Potter, D. D., & Tatler, B. W. (2013)	Fixation durations Saccade Amplitude Saccade peak velocity Horizontal spread of fixation positions Vertical spread of fixation positions Blink frequency Blink duration	17 participants, inclusion of a puzzle to complete during the trial	High vs. no cognitive load conditions	Laboratory	EyeLink 1000 eye-tracker
Steinhauer, S. R., Condray, R., & Kasparek, A. (2000)	Pupil Diameter (extent of constriction)	33 subjects, arithmetic task	High vs. no load	Laboratory	ISCAN, Inc., Model RK-406 pupillometer
Steinhauer, S. R., Siegle, G. J., Condray, R., & Pless, M. (2004)	Pupil Diameter	22 subjects, arithmetic task	High vs. low	Laboratory	ISCAN, Inc., Model RK-406 Pupillometer
Tokuda, S., Obinata, G., Palmer, E., & Chaparro, A. (2011)	Saccadic intrusions Pupil dilation	16 participants, dual task: auditory N-back task and a free-viewing task	Low, medium, and high mental workload conditions	Laboratory	Tobii 1750 eye tracker

Reference	Eye Measure(s) Studied	Participants Studied / Task Type	Levels of Workload	Real World vs. Laboratory Setting	System Used
Van Orden, K. F., Limbert, W., Makeig, S., & Jung, T. (2001)	Blink frequency Blink duration Fixation frequency Fixation duration Saccadic extent Mean pupil size	11 participants, mock air warfare task	Target Density, 9 levels	Laboratory	Applied Sciences Laboratory SU4000 eye-tracking system
Veltman, J. A., & Gaillard, A. K. (1996)	Blink duration Number of Blinks	14 subjects, primary and secondary task, simulated flight tasks	6 segments of varying task difficulty in flight task	Flight Simulator	CODAS system
Wang, Q., Yang, S., Liu, M., Cao, Z., & Ma, Q. (2014)	Fixation Count Fixation Duration	42 subjects, Online shopping tasks	Simple vs. Complex tasks	Laboratory	Hi-Speed iView X eye-tracker
Wilson, G. F., Fullenkamp, P., & Davis, I. (1994)	Blink Rate Blink duration	7 pilots, flight tasks	5 flight tasks	Laboratory	Del Mar Avionics miniature physiological data recorders
Yang, Y., McDonald, M., & Zheng, P. (2012)	Fixation on touch screen Blink Rate Saccades	41 Drivers, driving with touch screen tasks	Three task levels of difficulty and baseline	Real-road driving environment	FaceLAB Eye Monitoring System

When performing a meta-analysis, it is also important to compare the variation in the observed effect sizes with the variation expected from sampling errors (Cooper, 2010). By testing the homogeneity of effect sizes, a calculation for the probability of only sampling error having caused the variation between the observed and expected effect sizes can be performed. This test allows a conclusion for any variation in effect sizes to be explained by sampling error, or chance, and prevents the need for additional analyses to be performed within the meta-analysis. According to Cooper (2010), the explanation of sampling error is the simplest explanation for a difference in effect sizes to occur. If it is determined that there is a greater variability in effect sizes than by sampling error alone, an analysis should be completed to examine whether different characteristics within the study could be associated with the variance in effect sizes. Although, many meta-analysts will still examine any potential moderator variables, with or without identifying sampling error as a plausible cause in variation, when theoretical or practical reasons can be recognized (Cooper, 2010). This meta-analysis took the approach of examining moderator variables regardless of the result of the homogeneity analysis.

The examination of moderator variables adds to theory development and increases the richness of empirical work (Rosenthal & DiMatteo, 2001). Identifying these moderator variables could be important for modeling the effect of cognitive workload on eye-related measurements. With the inclusion criteria already taking into account the year of publication, it may be additionally beneficial to identify the eye-related measures as moderating variables to further analyze which eye-related measurements are related to identifying cognitive workload. This analysis would be similar to examining each of the dependent variables to further identify those most and least affected by varying levels of cognitive workload (Rosenthal, 1994). This may be important since some eye movement measurements have been observed to increase under higher cognitive workload while other measurements have been observed to decrease under higher cognitive workload.

A second moderator variable to examine was the type of task used in each study. Some tasks previously linked to cognitive workload include "short and long-term memory access, mental arithmetic, sentence comprehension, vigilance, and visual and auditory perception tasks" (Klinger et al., 2008). These task types can be classified as simplistic tasks when being compared to the application type tasks performed in real or simulated expert-driven tasks. It would be important for future modeling and design to determine if there is an influence on the effect of cognitive workload with eye-related measurements between the different types of tasks performed by an operator.

A final moderator variable examined includes the relationship between the eye tracking system and the operator. Prior research utilizing eye trackers to estimate a user's cognitive workload used only head-mounted cameras; although this method provided high precision for results, it also proved to be distracting and burdensome to the operators (Klinger et al., 2008). The development of remote eye trackers, which use display-mounted cameras, provide an attractive alternative to the head-mounted cameras. A remote eye tracker allows for a less obtrusive measurement of a user's eye movements. It is a familiar environment for subjects because it uses a computer system resembling a standard desktop computer. Unfortunately, remote eye trackers are subject to a greater amount of measurement noise compared to head-mounted systems (Klinger et al., 2008). The need to limit the physical relationship between the system and the operator continues to be one of the most significant obstacles to address before widespread use of eye tracking devices in system designs (Jacob & Karn, 2003). By investigating the type of eye tracking system utilized, we can further evaluate the relative efficiency and reliability of head- versus display-mounted systems.

3.3 COMPUTATIONS FOR META-ANALYSIS

For this meta-analysis, the statistical results from the included studies were converted into effect sizes. With the use of effect sizes, it is possible to "compare the magnitude of experimental treatments from one experiment to another" (Thalheimer & Cook, 2002). Using effect sizes answers the question "how much?" instead of the test for significance answering yes or no (Cooper, 2010). In fact, From Horrey and Wickens

(2004), "effect sizes are advantages because they focus on how large a particular effect is (as opposed to whether or not it differs from zero)." In other words, whereas statistical tests of significance are informative for the likelihood that the results from an experiment differ from chance expectations, the use of effect sizes focuses on the relative magnitude of the experimental treatment, or the size of the experimental effect. In general terms, effect sizes are calculated as the difference between treatment means divided by the standard deviation of the conditions (Thalheimer & Cook, 2002).

From Cooper (2010), one of the steps in a meta-analysis involves determining the effect size metric to utilize. According to Rosenthal and DiMatteo (2001), the product moment correlation (*r*) as a measure of effect size has a number of advantages over other measures such that it is more easily interpreted in terms of practical importance than are Cohen's *d* or Hedges' *g*. In particular, the product moment correlation, also referred to as Pearson's *r*, can quantify the strength of relationships and not just the size of the experimental effects (Nakagawa & Cuthill, 2007). In addition, the use of correlational effects, such as Pearson's *r*, will allow a representation of an association between eye-related measurements and cognitive workload (Schaefer et al., 2016). Using Cohen (1992), we can compare effect size *r* results to the known benchmarks of 0.10 as small, 0.30 as medium, and 0.50 as large effect sizes.

The effect size r was calculated using the test statistics reported in each study. The statistical results were converted into effect size r, using the conversions and procedures described in Rosenthal and DiMatteo (2001), Cohen (1988), Rosenthal (1984) and Rosnow and Rosenthal (1996). It should be noted that $t_0^2 \approx F_0$ for simple linear regression, where t-tests from comparing two means and an F-test from analysis of variance are equivalent and supported by Cochran's theorem for this special case (Kutner, Nachtsheim, Neter, & Li, 2005). This approach is based on the fact that the square of a *t* random variable with *v* degrees of freedom is an F random variable with 1 numerator and *v* denominator degrees of freedom (Montgomery, 2013; Kutner et al., 2005). The test statistics (*t*- or *F*-values) were converted into effect sizes using the computations provided. From Rosenthal and DiMatteo (2001), the computation for effect size r from a t statistic is:

$$r = \sqrt{\frac{t^2}{t^2 + df}} \quad . \tag{1}$$

Also, Rosenthal and DiMatteo (2001) identify the computation for effect size r from an *F* statistic with 1 degree of freedom (df) in the numerator as:

$$r = \sqrt{\frac{F}{F + df_{\text{error}}}} \quad . \tag{2}$$

To calculate the effect size r from an F statistic with different number of levels, we need to first perform and obtain additional information from the statistical results. From Cohen (1988), we can define f^2, or a ratio of variances, as

$$f^2 = \frac{SS_{\text{treatment}}}{SS_{\text{error}}} \quad . \tag{3}$$

With this value we can directly compute the power, η^2, via

$$\eta^2 = \frac{f^2}{1 + f^2} \quad . \tag{4}$$

To obtain the same relationship above, the following equation from Rosenthal (1984) can be converted as below.

$$F = \frac{\eta^2}{1 - \eta^2} \frac{df_{\text{error}}}{df_{\text{treatment}}} \tag{5}$$

$$F \frac{df_{\text{treatment}}}{df_{\text{error}}} = \frac{\eta^2}{1 - \eta^2} \tag{6}$$

$$F = f^2 \frac{df_{\text{error}}}{df_{\text{treatment}}} \tag{7}$$

$$f^2 = \frac{\eta^2}{1 - \eta^2} \tag{8}$$

Using Cohen (1988), we can convert f to its d estimate

$$d = 2f \tag{9}$$

and since both r and d estimates can be readily converted to one another (Rosenthal & DiMatteo, 2001), we can convert our d value to our effect size r as follows:

$$r = \sqrt{\frac{d^2}{d^2 + 4}} \quad . \tag{10}$$

For studies only containing p values, it is possible to convert the p value to its associated one-tailed standard normal deviate Z and use the following conversion equation from Rosenthal and DiMatteo (2001) to convert to effect size r.

$$r = \frac{Z}{\sqrt{N}} \tag{11}$$

However, if only a range is given, the following one-tailed standard normal deviate Z can be used: $p < 0.05$, Z = 1.645; $p < 0.01$, Z = 2.326; and $p < 0.001$, Z = 3.090 (Rosenthal & DiMatteo, 2001).

These different conversions and procedures were utilized to calculate the effect size from the statistical results for each study included in the meta-analysis. A summary of the statistical results and the computed r values for each study in the meta-analysis is provided in Table 4.

Table 4. *Statistical Results and Computed Effect Sizes*

Reference	Eye Measure(s) Studied	Statistical Results	Effect Size (r)
Backs & Walrath, (1992)	Pupil Dilation	F(1,6) = 27.46, p<0.01	0.9059
	Blink Duration	F(1,6) = 12.51, P<0.05	0.8221
Benedetto et al., (2011)	Blink Duration	F(2,28) = 4.78, p< 0.05	0.5045
	Blink Rate	No significant results were obtained for blink rate.	0.0000
	Average Pupil Size	F(2,28) = 33.27, p<0.001	0.8389
Brookings et al., (1996)	Blink rate	F(2,14) = 9.37, p<0.01.	0.7566
	Saccade Rate	Saccade measures were not significant.	0.0000
	Saccade Amplitude	Saccade measures were not significant.	0.0000
De Rivecourt et al., (2008)	Mean dwell time	(p < 0.001)	0.3544
	Fixation Duration	(p < 0.001)	0.3544
Di Nocera et al., (2007)	Eye Fixations	F(4,36) = 25.85, p<.0001	0.8614
May et al., (1990)	Saccadic eye movement	F(2,8) = 4.22, p=0.056	0.7165
	Saccadic extent	F(2,18) = 16.06, p<0.0001	0.8005
	Saccadic extent	F(2,18) = 5.49, p=0.026	0.6155
Niezgoda et al., (2015)	Blink Rate	F(2.63,118.16) = 2.96, p<0.05	0.2490
	Pupil Size Diameter	F(2.52,101.36) = 71.31, p<0.01	0.7994
	Fixation Durations	F(2.25,101.33) = 3.66, p<0.05	0.9970
	Fixation location on the vertical axis	F(2.66,119.70) = 33.98, p<0.01	0.6557
	Fixation location on the horizontal axis	F(3,135) = 8.24, p<0.01	0.3937
Pomplun & Sunkara, (2003)	Pupil Size	F(2,18) = 35.13, p<0.001	0.8922
Rantanen & Goldberg, (1999)	Area of Visual Field	F(2,39) = 15.21, p<0.001	0.6620
Recarte et al., (2008)	Pupil dilation	Single-task, F(3,84) = 78.93, p=0.000 and Dual-task, F(3,84) = 51.49, p=0.000.	0.8591 0.8050
	Blink Rate	Single-task, F(3,84) = 42.66, p=0.000 and Dual-task, F(3,84) = 4.01, p=0.010.	0.7772 0.3536

Reference	Eye Measure(s) Studied	Statistical Results	Effect Size (r)
Reyes & Lee, (2008)	Fixation Duration	F(2,16) = 2.9, p=0.08	0.5158
	Horizontal Fixation Position	F(2,16) = 4.0, p= 0.038	0.5774
	Vertical Fixation Position	F(2,16) = 24.1, p<0.0001	0.8665
	Gaze Concentration (dwell)	F(2,16) = 69.00, p<0.0001	0.9466
	Saccade Duration	F(2,16) = 60.00, p<0.0001	0.9393
	Saccade Speed	F(2,16) = 64.10, p<0.0001	0.9429
	Saccade Distance	F(2,16) = 132.8, p<0.0001	0.9712
Rosenfield et al., (2015)	Blink rate	F(1,30) = 3.87, p=.05	0.3376
Ryu, & Myung, (2005)	Blink interval	F(2,18) =7.64, p<.01	0.6775
Savage et al., (2013)	Fixation durations	No other significant oculomotor differences between conditions.	0.0000
	Saccade Amplitude	No other significant oculomotor differences between conditions.	0.0000
	Saccade peak velocity	t(16) = 2.29, p=0.036	0.4970
	Horizontal spread of fixation positions	t(16) =3.06, p=0.008	0.6080
	Vertical spread of fixation positions	No other significant oculomotor differences between conditions.	0.0000
	Blink frequency	t(16) = 3.01, p=0.008	0.6010
	Blink duration	No other significant oculomotor differences between conditions.	0.0000
Steinhauer et al., (2000)	Pupil Diameter (extent of constriction)	F(1,32) = 58.2, p<0.0001	0.8033
Steinhauer et al., (2004)	Pupil Diameter	F(1,21) = 4.6, p=0.043, η^2 = 0.181	0.4240
Tokuda et al., (2011)	Saccadic intrusions	F(2,39) = 41.8, p<.05	0.8258
	Pupil dilation	F(2,39) = 23.07, p<.05	0.7362

Reference	Eye Measure(s) Studied	Statistical Results	Effect Size (r)
Van Orden et al., (2001)	Blink frequency	F(8,360) = 13.00, p<0.001	0.4733
	Blink duration	F(8,360) = 7.2, p<0.001	0.3715
	Fixation frequency	F(8,360) = 6.37, p<0.001	0.3521
	Fixation duration	F(8,360) = 0.15, p<0.5	0.0548
	Saccadic extent	F(8,360) = 3.15, p<0.005	0.2550
	Mean pupil size	F(8,360) = 2.14, p<0.05	0.2121
Veltman & Gaillard, (1996)	Blink duration	F(5,65) = 61.86, p<0.01	0.9090
	Number of Blinks	F(5,65) = 13.75, p<0.01	0.7170
Wang et al., (2014)	Fixation Count	F(1,41) = 115.051, p<0.0001	0.8586
	Fixation Duration	F(1,41) = 134.046, p<0.0001	0.8751
Wilson et al., (1994)	Blink Rate	F(4,24) = 14.09, p<0.0001	0.8375
	Blink duration	F(4,24) = 3.26, p=0.029	0.5933
Yang et al., (2012)	Fixation on touch screen	F(3,437) = 31.29	0.4207
	Blink Rate	F(3,437) = 2.066	0.1183
	Saccades	F(3,437) = 12.01	0.2757

Before continuing with the analysis, the effect sizes from these multiple studies need to be combined. In order to combine effect sizes in *r* from multiple studies, we first need to normalize our individual effect sizes. This step is essential since *r*-indexes can exhibit non-normal sampling distributions when estimating population values or the distribution of the *r*-indexes sampled will become more and more skewed (Cooper, 2010; Rosenthal, 1994). This is completed by converting the effect size *r* scores to z-scores using Fisher's *r*-to-*z* transformation (Rosenthal & DiMatteo, 2008; Rosenthal, 1994). From Rosenthal (1994), we can perform this transformation through the relationship between *r* and Z_r of:

$$Z_r = \frac{1}{2} log_e \left[\frac{1+r}{1-r}\right] \quad . \qquad (12)$$

According to Cooper (2010), the *z*-scores have no limiting value and are normally distributed. According to Rosenthal (1994), practically all meta-analytic procedures interested in *r* require most of the computations to be carried out on the transformation,

Z_r, and not actually on *r*.

Once normalized by the transformation, the means, both weighted and unweighted, of these transformed values must be calculated. According to Cooper (2010), a meta-analysis typically presents both weighted and unweighted average effect sizes. Although both will be calculated, this meta-analysis will utilize each of the average effect sizes for different computations and analyses. The weighted and unweighted average effect sizes for this meta-analysis are shown in Table 5.

Table 5. *Sample Sizes and Weighted and Unweighted Average Effect Sizes for Measurements and Moderator Variables.*

			Number of Studies	Total Sample Size	Weighted Average Effect Size	Unweighted Average Effect Size
Overall			60	1158	0.8407	0.8066
Measures	**Blinks**	**Overall**	18	315	0.5020	0.6434
		Blink Duration	6	72	0.6495	0.7189
		Blink Rate	8	191	0.3977	0.5417
		Blink Interval	1	10	0.8245	0.8245
		Blink Frequency	3	42	0.7199	0.7035
	Saccades	**Overall**	13	179	0.7196	0.8147
		Saccade Extent	6	89	0.6050	0.8323
		Saccade Rate	3	36	1.1535	0.9795
		Saccade Peak Velocity	1	17	0.5453	0.5453
		Saccade Amplitude	3	37	0.6792	0.7043
	Pupils	**Overall**	10	219	1.0455	1.0372
		Pupil Size	6	137	0.9694	0.9907
		Pupil Dilation	4	82	1.1748	1.2232
	Mean Dwell Time		2	31	0.8844	1.0842
	Fixations	**Overall**	16	401	1.0276	0.8055
		Fixation Duration	7	188	1.3242	0.8642
		Fixation Frequency	3	63	1.1531	0.9849
		Horizontal Fixation	3	75	0.5107	0.5935
		Vertical Fixation	3	75	0.6914	0.7014
	Area of Visual Field		1	13	0.7964	0.7964
Task Type	**Simplistic**		23	470	0.8374	0.7554
	Application		37	688	0.8429	0.8385
Eye-Tracking Method	**Head-Mounted**		19	399	0.9872	0.8358
	Display-Mounted		41	759	0.7619	0.7931

All final results will be reported in *r*, meaning that after the calculations are performed, the transformed values need to be converted back to the *r* correlation units. As a result, it is the unweighted mean of these transformed values that is then converted back to *r*, representing the unweighted mean *r* (Rosenthal & DiMatteo, 2001; Cooper, 2010). The equation to convert each of these values back to the *r* correlation units based on Borenstein et al. (2009) and Rosenthal (1994) is:

$$r = \frac{e^{2z} - 1}{e^{2z} + 1} \quad . \tag{13}$$

A 95% confidence interval can be estimated to determine whether the combined effect sizes differ significantly from zero using the following equation from Rosenthal and DiMatteo (2001):

$$\overline{Z_{\mathrm{r}}} \pm \frac{t(_{.05})S}{\sqrt{k}} \quad , \tag{14}$$

where $\overline{Z_{\mathrm{r}}}$ is the unweighted mean of the transformed *r* values, $t_{(.05)}$ is the appropriate t value at the 0.05 probability level, *S* is the standard deviation of the transformed r values, and k is the number of studies. Rosenthal and DiMatteo (2001) point out, that with the unweighted mean *r*, a random effects confidence interval is usually preferred, even when it yields wider confidence intervals, to allow generalization for studies other than those included in the analysis. Once computed, these lower and upper values of the 95% confidence intervals are transformed back to *r* values, using equation 13, defining the confidence intervals around the effect (Rosnow & Rosenthal, 1996; Cooper, 2010). An example using each of the computations and procedures described can be found in the Appendix.

It is then important to perform a homogeneity analysis on the *r* values that have been transformed to the appropriate *z*-scores. The test for homogeneity against z-transformed *r* values is provided by Cooper (2010) and involves the following formula:

$$Q_{\text{total}} = \sum_{i=1}^{k} (n_{\text{i}} - 3) z_{\text{i}}^2 - \frac{\left[\sum_{i=1}^{k} (n_{\text{i}} - 3) z_{\text{i}}\right]^2}{\sum_{i=1}^{k} (n_{\text{i}} - 3)} \quad , \tag{15}$$

where n_i is the total sample size for the *i*th comparison, k is the number of studies and z_i is the transformed *r* values.

IV. RESULTS AND ANALYSIS

The effect sizes and confidence intervals were computed for each measure considered in the overall meta-analysis as well as for the three moderator variables. The results from the meta-analysis are shown in Table 6 with bold results indicating non-significant findings or those results where the confidence interval includes zero indicating there could be no relationship between cognitive workload and eye-related measurements. The combined effect size for the meta-analysis resulted in an effect size of 0.668. By examining all of the studies included in the meta-analysis collectively, without factoring in any of the moderator variables, there is a large effect size, or relationship, between cognitive workload and eye-related measurements. The 95% confidence interval was estimated as [0.569, 0.748].

The homogeneity analysis performed resulted in a value of 489.871, which is a highly significant result based on a chi-square test with 59 degrees of freedom from a critical value of chi-square at $p < 0.05$. The interpretation of this significant result implies that given the sizes of the samples on which these variance estimates are based, the variation calculated in effect sizes is too great to be explained by only sampling error. An analysis on moderator variables can identify other possible distinctions between the studies that are contributing to the difference in variances. Furthermore, with the inclusion of the moderator variables, we can identify possible variables that could be important to consider for future modeling and design.

Table 6. *Summary Table for the Meta-Analysis, Including the Examination of Moderator Variables.*

			Number of Studies	**Combined Effect Size**	**95% Confidence Interval**
Overall			60	0.668	[0.569, 0.748]
Measures	**Blinks**	**Overall**	18	0.567	[0.402, 0.697]
		Blink Duration	6	0.616	[0.144, 0.860]
		Blink Rate	8	0.494	[0.151, 0.731]
		Blink Interval	1	0.678	[ ]
		Blink Frequency	3	0.607	[0.219, 0.829]
	Saccades	**Overall**	13	0.672	[0.305, 0.865]
		Saccade Extent	6	0.682	[0.247, 0.888]
		Saccade Rate	3	**0.753**	**[-0.849, 0.997]**
		Saccade Peak Velocity	1	0.497	[ ]
		Saccade Amplitude	3	**0.607**	**[-0.981, 0.999]**
	Pupils	**Overall**	10	0.777	[0.632, 0.869]
		Pupil Size	6	0.726	[0.398, 0.890]
		Pupil Dilation	4	0.837	[0.680, 0.921]
	Mean Dwell Time		2	**0.795**	**[-1.000, 1.000]**
	Fixations	**Overall**	16	0.667	[0.283, 0.867]
		Fixation Duration	7	**0.698**	**[-0.191, 0.958]**
		Fixation Frequency	3	**0.755**	**[-0.330, 0.981]**
		Horizontal Fixation	3	0.532	[0.205, 0.753]
		Vertical Fixation	3	**0.605**	**[-0.738, 0.982]**
	Area of Visual Field		1	0.662	[ ]
Task Type	**Simplistic**		23	0.638	[0.503, 0.743]
	Application		37	0.685	[0.539, 0.791]
Eye-Tracking System	**Head-Mounted**		19	0.684	[0.441, 0.833]
	Display-Mounted		41	0.660	[0.548, 0.749]

(Bold results indicate non-significant findings, i.e., confidence interval includes zero).

The first moderator variable examined was of the individual eye-related measurements observed in the studies included in the meta-analysis. By performing this analysis, we can identify which of the specific eye-related measurements are most and least affected by cognitive workload. Those eye-related dependent variables analyzed from each study were grouped into the categories of blinks, saccades, pupils, mean dwell time, fixations and the area of visual field. These categories were further broken down to analyze each type of measurement observed under each category.

From this analysis, large effect sizes were calculated for each of the categories, with the highest significant effect size resulting from measurements of the pupil with an effect size of 0.777 and an estimated 95% confidence interval of [0.632, 0.869]. Investigating the individual measurement types for the pupil, measuring pupil dilation appears to have the largest relationship with cognitive workload based on the effect size of 0.837 and the tight 95% confidence interval [0.680, 0.921]. Although a large overall effect size of 0.795 was calculated for mean dwell time, the effect size was non-significantly different from zero; that is, the estimated 95% confidence interval included zero, [-1.000, 1.000].

Even with most measurements resulting in large effect sizes, there were still a few variables that appear to be less affected by cognitive workload than others based on the inclusion of zero within the estimated 95% confidence intervals. These measurements included saccade rate, saccade amplitude, fixation duration, fixation frequency, and vertical fixation. There are also three variables for which a 95% confidence interval could not be calculated. These variables are blink interval, saccade peak velocity and area of visual field since there was only one study observing these variables included in the meta-analysis.

The results for the different eye-related measurements are also presented as a forest plot, where all computed effect sizes are marked with a symbol, either a square or a diamond, and are shown with the estimated 95% confidence interval. The forest plot is shown in Figure 2.

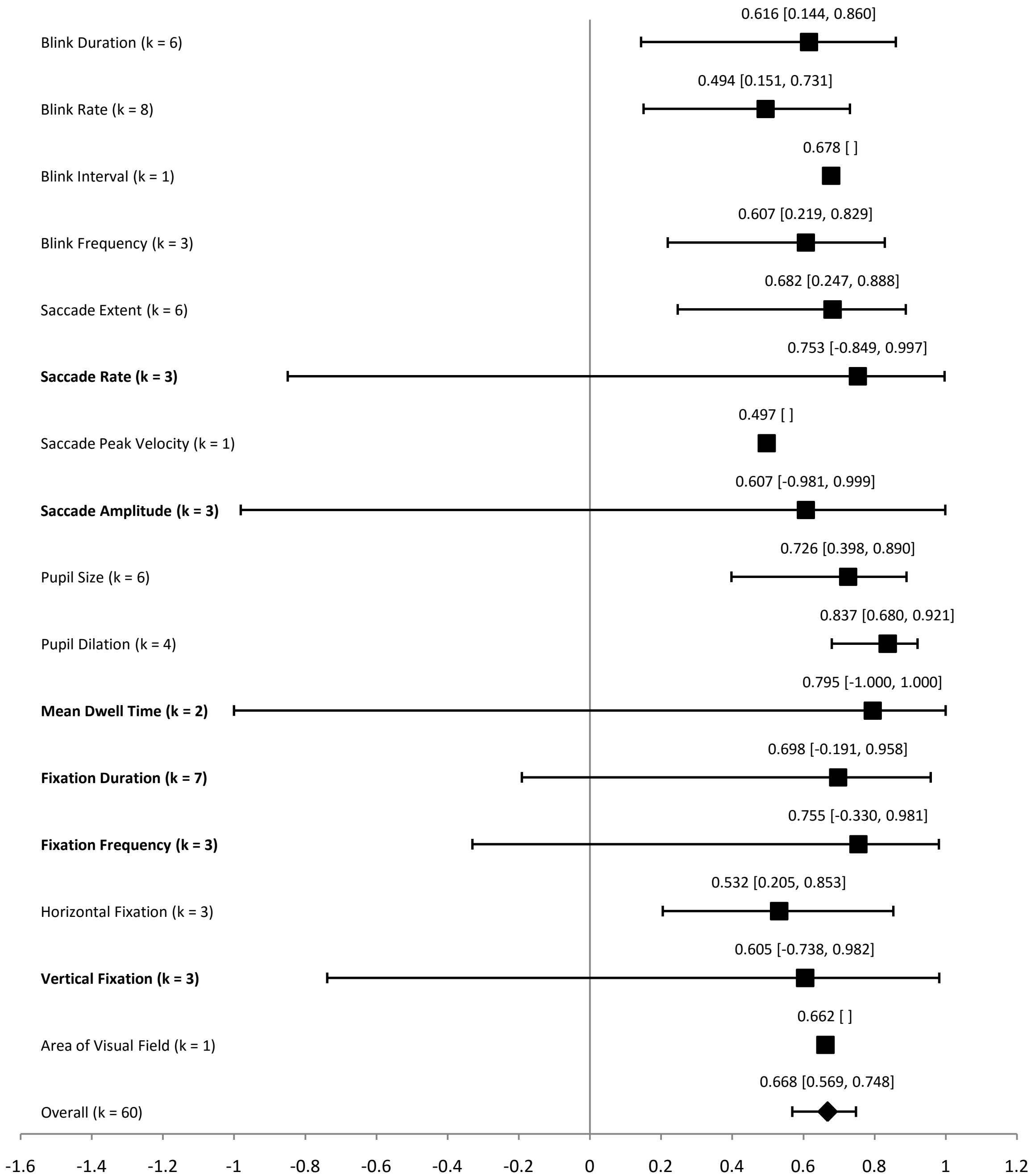

Figure 2. Forest Plot Representing the Effects of Cognitive Workload on Eye-related Measurements.

The forest plot provides a visual depiction of the results, not only providing the capability to easily and quickly identify those measurements of cognitive workload, but to also identify those areas where continued research is needed. Those areas would be indicated by large confidence intervals, those non-significant confidence intervals including zero, or the use of a small sample size (Schaefer et al., 2016).

The second moderator variable of task type was also analyzed. With both simplistic and application task types resulting in large effect sizes, it can be determined that there is no difference between the type of task being performed and using an eye-related measurement to measure the effect of cognitive workload. In other words, both types of tasks result in similar effects with cognitive workload on eye-related measurements. In fact, there is relatively no statistical difference between the two types of tasks with estimated 95% confidence intervals being [0.503, 0.743] and [0.539, 0.791] respectively.

A similar conclusion can be drawn from the third moderator variable of eye-tracking system utilized. Not only did both systems result in large effect sizes, but the effect sizes were roughly equivalent at 0.684 for head-mounted systems and 0.660 for display-mounted systems with the 95% confidence intervals being [0.441, 0.833] and [0.548, 0.749] respectively.

V. CONCLUSIONS AND APPLICATIONS

There are some important findings to be observed from the current meta-analysis. First, the large, significant effect size achieved for the examination of the studies collectively indicates that the use of physiological, or more specifically measurements from the eye, provide a reliable measurement of cognitive workload. From this conclusion, the monitoring and evaluating of eye-related measurements in systems would allow for identifying and handling the levels of cognitive workload imposed on the operator.

It can also be important to further discuss the previously mentioned notion of the number of unique contributing authors and selections of unique journals. From the studies included in the meta-analysis, two authors appeared as contributing authors on multiple studies, contributing to the research performed on two unique studies each, totaling four studies with similar authors. This practice is not uncommon in the research community, with similar contributing authors appearing on multiple studies examining similar topics, either based on their expertise or their designated research domain. It would be important to examine if similar measurements were examined by those overlapping authors, which could provide further support for the results obtained against those specific eye-related measurements. On the other hand, the inclusion of multiple studies from similar authors could hinder the results obtained if inconsistent or incorrect methods were proven to be performed by the author. Similar biases could be observed if the studies included in the meta-analysis were published from a limited diversity of journals. For instance, the aims and scope of each journal are specific to a range of topics and could limit the discovery and inclusion of other important studies found within a different journal. The studies included in this meta-analysis range from journals in the domains of ergonomics, psychology, and engineering; thus emphasizing the vast importance and attempts to understand cognitive workload in a variety of different domains. With a sufficient uniqueness obtained from the journals and authors collected for this meta-analysis, any inconsistencies based on those similarities should be excluded from the results.

With the inclusion of multiple eye-related measurements, the additional intent of this meta-analysis was to attempt to further differentiate or identify those specific measurements that have a significant link to measuring cognitive workload. Through the examination of the dependent variables as moderators, specific measurements with a relationship to measuring cognitive workload were identified or further confirmed. With this knowledge, the selection of measurements to aid in future research can be simplified. This allows the researchers to select from known measurements of workload where one measure may be more obtainable or measureable than another within a particular system.

A discussion in regards to the classification of those dependent eye-related measurements identified from each study included in the meta-analysis and the reasoning for such classifications is also important. Without having direct knowledge of the individual research experiments conducted, any specific assumptions as to the intention of how the authors wished to utilize the specific measurements selected for observation could not be made. To prevent any misinterpretation for its intention, those eye-related measurements were categorized based on the wording selection of the authors.

Although individual studies have shown different outcomes in terms of the relationship of a specific measurement, the purpose of performing a meta-analysis is to provide a single estimate of the reliability and magnitude for a measurement. Those significant and reliable measurements identified through this meta-analysis are blink duration, blink rate, blink interval, blink frequency, saccade extent, saccade peak velocity, pupil size, pupil dilation and horizontal fixation. It is important to note the inclusion of one study that did not identify the data for non-significant results. For this meta-analysis, these non-significant results were estimated as zero, which could lead to an underestimate of the average effect size (Pigott, 1994).

In addition, the number of included studies in this meta-analysis may not provide the most stable results. A previous meta-analysis only analyzed effect sizes where at least three correlation coefficients were available, based on the determination of the typical minimum standard indentified in past meta-analysis methodologies. With fewer studies, the values obtained based on the combined effect size can be unstable (Caird, Willness, Steel, & Scialfa, 2008). Therefore, estimates based on limited information should be interpreted with the appropriate caution. Nevertheless, it can be inferred by the results of this meta-analysis, that simply assuming that the small number of studies observing one measurement compared to another would not prevent the outcome of observing a large, significant effect size. This is shown since some measurements with an equal number of studies conclude both significant and non-significant results. However, these analyses based on limited information can serve to reveal the scarcity of studies that examine cognitive workload and eye-related measurements, pointing toward the need for further experimentation to reliably identify these effects. In particular, from the number of studies observing each eye-related measurement, we can identify which measurements are being observed most often in experiments or studies compared to those measurements observed the least.

For future analyses, the statistical results and analyses missing from some studies should be requested from the study authors to potentially include the study in the meta-analysis. This would prevent the exclusion of studies based on the criteria of unreported effect sizes. This would limit the need to justify in the inclusion criteria that sufficient information be included to determine effect size estimates.

It is also important to recognize those estimated 95% confidence intervals for variables that included zero. This result could represent the potential for a null effect for that measure or it could be explained through other limitations such as the number of studies included in the analysis or the existence of a large diversity in the characteristics of the findings within those included studies (Schaefer et al., 2016). These non-significant confidence intervals should not condemn the use of these measurements, but instead should encourage additional research observing those particular measurements.

A second finding from this meta-analysis refers to the similarities observed between both simplistic and application task types or between those memory or arithmetic type tasks and those types of task performed in real or simulated expert-driven tasks. This is an important finding that allows for the designing or redesigning of systems to be non-restrictive to the types of tasks being performed and subsequently analyzed in terms of cognitive workload.

Findings based on the moderator variables of eye-tracking systems also resulted in a similarity between the two systems. Although this meta-analysis did not indicate a difference between the type of eye-tracking system used, this does not imply that both systems are comparable. Since equipment could result in a lower reliance on actual measurements, this could have reduced the overall effect size for either head-mounted or display-mounted systems. Until equipment can be standardized and show consistent results for similar measurements, it can be assumed that there could be implications from using the current existing systems. For instance, Jacob and Karn (2003) note that more work is needed to resolve technical issues with the current eye tracking systems and in terms of the analysis of the produced data. These issues include "constraints on participant movement; tracker accuracy, precision, ease of setup; dealing with dynamic stimuli; and labor-intensive data extraction" (Jacob & Karn, 2003).

According to Cegarra and Chevalier (2008), there are no methods that can evaluate and measure cognitive workload alone, which is why it is not atypical for studies investigating techniques for evaluating and measuring cognitive workload to incorporate the use of more than one measurement from performance, subjective, and physiological techniques. However, with no method perfectly measuring cognitive workload by itself, the addition of eye-related measurements with other proven measurements can strengthen the design and implementation of a system for measuring and identifying cognitive workload.

Appendix

Example of Conversions and Procedures Utilized in the Meta-Analysis

An example is discussed utilizing the conversions and procedures described in this meta-analysis from Rosenthal and DiMatteo (2001), Cohen (1988), Rosenthal (1984) and Rosnow and Rosenthal (1996) to compute the effect size r from Brookings et al. (1996), one of the studies included in this meta-analysis.

The statistical results from Brookings et al. (1996) are F(2,14) = 9.37. Using the equations we can calculate the effect size r to be 0.7566.

$$F \frac{df_{\text{treatment}}}{df_{\text{error}}} = \frac{\eta^2}{1-\eta^2} \quad \rightarrow \quad 9.37\,\frac{2}{14} = \frac{\eta^2}{1-\eta^2}$$

$$f^2 = \frac{\eta^2}{1-\eta^2} \quad \rightarrow \quad f^2 = 9.37\,\frac{2}{14} \quad \rightarrow \quad f^2 = 1.3386$$

$$d = 2f \quad \rightarrow \quad d = 2\left(\sqrt{1.3386}\right) \quad \rightarrow \quad d = 2.3139$$

$$r = \sqrt{\frac{d^2}{d^2+4}} \quad \rightarrow \quad r = \sqrt{\frac{(2.3139)^2}{(2.3139)^2+4}} \quad \rightarrow \quad r = 0.7566$$

Then, using Fisher's *r*-to-*z* transformation, we can compute Z_r to be 0.9882.

$$Z_r = \frac{1}{2} log_e \left[\frac{1+r}{1-r}\right] \quad \rightarrow Z_r = \frac{1}{2} log_e \left[\frac{1+0.7566}{1-0.7566}\right] \quad \rightarrow \quad Z_r = 0.9882$$

To continue with the analysis, the effect sizes from the multiple studies needed to be combined, or averaged, based on the separate eye-related measurements. After computing the average, $\overline{Z_r}$, for those studies observing blink rate as 0.5417 with S = 0.4655, we can calculate the estimated 95% confidence interval for blink rate.

$$\overline{Z_{\mathrm{r}}} \pm \frac{t(_{.05})S}{\sqrt{k}} \quad \rightarrow \quad 0.5417 \pm \frac{(2.365)0.4655}{\sqrt{8}} \quad \rightarrow \quad 0.5417 \pm 0.3892$$

After computing $\overline{Z_{\mathrm{r}}}$ for blink rate of 0.5417, we can convert this value back to *r* units to express the combined effect size for blink rate in terms of the effect. This computation results in an effect size r of 0.494.

$$r = \frac{e^{2z} - 1}{e^{2z} + 1} \quad \rightarrow \quad r = \frac{e^{2(0.5417)} - 1}{e^{2(0.5417)} + 1} \quad \rightarrow \quad r = 0.494$$

After computing the lower and upper values for the 95% confidence interval for blink rate as [0.1524, 0.9309], we can convert these values back to *r* units to define the 95% confidence interval around the effect for blink rate. The estimated 95% confidence interval computed for blink rate is [0.151, 0.731].

$$r_{\mathrm{lower}} = \frac{e^{2z} - 1}{e^{2z} + 1} \quad \rightarrow \quad r_{\mathrm{lower}} = \frac{e^{2(0.1524)} - 1}{e^{2(0.1524)} + 1} \quad \rightarrow \quad r_{\mathrm{lower}} = 0.151$$

$$r_{\mathrm{upper}} = \frac{e^{2z} - 1}{e^{2z} + 1} \quad \rightarrow \quad r_{\mathrm{upper}} = \frac{e^{2(0.9309)} - 1}{e^{2(0.9309)} + 1} \quad \rightarrow \quad r_{\mathrm{upper}} = 0.731$$

References

*Backs, R. W., & Walrath, L. C. (1992). Eye movement and pupillary response indices of mental workload during visual search of symbolic displays. *Applied Ergonomics*, *23*(4), 243-254. doi:10.1016/0003-6870(92)90152-L

Beatty, J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological Bulletin*, *91*(2), 276-292. doi:10.1037/0033-2909.91.2.276

Beatty, J., & Kahneman, D. (1966). Pupillary changes in two memory tasks. *Psychonomic Science*, *5*(10), 371-372.

*Benedetto, S., Pedrotti, M., Minin, L., Baccino, T., Re, A., & Montanari, R. (2011). Driver workload and eye blink duration. *Transportation Research Part F: Psychology And Behaviour*, *14,* 199-208. doi:10.1016/j.trf.2010.12.001

Bihl, T. J. (2015). Feature selection and classifier development for radio frequency device identification (Doctoral dissertation, Air Force Institute of Technology).

Bihl, T. J., Bauer, K. W., Temple, M. A., & Ramsey, B. (2015). Dimensional reduction analysis for Physical Layer device fingerprints with application to ZigBee and Z-Wave devices. *MILCOM 2015 - 2015 IEEE Military Communications Conference*, 360-365. doi:10.1109/MILCOM.2015.7357469

Bommer, S. C., & Fendley, M. (2015). Assessing the effects of multimodal communications on mental workload during the supervision of multiple unmanned aerial vehicles. *International Journal of Unmanned Systems Engineering, 3*(1), 38-50. http://dx.doi.org/10.14323/ijuseng.2015.4

Borenstein, M., Hedges, L. V., Higgins, J. P. T. & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Chichester, UK: John Wiley & Sons, Ltd.

*Brookings, J. B., Wilson, G. F., & Swain, C. R. (1996). Psychophysiological responses to changes in workload during simulated air traffic control. *Biological Psychology*, *42*(3), 363-377. doi:10.1016/0301-0511(95)05167-8

Buettner, R. (2013). Cognitive Workload of Humans Using Artificial Intelligence Systems: Towards Objective Measurement Applying Eye-Tracking Technology. *KI 2013: Advances In Artificial Intelligence*, 37-48. doi:10.1007/978-3-642-40942-4_4

Caird, J. K., Willness, C. R., Steel, P., & Scialfa, C. (2008). A meta-analysis of the effects of cell phones on driver performance. *Accident Analysis And Prevention*, *40,* 1282-1293. doi:10.1016/j.aap.2008.01.009

Cegarra, J., & Chevalier, A. (2008). The use of Tholos software for combing measures of mental workload: toward theoretical and methodological improvements. *Behavior Research Methods*, *40*(4), 988-1000. doi: 10.3758/BRM.40.4.988

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: L. Erlbaum Associates.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155-159. doi:10.1037/0033-2909.112.1.155

Cooper, H. (2010). *Research synthesis and meta-analysis: A step-by-step approach (4th ed.)*. Thousand Oaks, CA: Sage Publications, Inc.

*De Rivecourt, M., Kuperus, M., Post, W., & Mulder, L. (2008). Cardiovascular and eye activity measures as indices for momentary changes in mental effort during simulated flight. *Ergonomics*, *51*(9), 1295-1319.

*Di Nocera, F., Camilli, M., & Terenzi, M. (2007). A random glance at the flight deck: Pilots' scanning strategies and the real-time assessment of mental workload. *Journal Of Cognitive Engineering And Decision Making*, *1*(3), 271-285. doi:10.1518/155534307X255627

Di Stasi, L. L., Antolí, A., Gea, M., & Cañas, J. J. (2011). A neuroergonomic approach to evaluating mental workload in hypermedia interactions. *International Journal Of Industrial Ergonomics*, *41,* 298-304. doi:10.1016/j.ergon.2011.02.008

East, J. A., Bauer Jr., K. W., & Lanning, J. W. (2002). Feature selection for predicting pilot mental workload: A feasibility study. *International Journal Of Smart Engineering System Design*, *4*(3), 183-193. doi:10.1080/10255810290008081

Endsley, M. R. (1995). Measurement of situation awareness in dynamic systems. *Human Factors*, *37*(1), 65-84. doi:10.1518/001872095779049499

Endsley, M. R. (2012) Situation Awareness. In G. Salvendy (Eds.), *Handbook of Human Factors and Ergonomics*, (pp. 553-568). Hoboken, NJ: John Wiley & Sons, Inc.

Endsley, M. R., & Garland, D. J. (2000). *Situation awareness: Analysis and measurement*. Mahwah, NJ: CRC Press.

Engström, J., Johansson, E., & Östlund, J. (2005). Effects of visual and cognitive load in real and simulated motorway driving. *Transportation Research Part F: Psychology And Behaviour*, *8*, 97-120. doi:10.1016/j.trf.2005.04.012

Fogarty, C., & Stern, J. A. (1989). Eye movements and blinks: Their relationship to higher cognitive processes. *International Journal Of Psychophysiology*, *8*(1), 35-42. doi:10.1016/0167-8760(89)90017-2

Glioma Meta-analysis Trialists (GMT) Group. (2002). Chemotherapy in adult high-grade glioma: A systematic review and meta-analysis of individual patient data from 12 randomized trials. *The Lancet, 359*(9311), 1011-1018.

Haapalainen, E., Kim, S., Forlizzi, J. F., & Dey, A. K. (2010). Psycho-physiological measures for assessing cognitive load. *Proceedings Of The 12th ACM International Conference: Ubiquitous Computing*, 301-310. doi:10.1145/1864349.1864395

Hall, J. A., & Rosenthal, R. (1995). Interpreting and evaluating meta-analysis. *Evaluation & The Health Professions*, *18*(4), 393-407.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature Methods*, *12*(3), 179-185. doi:10.1038/nmeth.3288

Hess, E. H., & Polt, J. M. (1964). Pupil size in relation to mental activity during simple problem-solving. *Science*, *143*(3611). 1190-1192.

Higgins, J. P., & Green, S. (2008). Analysing data and undertaking meta-analyses. *Cochrane Handbook For Systematic Reviews Of Interventions*, 243-296. doi:10.1002/9780470712184.ch9

Hogervorst, M., Brouwer, A., & van Erp, J. (2014). Combining and comparing EEG, peripheral physiology and eye-related measures for the assessment of mental workload. *Frontiers In Neuroscience*, *8.*

Holland, M. K., & Tarlow, G. (1975). Blinking and thinking. *Perceptual And Motor Skills*, *41*(2), 403-406. doi:10.2466/pms.1975.41.2.403

Holmquist, K., Nystrom, N., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). *Eye tracking: A comprehensive guide to methods and measures*. Oxford, UK: Oxford University Press.

Horrey, W., & Wickens, C. (2004). Cell phones and driving performance: A meta-analysis. In *Proceedings Of The Human Factors And Ergonomics Society Annual Meeting*, *48*(19), 2304-2308.

Iqbal, S. T., Zheng, X. S., &Bailey, B. P. (2004). Task-evoked pupillary response to mental workload in human-computer interaction. *Conference on Human Factors in Computing Systems Proceedings*, 1477-1480.

Jacob, R. J., & Karn, K. S. (2003). Commentary on section 4: Eye tracking in human-computer interaction and usability research. Ready to deliver the promises. *The Mind's Eye*, 573-605. doi:10.1016/B978-044451020-4/50031-1

Kahneman, D. (1973). *Attention and effort*. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Kahneman, D., Beatty, J., & Pollack, I. (1967). Perceptual Deficit during a Mental Task. *Science*, (3785). 218-219.

Klingner, J., Kumar, R., & Hanrahan, P. (2008). Measuring the task-evoked pupillary response with a remote eye tracker. *Eye Tracking Research & Application*, 69. doi:10.1145/1344471.1344489

Kramer, A.F. (1990). Physiological Metrics of Mental Workload: A Review of Recent Progress. Navy Personnel Research and Development Center.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2005). *Applied linear statistical models*. New York, NY: McGraw Hill/Irwin.

Laine, T. I., Bauer Jr., K. W., Lanning, J. W., Russell, C. A., & Wilson, G. F. (2002). Selection of input features across subjects for classifying crewmember workload using artificial neural networks. *IEEE Transactions On Systems, Man & Cybernetics: Part A*, *32*(6), 691-704.

Maier, A., Baltsen, N., Christoffersen, H., & Storrle, H. (2014). Towards diagram understanding: A pilot study measuring cognitive workload through eye-tracking. *Proceedings of International Conference on Human Behaviour in Design 2014.*

Marquart, G., Cabrall, C., & de Winter, J. (2015). Review of eye-related measures of drivers' mental workload. *Procedia Manufacturing*, *3*(6th International Conference on Applied Human Factors and Ergonomics (AHFE 2015) and the Affiliated Conferences, AHFE 2015), 2854-2861. doi:10.1016/j.promfg.2015.07.783

Marshall, S. P. (2007). Identifying cognitive state from eye metrics. *Aviation, Space, And Environmental Medicine*, *78*(5), B165-B175.

*May, J. G., Kennedy, R. S., Williams, M. C., Dunlap, W. P., & Brannan, J. R. (1990). Eye movement indices of mental workload. *Acta Psychologica*, *75*(1), 75-89. doi:10.1016/0001-6918(90)90067-P

Montgomery, D. C. (2013). *Design and analysis of experiments*. Hoboken, NJ: John Wiley & Sons, Inc.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews Of The Cambridge Philosophical Society*, *82*(4), 591-605.

*Niezgoda, M., Tarnowski, A., Kruszewski, M., & Kamiński, T. (2015). Towards testing auditory–vocal interfaces and detecting distraction while driving: A comparison of eye- movement measures in the assessment of cognitive workload. *Transportation Research Part F: Psychology And Behaviour*, *32,* 23-34. doi:10.1016/j.trf.2015.04.012

Noel, J. B., Bauer, J. W., & Lanning, J. W. (2005). Improving pilot mental workload classification through feature exploitation and combination: A feasibility study. *Computers And Operations Research*, *32*(10), 2713-2730. doi:10.1016/ j.cor.2004.03.022

Onnasch, L., Wickens, C. D., Li, H., & Manzey, D. (2014). Human performance consequences of stages and levels of automation: an integrated meta-analysis. *Human Factors*, *56*(3), 476-488.

Pigott, T. D. (1994). Methods for handling missing data in research synthesis. In Cooper, H., Hedges, L. V. (Eds.), *The handbook of research synthesis* (pp. 163-175). New York, NY: Russell Sage Foundation.

*Pomplun, M., & Sunkara, S. (2003). Pupil dilation as an indicator of cognitive workload in human-computer interaction. In *Proceedings of the International Conference on HCI*, 542-546.

Poole, A., & Ball L. J. (2005). Eye tracking in human-computer interaction and usability research: Current status and future prospects. In Ghaoui, C. (Eds.), *Encyclopedia of Human-Computer Interaction.* Hershey, PA: Idea Group, Inc.

*Rantanen, E., & Goldberg, J. (1999). The effect of mental workload on the visual field size and shape. *Ergonomics*, *42*(6), 816-834.

*Recarte, M. Á., Pérez, E., Conchillo, Á., & Nunes, L. M. (2008). Mental workload and visual impairment: Differences between pupil, blink, and subjective rating. *The Spanish Journal Of Psychology*, *11*(2), 374-385.

*Reyes, M. L., & Lee, J. D. (2008). Effects of cognitive load presence and duration on driver eye movements and event detection performance. *Transportation Research Part F: Psychology And Behaviour*, *11,* 391-402. doi:10.1016/ j.trf.2008.03.004

*Rosenfield, M., Jahan, S., Nunez, K., & Chan, K. (2015). Cognitive demand, digital screens and blink rate. *Computers In Human Behavior*, *51*, 403-406. doi:10.1016/j.chb.2015.04.073

Rosnow, R. L., & Rosenthal, R. (1996). Computing contrasts, effect sizes, and counternulls on other people's published data: General procedures for research consumers. *Psychological Methods*, *1*(4), 331-340. doi:10.1037/1082-989X.1.4.331

Rosenthal, R. (1984). *Meta-analytic procedures for social research*. Beverly Hills, CA: Sage Publications.

Rosenthal, R. (1994). Parametric measures of effect size. In Cooper, H., Hedges, L. V. (Eds.), *The handbook of research synthesis* (pp. 231-244). New York, NY: Russell Sage Foundation.

Rosenthal, R., & DiMatteo, M. R. (2001). Meta-analysis: Recent developments in quantitative methods for literature reviews. *Annual Review Of Psychology*, *52*(1), 59-82.

Rozado, D., & Dunser, A. (2015). Combining EEG with pupillometry to improve cognitive workload detection. *Computer, 48*(10), 18-25. doi:10.1109/MC.2015 .314

*Ryu, K., & Myung, R. (2005). Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *International Journal Of Industrial Ergonomics*, *35,* 991-1009. doi:10.1016/j.ergon.2005.04.005

*Savage, S. W., Potter, D. D., & Tatler, B. W. (2013). Does preoccupation impair hazard perception? A simultaneous EEG and eye tracking study. *Transportation Research Part F: Psychology And Behaviour*, *17,* 52-62. doi:10.1016/ j.trf.2012.10.002

Schaefer, K. E., Chen, J. C., Szalma, J. L., & Hancock, P. A. (2016). A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. *Human Factors*, *58*(3), 377-400. doi:10.1177/0018720816634228

Stanton, N. A., Salmon, P. M., Walker, G. H., Baber, C., & Jenkins, D. P. (2005). *Human factors methods: A practical guide for engineering and design*. Burlington, VT: Ashgate Publishing Company.

*Steinhauer, S. R., Condray, R., & Kasparek, A. (2000). Cognitive modulation of midbrain function: task-induced reduction of the pupillary light reflex. *International Journal Of Psychophysiology*, *39,* 21-30. doi:10.1016/S0167-8760(00)00119-7

*Steinhauer, S. R., Siegle, G. J., Condray, R., & Pless, M. (2004). Sympathetic and parasympathetic innervation of pupillary dilation during sustained processing. *International Journal Of Psychophysiology*, *52*, 77-86. doi:10.1016/j.ijpsycho.2003.12.005

Telford, C. W., & Thompson, N. (1933). Factors influencing eyelid responses. *Journal of Experimental Psychology, 16*, 524-539.

Thalheimer, W., & Cook, S. (2002). How to calculate effect sizes from published research: A simplified methodology. *Work-Learning Research*, 1-9.

*Tokuda, S., Obinata, G., Palmer, E., & Chaparro, A. (2011). Estimation of mental workload using saccadic eye movements in a free-viewing task. *Conference Proceedings: ... Annual International Conference Of The IEEE Engineering In Medicine And Biology Society. IEEE Engineering In Medicine And Biology Society. Annual Conference*, *2011,* 4523-4529. doi:10.1109/IEMBS. 2011.6091121

Uttal, D. H., Meadow, N. G., Tipton, E., Hand, L. L., Alden, A. R., Warren, C., & Newcombe, N. S. (2013). The malleability of spatial skills: A meta-analysis of training studies. *Psychological Bulletin*, *139*(2), 352-402.

*Van Orden, K. F., Limbert, W., Makeig, S., & Jung, T. (2001). Eye activity correlates of workload during a visuospatial memory task. Human Factors, 43(1), 111-121. doi:10.1518/001872001775992570

*Veltman, J. A., & Gaillard, A. K. (1996). Physiological indices of workload in a simulated flight task. *Biological Psychology*, *42*(3), 323-342. doi:10.1016/0301-0511(95)05165-1

Wang, L., He, X., & Chen, Y. (2014). Distinguishing analysis on workload peak and overload under time pressure with pupil diameter. *2014 IEEE International Inter-Disciplinary Conference On Cognitive Methods In Situation Awareness & Decision Support (Cogsima)*, 151. doi:10.1109/CogSIMA.2014.6816555

*Wang, Q., Yang, S., Liu, M., Cao, Z., & Ma, Q. (2014). An eye-tracking study of website complexity from cognitive load perspective. *Decision Support Systems*, *62,* 1-10.

Wasserstein, R. L., & Lazar, N. A. (2016). The ASA's statement on p-values: Context, process, and purpose. *The American Statistician*. doi:10.1080/00031305.2016.1154108

Wickens, C. D. (2008). Multiple Resources and Mental Workload. (cover story). *Human Factors*, *50*(3), 449-455. doi:10.1518/001872008X288394

Wickens, C. D., Hollands, J. G., Banbury, S., & Parasuraman, R. (2013). *Engineering psychology and human performance* (Fourth ed.). Boston, MA: Pearson Education, Inc.

Wierwille, W. W., & Eggemeier, F. T. (1993). Recommendations for mental workload measurement in a test and evaluation environment. *Human Factors, 35*(2), 263-281.

*Wilson, G. F., Fullenkamp, P., & Davis, I. (1994). Evoked potential, cardiac, blink, and respiration measures of pilot workload in air-to-ground missions. *Aviation, Space, And Environmental Medicine*, *65*(2), 100-105.

Xie, B., & Salvendy, G. (2000). Prediction of metal workload in single and multiple task environments. *International Journal Of Cognitive Ergonomics*, *4*(3), 213-242.

*Yang, Y., McDonald, M., & Zheng, P. (2012). Can drivers' eye movements be used to monitor their performance? A case study. *IET Intelligent Transport System*, *6*(4), 444-452. doi:10.1049/iet-its.2012.0008

Young, A. H., & Hulleman, J. (2013). Eye movements reveal how task difficulty moulds visual search. *Journal Of Experimental Psychology: Human Perception And Performance*, *39*(1), 168-190.