Toward Objective Evaluation of Working Memory in Visualizations: A Case Study Using Pupillometry and a Dual-Task Paradigm

Lace M.K. Padilla, Spencer C. Castro, P. Samuel Quinan, Ian T. Ruginski, and Sarah H. Creem-Regehr

Abstract— Cognitive science has established widely used and validated procedures for evaluating working memory in numerous applied domains, but surprisingly few studies have employed these methodologies to assess claims about the impacts of visualizations on working memory. The lack of information visualization research that uses validated procedures for measuring working memory may be due, in part, to the absence of cross-domain methodological guidance tailored explicitly to the unique needs of visualization research. This paper presents a set of clear, practical, and empirically validated methods for evaluating working memory during visualization tasks and provides readers with guidance in selecting an appropriate working memory evaluation paradigm. As a case study, we illustrate multiple methods for evaluating working memory in a visual-spatial aggregation task with geospatial data. The results show that the use of dual-task experimental designs (simultaneous performance of several tasks compared to single-task performance) and pupil dilation can reveal working memory demands associated with task difficulty and dual-tasking. In a dual-task experimental design, measures of task completion times and pupillometry revealed the working memory demands associated with both task difficulty and dual-tasking. Pupillometry demonstrated that participants' pupils were significantly larger when they were completing a more difficult task and when multitasking. We propose that researchers interested in the relative differences in working memory between visualizations should consider a converging methods approach, where physiological measures and behavioral measures of working memory are employed to generate a rich evaluation of visualization effort.

Index Terms—Working Memory, Cognitive Effort, Evaluation Methods, Pupillometry, Geographic/Geospatial Visualization, Quantitative Evaluation

1 Introduction

What makes one visualization *better* than another? Although seemingly simple, the question of how to objectively evaluate visualization quality is far from answered. In addition to important *user experience goals*, such as memorability [34], engagement [5], and enjoyment [76], one way to evaluate visualizations is to consider task-specific *usability criteria* [75], such as how easy or difficult it is for a user to complete the given task. For example, accuracy is a commonly used usability criterion for visualization quality, because a highly memorable or enjoyable visualization that misleads viewers is of poor quality [86]. Visualizations that elicit a prompt and accurate understanding of the data have clear user experience benefits, but what can these measures actually tell us about how hard or easy it is for users to complete their goals with the visualization? In this paper, we take a critical look at the conclusions one can and cannot draw from measures of speed and accuracy, based on cognitive processing mechanisms. We also detail less frequently used measures that focus on examining the mental effort associated with completing a visualization task. Further, we advocate for a *converging methods* approach (i.e., using multiple measures to examine phenomena that cannot be measured directly, such as mental effort), to create a clearer and more objective picture of visualization quality.

In cognitive and visualization science, there is no consensus on how effectively tests of accuracy and speed measure mental effort. Some researchers propose that accuracy and speed reflect the cognitive effort required to complete a task [39, 49, 80]. Borrowing the physical analogy used by Shenhav et al. [80], if one's task is to lift an object to a given height, the task demands would include the weight of that object, the height, and the *affordance* of the object to the task of being lifted. It is easier, for example, to lift a 20kg barbell over your head than a 20kg fish tank because the barbell was designed for and better affords lifting. Similarly, different visualization techniques vary in how effectively they enable the completion of a given task (i.e., their task affordances). Continuing with the analogy, the task demands and the capabilities of one's muscles provide a range of attainable performance outcomes for lifting the object. The amount of *effort* one applies to lift the object modulates the actual outcome, which is measured by both one's success in lifting the object to the required height and the speed at which this objective is achieved. Using this logic, it could be argued that *cognitive effort* modulates the speed and accuracy of a person's performance related to a given visualization task.

However, numerous visualization researchers (e.g., [33, 34]) and cognitive scientists (e.g., [38, 44, 94]) have pointed out issues with measures of speed and accuracy. For example, trade-offs between speed and accuracy are not fixed across tasks or users [44], speed and accuracy can lack the required precision to measure cognitive effort [94], and different levels of effort can produce the same accuracy and speed responses [33] (reviewed in Section 2.1). As Kyllonen and Zu state in a review of response time measures:

> A respondent may respond quickly and correctly because of a lucky guess, or slowly and correctly, but could have answered correctly quickly if incentivized to do so. If a respondent does not answer correctly, it could be due to not knowing the answer, not spending enough time to process the information fully, or having gotten confused while answering and quitting. [44, p. 1].

Measuring speed and accuracy alone can be problematic, but these measures are still necessary metrics for many contexts in visualization research. To offset some of the issues associated with speed and accuracy, we advocate for a converging measures approach where validated measures of cognitive effort are used in conjunction with speed and accuracy.

To use converging measures to evaluate mental effort in a visualization context, first it is important to clearly define effort. Psychologists commonly use the term *working memory* to describe a large component of mental effort, and the definition of this term remains a hotly debated topic. In line with a recent review on decision-making with visualizations [62], one definition useful for visualization research sug-

- *L.Padilla, is with the University of California Merced in Cognitive + Information Sciences. E-mail: Lace.Padilla@UCMerced.edu*
- *S. Castro, I. Ruginski, and S. Creem-Regehr are with the University of Utah Department of Psychology*
- *S. Quinan is with the University of Utah School of Computing*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org.
Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx/

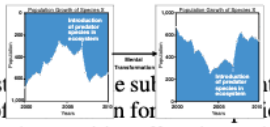

Fig. 1: The mental transformation required to correctly order the Y-axis of a figure originally published in Padilla et al. [61].

gests that working memory consist e sul ıts of the mind that hold a limited amount of n for iod [20]. For example, in a study comparing the cognitive effort demands of network diagrams in which some of the diagrams were filtered to show only the task-relevant connections, Huanga et al. [33] found that individuals who viewed non-filtered diagrams had greater self-reported cognitive load than those who viewed filtered diagrams. Presumably, individuals who were shown all of the connections in the diagrams had to direct their attention to only the task-relevant information in the display. Cognitive science theories propose that this type of cognitive control and attention direction requires significant working memory (e.g., [42, 81]).

Cognitive science has developed and systematically tested various measures of working memory fluctuations during a task, including dual-task experimental designs (simultaneous performance of multiple tasks compared to single-task performance), pupillometry (the dilation of one's pupil), and neurological changes (for reviews, see [3, 48, 52]). As summarized in Table 1, in the context of visualization research, there are pros and cons to each of these working memory evaluation techniques. One such consideration is the *validity* of the measure, which can be used to formalize comparisons across empirical measurement techniques [61]. Validity refers to how closely the conditions of an experiment match real-world conditions (ecological validity), the generalizability of the findings to other contexts (external validity), and its ability to measure what it claims to measure (construct validity) [61].

The goal of this work is to provide practical cross-domain methodological guidance for objectively evaluating working memory demands in data visualizations. We focus on the most feasible measures of working memory for visualization researchers with high construct validity: pupillometry and dual-task experimental designs. The key contributions of this work include a critical discussion of working memory evaluation techniques, a detailed outline of empirically tested working memory demanding tasks, a case study comparing multiple methods for measuring working memory effort in a complex visualization task, and a discussion of open questions concerning objective evaluations of visualization quality.

2 Related Work

Utilizing pupillometry and dual-task experimental designs requires a foundational understanding of working memory theory. The study of working memory has been a vibrant topic in cognitive science for over half a century [7], and thus an exhaustive discussion of the nuances in this field is beyond the scope of this paper. However, in the following sections, we detail the key concepts in working memory theory and demonstrate how visualization researchers can use working memory theory to evaluate visualization quality with pupillometry and dual-task experimental designs.

2.1 Working Memory

One of the most hotly debated topics in cognitive science centers around the exact definition of working memory [20]. The broadest definition suggests that working memory consists of various components that can hold a limited amount of transformable information for a finite period [20, 62]. There are two critical concepts in this definition: 1) working memory is capacity limited [19, 23], and 2) working memory functions decay over time (e.g., [19, 45]). For example, historical research by Miller [56] found that people can remember seven numbers +-two. Researchers have updated [57], refined [25], and challenged [25] this work since its introduction in 1956. This follow-up work collectively illustrates that the amount of information that we can hold in our mind for a short time is limited in capacity. The second concept in this broad definition of working memory theory is that it diminishes over time. Researchers such as Cowan et al. [22] suggest that our ability to store information begins to decay after approximately 5-10 seconds. The specifics of temporal decay vary based on the task, type of information, and capacities of the participant [22]. Active rehearsal of information increases temporal capacity, but most definitions of working memory include a passive storage component that is time-limited without rehearsal or reactivation [20]. For example, Cowan et al. [21] demonstrated that when rehearsal was prevented, working memory task performance still correlated highly with cognitive aptitude measures. Finally, contemporary theories also stress that we use working memory to guide our attention to task-relevant information and suppress automatic responses [42, 81]. Researchers have integrated working memory theory into models of visualization comprehension [66, 70], which have been updated by the Padilla et al. [62] Model of Visualization Decision-making to include modern approaches to working memory and decision-making theory.

Visualization studies commonly assert that some visualization techniques can reduce cognitive effort (e.g., [31, 88, 8]). Researchers commonly use cognitive load theory to describe the influence of visualizations on working memory, which proposes that increases in working memory are due to extraneous cognitive load (i.e., load associated with the information communication method [60]). Researchers propose that different visualization techniques vary in the amount of extraneous load they demand of the viewer and thus differ in working memory demand [33, 87, 36].

Researchers have attempted to understand the cognitive mechanisms and processes that result in increases to cognitive load associated with visualizations. For example, numerous studies find that visualizations help users to more quickly and/or accurately perform spatial tasks compared to textual representations of the same information (e.g., [77, 83, 93]). Vessey and Galletta [93] propose that the reason visualizations are more useful for spatial tasks involves the concept of *cognitive fit*. Cognitive Fit Theory asserts that visualizations can *match* how users naturally think about the data, and when there is a match, little working memory is needed [93]. For example, when spatial data is represented spatially (e.g., geospatial data shown in a map), there is a match, and little effort or working memory is required to use the visualization effectively. However, when spatial data is represented textually (e.g., geospatial data depicted in a table), the viewer needs to mentally transform the tabular data into a geospatial context [46], which requires significant working memory. When there is a mismatch among any configuration of the users' mental representation (mental schema) [46], the visualization, and the task, the users must mentally align the mismatched components, which requires significant working memory [93]. Take, for example, the deceptive visualization practice of Y-axis reversal, shown in Figure 1 [61, 62, 64]. Figure 1 shows the population of Species X over time and indicates the point at which a predator was introduced into the ecosystem. If viewers make a snap judgment, they will wrongly assume that the predator species was responsible for a drop in the population of Species X [64]. However, viewers can apply working memory to complete a mental transformation such that the Y-axis is correctly ordered (Figure 1 right), and draw inferences from this updated mental representation. Then viewers would see that the introduction of the predator species correlates with a rise in the population of Species X [61, 62].

In addition to increasing with mental transformations, working memory demand also increases with task difficulty [41]. For example, we postulate that low-level tasks such as *retrieve value*, *find extremum*, or *find anomalies* [2] likely require less working memory than higher order tasks such as discovering novel patterns in the data. This assertion is in line with work by Kahneman and Beatty [41], where pupil dilation was found to increase along with difficulty for detect-

Table 1: Comparison of working memory evaluation techniques, in the context of visualization research.

Approach	Measures	Pros	Cons and Practical Considerations
Standard	Accuracy, speed [31, 93], and self-report data [97, 33]	Easy to implement and compare findings to prior work. Measures have real-world implications. Can have high ecological validity.	Possible confounds (speed and accuracy trade-offs and motivation). Users are often unaware of working memory demand. Can lack sensitivity to working memory differences. Possibility of low construct validity. External validity will vary with experimental conditions.
Dual-task designs	Dual-task cost associated with speed and accuracy [8, 88]	Empirically validated measure of cognitive effort. Possibility of high ecological validity in contexts where users are distracted. Relatively greater construct validity than basic experimental designs that measure speed and accuracy.	Possible confounds (speed and accuracy trade-offs and motivation). Some inconclusive dual-task cost effects in prior work [8, 88]. External validity will vary with experimental conditions.
Pupillometry	Pupil dilation	Empirically validated measure of working memory, with relatively greater construct validity than speed and accuracy measures and possibility of high external validity. Straightforward to measure with most eye trackers. Highly sensitive to working memory changes.	Eye-tracker needed and eye tracking data processing required. Highly sensitive to image luminance. Possibility of high external validity, which can vary with experimental conditions.
Neurological changes	Electrical signals in the brain using electroencephalography (EEG) [4], and hemodynamic responses with near-infrared spectroscopy [68]	Greater construct validity than speed and accuracy measures. Possibility of high external validity.	Requires access to expensive equipment and equipment technicians. Findings can be difficult to interpret. External validity may vary based on the hypotheses of the experiment.

ing an auditory tone. Numerous subsequent studies similarly confirm that working memory increases with task difficulty (e.g.,[1, 90]). For a given task, there are a variety of ways to manipulate task difficulty. For example, using *identify* and *compare* from the query phase of Brehmer and Munzner's visualization task typology [13], an easy identification task would include identifying distinct patterns, and a difficult version would involve identifying complex or obscured patterns. Likewise, an easy comparison task would be to compare two values, and adding additional comparisons would increase the task difficulty.

Researchers often use behavioral measures, such as time to complete a task, accuracy, or subjective measures like self-report survey data as a comparable metric for working memory (e.g., [11, 33, 88, 97]). A variety of issues can arise when using speed, accuracy, or self-report measures alone as the only tests of working memory [48, 91], particularly in a visualization context. Concerning speed and accuracy, there are widely documented trade-offs that can make it difficult to attribute findings to any one source [44, 50]. Take, for example, a simple illustration proposed by Just et al. [38], where an addition problem (3 + 2 + 5 + 2 + 3 + 1 + 4 + 6 + 1 = ?) is compared to a multiplication problem (63 x 5 = ?). Just et al. [38] argue that the addition problem requires fewer mental resources but may take longer to complete. In a visualization context, users may enjoy a visualization and engage with it more, or the visualization could spur curiosity and additional data exploration, which could slow down the users but could also improve performance outcomes (reviewed in [34]). Additionally, trade-offs between speed and accuracy are not fixed across tasks or users [44]. For example, in one study, occupational psychologists found that in addition to working memory, a person's job-perception, alertness, and time pressure ultimately drive the outcomes of time on task, accuracy, and the sustainability of mental effort [55]. Finally, significant performance differences might not be observable in statistical analyses even if there are significant differences in working memory demands, as speed and accuracy can lack the required precision to measure working memory demands, particularly for binary decisions [47, 94]. Speed or reaction time distributions are particularly problematic for measures of central tendency as they tend to be positively skewed with outliers, reducing the ability of Analyses of Variance (ANOVA) to detect significant differences [72].

Self-report measures of cognitive effort [33] such as the NASA-TLX [97] include their own caveats and limitations. For example, when users can no longer monitor the working memory required by a task at a meta-cognitive level, self-report measures under-represent working memory demands [48, 96]. In fact, McKendrick and Cherry [54] demonstrated that various NASA-TLX sub-scales, such as perceived effort and perceived performance, correlated more with individual participants' random variation than behavioral outcomes of either in a spatial-memory task.

Some visualization researchers have sought to employ physiological measures of working memory other than speed and accuracy in a visualization task [3, 4]. Notably, Anderson et al. [4] demonstrated the use of electroencephalography (EEG) to measure voltage fluctuations resulting from ionic current within the neurons in the brain, while comparing several visualization techniques of boxplots. Although measuring electrical activity in the brain has greater construct validity than measures of speed and accuracy, this approach can be impractical for most visualization researchers. In this paper, we focus on measures that are both precise and feasible for visualization researchers: pupillometry and dual-task paradigms.

2.2 Pupillometry

Guillaume de Salluste proposed that the eyes are the windows to the soul [35], and more scientifically Michel Pierre Janisse described them as "the only visible part of the brain" [35, p. 1]. Researchers in psychology have been investigating the relationship between pupillometry and mental effort since the 60s [40], and this work has seen a resurgence in recent years (e.g., [82, 90]). Recently, HCI researchers have also called for the use of eye-tracking measures to evaluate cognitive load [36, 87]. The high-level conclusions from pupillometry research suggest that pupil dilation is highly correlated with working memory [90]. Pupillometry is currently a commonly used evaluation method of working memory in numerous applications, such as the cognitive state of drivers [48]. this paper, the term *pupil dilation* refers to the increase in pupil diameter associated with the execution of a task compared to the baseline pupil diameter measured when the viewer is not completing the task [90]. For example,

> Face a mirror, look at your eyes and invent a mathematical problem, such as 81 times 17. Try to solve the problem and watch your pupil at the same time, a rather difficult exercise in divided attention. After a few attempts, almost everyone is able to observe the pupillary dilation that accompanies mental effort... [39, p. 24].

As Daniel Kahneman illustrates in this exercise, a long lineage of research has revealed that our pupils dilate when we exert effort (e.g., [40]; for review, see [52, 90]).

Table 2: Summary of secondary tasks.

Task	Examples	Difficulty	Modality	Mental Process
Memory Span	Remember a series of N numbers, words [84], or sounds [18]	*Easy* ($N \leq 3$) to *hard* ($N \geq 6$)	Visual or auditory	Likely *semantic* but potentially spatial[†]
Operation Span	Reading or listening to a series of mathematical operations and/or logical statements [20]	*Medium* to *hard*[‡]	Visual or auditory	Likely *semantic* but potentially spatial[†]
Visuospatial Memory Span	Remember a visual array of items [28] or sequence of visual information [78]	*Easy* to *hard*[§]	Visual	Likely *spatial* but potentially semantic[†]
Continuous Sequence	Listening to a series of words or digits, then recalling an item from N places back [52]; or counting backwards by threes [65]	*Easy* (0-back) to *hard* (+2-back)	Visual or auditory	Likely *semantic* but potentially spatial[†]

[†] Depends on strategy employed
[‡] Depends on series complexity
[§] Depends on visual information complexity

Using neuroimaging techniques such as fMRI and two-photon microscopy, numerous studies find correlations between pupil dilation and working memory (e.g., [38, 59]). There has been some debate concerning the exact brain regions that are responsible for the pupil dilation associated with working memory (for a full discussion of this topic see [90]). Suffice to say, mounting evidence indicates that the mental effort induced dilation response created by the sphincter and dilator muscles in the pupil is highly related to the *locus-coeruleus norepinephrine* (LC-NE) system [37, 52]. Researchers suggest that the LC-NE system monitors the environment of cognitive demands [12] and optimizes effort [79], among other functions.

In addition to the neurobiological support for pupillometry, measures of pupil dilation are relatively easy to collect with most eye trackers [90], making eyetracking methodology well suited to visualization research [36]. However, as with all physiological measures, pupil size is influenced by numerous factors, such as drowsiness, stress, drug use, or luminance in the environment. Small changes in luminance can have a significant effect on pupil dilation and, therefore, the luminance of the stimuli, the lighting conditions of the room, and the calibration of the monitor(s) should be carefully controlled. By systematically controlling for variations in conditions, either within the experiment or with statistical procedures, researchers can gain a relatively objective measure of working memory [9].

2.3 Dual-Task Paradigms

Cognitive science has developed a *dual-task paradigm* for comparing the relative differences in cognitive workload between tasks, which involves the simultaneous completion of two tasks [43]. In dual-task paradigms, speed and accuracy measures are compared during primary task performance and dual-task performance. During dual-task performance, the participant completes the primary task and a working memory demanding secondary task at the same time (see Step-by-step Guide 1: Dual-task Designs). If significant working memory is required for the primary task, adding a working memory demanding secondary task will overload the capacity limited working memory system resulting in longer task completion times and more errors. *Dual-task cost* is the relative decrease in performance between the single- and dual-tasks. Tasks that require more working memory will demonstrate significantly larger dual-task costs. Dual-task paradigms have been used to demonstrate performance decrements in many applied settings, including visual and cognitive distractions while driving [16], interacting with technology [14], visually searching for remembered objects [26], and performing visualization tasks [8, 88].

In a dual-task experimental design, the secondary task should be assessed for three key factors: the relative difficulty, the modality of the task (e.g., visual or auditory), and the mental process (e.g., spatial or semantic) (see Table 2 for an overview of common secondary tasks). It is essential to select a secondary task that is an appropriate level of difficultly in relationship to the primary visualization task to gain relevant information about that primary task. If the secondary task is too difficult, the user will likely be unable to complete both tasks at the same time. For example, in driving when the secondary task becomes too difficult, the driver must either stop the secondary task or crash [15]. However, if the secondary task is too easy, the dual-task cost will not be observed. Listening to the radio while driving, for instance, does not have much of an effect on lane deviation, speed changes, or braking times [85]. The goal is to provide users with a secondary task that requires some of their pool of cognitive resources but does not prevent them from accomplishing the primary task. To simultaneously tax users and avoid inordinately impacting the primary task, researchers have developed a set of commonly used secondary tasks [43]. In the following passages, we detail previously used secondary tasks that are well suited to visualization primary tasks.

Simple Memory Span Secondary Tasks
Mental operation tasks are one of the most commonly used classes of secondary tasks, and they are well suited to use with visualizations as they do not interfere with the response modality of visualization tasks. These tasks require the user to remember and/or mentally manipulate information in memory. The simplest of these tasks are *memory span* tasks, which require a user to remember a sequence of numbers (i.e., digit-span) [57], words [88], or sounds [20]. Within the memory span tasks, the difficulty of a task can be manipulated by increasing the number of items to remember. For example, in the digit-span task, a participant must remember a series of numbers, with seven to nine being the upper bound of digits we can remember [57].

A substantial body of research explores how we remember information, which suggests that we commonly chunk information together to reduce mental effort. Chunking has been shown to rely on long-term memory processes to aid retrieval [73], but has also been shown to occur in immediate memory (i.e., without long-term memory consolidation) based on encoding of patterns or order [17]. A strategy to discourage chunking involves the concept of *semantic distance*, which is how closely the meaning of words are related to one another [74]. Words may belong to salient categories such as 'fruit' or 'yellow objects' that can have a short semantic distance. Words within the same category can be chunked with other members of the category making them easier to remember. For numbers, consider checking that the numbers do not repeat and are not all even, multiples of 3, or other salient groupings.

Operation Span Secondary Tasks
Another solution to chunking is to have participants manipulate information in an operation span (OSPAN) task. These tasks usually involve reading or listening to a series of mathematical operations, logical statements, or a mix of both. Some researchers [20] consider reading and listening span tasks as the gold standard of working memory tasks, because these tasks require participants to both remember and manipulate information. In a listening span task, participants listen to sentences and judge the sentence to be true or false, while also remembering the last word in each sentence [23]. At the end of the span (i.e., four to five sentences), participants are asked to recall the last word of each sentence out loud in order. For reading and listening span tasks, difficulty can be modulated by the number of sentences participants need to remember before recall. One challenge with OSPAN tasks is the constantly changing number of items in memory across the span. As each new sentence is presented, the required working memory increases, making for a more variable working memory manipulation.

Visuospatial Memory Span Secondary Tasks
The counterparts of operation span tasks are visuospatial working memory tasks [78, 92]. Many working memory theories suggest that a dedicated component of working memory specifically stores and manipulates visuospatial information, commonly termed the *visuospatial sketchpad* [6]. Visualization studies likely involve visuospatial working memory. An advantage of using a spatial working memory sec-

ondary task is that the task will likely tax the main resource pool that participants will be using for the visualization task. However, there is the possibility that even a simple spatial working memory task will be too taxing to complete at the same time as a visualization task. A sub-group of dual-tasking researchers refers to this effect as *dual-task interference*, although members disagree about the extent of interference across difficulty and modality [65, 95]. Additional research is needed to evaluate the impact of using a visual-spatial working memory task as the secondary task while completing a visualization task.

One example of a visuospatial span task is to show participants an image with an array of items such as circles with various colors in various orientations. Participants must remember the color and location of the circles while completing the primary task. Once the primary task is complete, a circle of a particular color appears and the participants are asked to report if that circle is in the remembered location [30] (for more examples see, [78]).

Continuous Sequence Secondary Tasks

Continuous sequence secondary tasks are often referred to under the umbrella of *N-back tasks*. The simplest version (i.e., the 0-back task) requires participants to repeat visual or auditory information vocally or with a key press immediately after its presentation, which continues at a constant pace throughout the task. As one of the hardest secondary tasks, auditory N-back tasks beyond the 0-back task require the participants to listen to a continuous string of words or digits, remember their order, and then recall a digit from N places back in the order (usually out loud). For example, in a 2-back task, participants report the digit two places back from the most recent number provided. In a sequence of 5, 2, 8, 9, 1 . . ., participants would not respond to the first two digits, then after hearing 8 say 5, and after hearing 9 say 2 [58]. Another popular version of a continuous updating task requires participants to count backwards by 3s starting with a large number, such as 1986 [69].

Step-by-step Guide 1: Dual-task Designs

1. Select the primary visualization task based on your specific context.
2. To select an appropriate secondary task, identify the level of difficulty of the primary visualization task. See Section 2.1 for discussion of task difficulty.
3. Select a secondary task with an appropriate level of difficulty in relation to the primary task. If you have a hard primary task, select a medium or easy secondary task (see Table 2 for guidance).
4. Decide whether you will use a between-, within-, or mixed-subjects design. In all versions, when comparing two visualizations there will be four conditions: **(1)** completion of the primary task with visualization *A*. **(2)** simultaneous completion of the primary and secondary tasks with visualization *A*. **(3)** completion of the primary task with visualization *B*. **(4)** simultaneous completion of the primary and secondary tasks with visualization *B*.

   Between-subjects design: Participants are randomly assigned to groups 1–4 and each group completes one condition. Analyze the primary task performance between each group. **Within-subjects design**: Every participant completes all 4 conditions. Analyze how each user's performance changes due to both the visualization and dual-tasking. **Mixed-design**: Participants are randomly assigned to one of two groups. Group 1 completes conditions 1–2. Group 2 completes conditions 3–4. Analyze the changes in performance, per person, between the single and dual-tasks (i.e., dual-task cost) across groups.

3 Case Study

The goal of the current study is to illustrate a converging methods approach for evaluating working memory demands in a visualization task using a dual-task experimental design and pupillometry. To illustrate these methods, we draw on a geospatial visualization task from Padilla et al. [63] for our primary task. This geospatial visual aggregation task was selected to illustrate the application of a dual-task paradigm in the context of a complex visual-spatial task with medium difficulty. Medium difficulty was ideal to demonstrate the influence of a secondary task. The task consisted of presenting viewers with a Digital Elevation Model (DEM), which was visualized using a continuous grayscale encoding (see Figure 3). For each DEM, red squares were superimposed on the figure. Participants were tasked with deciding which area within the red squares contained the highest average elevation. This task required participants to mentally aggregate the elevation data within each square and then compare the average values among the squares. But a binary choice task is not representative of many visualization tasks, it was chosen because this design can be manipulated to require more or less working memory in order to illustrate dual-task cost effects in a controlled and predictable way. In the current study, the difficulty of the task was manipulated by having the viewer compare the average elevation between either two regions or four regions, as shown in Figure 3. This difficulty manipulation and binary task was selected because the increase in difficulty was known a priori (i.e., the probability of randomly selecting a correct response from two regions is 50% vs. 25% for four regions). Our goal was simply to illustrate the use of multiple working memory measures for visualization tasks, to provide an example of how to compare the working memory demands between visualizations. Working memory measures may be more informative for tasks that show a more substantial speed/accuracy trade-off, or those in which efficiency measures are less meaningful, such as tasks that focus on exploration and insight.

The secondary task that we selected was a version of the simple memory span task, in which participants were tasked with remembering seven non-repeating numbers while completing the primary visual-spatial aggregation task. The memory span was selected because it does not require spatial working memory, which might interfere with the spatial aggregation task. Further, the memory span task selected is of medium difficulty, which allows for appropriate dual-task cost during both easy and hard visual-spatial aggregation tasks.

3.1 Hypotheses and Methods

In line with prior dual-task research (e.g., [30, 65]), we predict that the comparison of the average elevations of two regions will require less working memory than four regions. The present work seeks to demonstrate the increased working memory associated with more complex visualization tasks by evaluating dual-task costs in speed, accuracy, and pupil dilation. Specifically, we predict the following:

H1. *Response times and errors will increase with task difficulty.* Observation of a main effect of task difficulty on speed and accuracy would provide behavioral evidence that working memory increases with task difficulty. This finding would also provide evidence that tests of speed and accuracy are sensitive enough to measure the rise in working memory required by spatially aggregating four regions compared to two regions.

H2. *Response times and errors will increase when completing both the primary and secondary tasks compared to completing only the primary task.* This finding would provide behavioral evidence that the secondary task required significant working memory. Further, this result would propose that measures of speed and accuracy are sensitive enough to pick up on the additional working memory required by a simple memory span task, in the present context.

H3. *There will be a greater increase in errors and response times from the easy to hard tasks for dual-taskers compared to single-taskers.* More substantial dual-task costs for harder tasks would provide converging evidence that harder visualization tasks require more working memory than easier ones, compared to simple measures of speed and accuracy. The contribution of evaluating the relative dual-task costs is that prior research finds that dual-task cost is highly correlated with working memory effort using neuroimaging measures (e.g., [32]).

H4. *Pupil diameter will increase with task difficulty.* A main effect of task difficulty on pupil dilation would demonstrate the capacity of pupillometry to measure differences in working memory when mentally aggregating four compared to two regions.

H5. *Pupil diameter will increase when completing both primary and secondary tasks compared to single-task performance.* The main effect of dual-tasking on pupil dilation would indicate that pupil dilation can be used to measure the increased working memory associated with completing a secondary task.

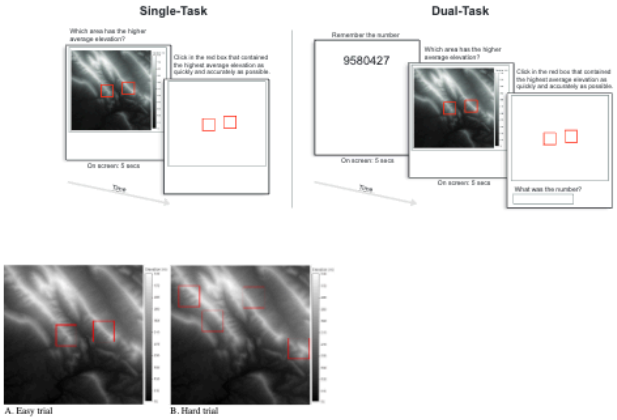

Fig. 2: Diagram depicting the experimental design sequence for the single-task (left) and dual-task (right) groups for an easy trial.

Fig. 3: Example stimuli representing easy trials (A) where two regions were indicated by red boxes and hard trials (B) with four regions. For each stimuli, participants were instructed to identify the red square that contained the highest average elevation.

H6. *Pupil diameter will be largest when dual-tasking and completing a hard task.* Larger pupil dilation dual-task cost for harder tasks compared to easier tasks would suggest that the combination of dual-tasking measures and pupillometry can successfully magnify the influence of working memory and measure it successfully. The combination of dual-tasking and pupillometry can be desirable if important differences in working memory are hard to detect due to the nature of the tasks or because of simplified experimental lab settings. Pupillometry should provide compelling evidence of working memory, as a large body of research consistently finds a correlation between pupil diameter and effort (for review, see [90]). Physiological measures also reduce the variability that is produced by subjective reports of mental processes, which can be highly variable [48].

The current study employs a mixed between- and within-subjects design, in which one group of undergraduate students from the Psychology Department at the University of Utah (n = 20, mean age = 26.2, SD = 8.9, male = 6, female = 14) completed the dual-task paradigm, while a second group completed only the primary visualization task (n = 20, mean age = 24.6, SD = 9.2, male = 5, female = 15). These naive participants received no directed training on how to read maps in the psychology curriculum. Participants received course credit for participation in this study, which received IRB approval prior to data collection. To test H4-H6, the current study also measured participants' pupil diameter using a Seeing Machines Fovio^{TM} Eye Tracker and EyeWorks^{TM} recording software [28].

3.2 Stimuli Generation

The experimental stimuli were generated from ten non-contiguous regions of a 1x1 degree tile of the 1/3 arc-second DEM from the USGS National Map 3D Elevation Program [89]. In line with similar work by Padilla et al. [63], we normalized the data of each selected DEM region and then mapped it to the lightness channel (L*) in CIELAB while leaving $a^* = b^* = 0$. This process effectively encodes the data with a perceptual grayscale color map. We then converted the resulting color-mapped DEM regions to sRGB images for display. An overview of the selected regions and their respective data ranges have been included as supplemental material (osf.io/6u8em).

The resulting ten sRGB images were used as *basemaps* for the experimental stimuli. From each basemap, we created both *easy* and *hard* stimuli, as shown in Figure 3. The *easy* stimuli highlighted two regions and the *hard* stimuli highlighted four. Using the Mahy et al. [51] benchmark for just-noticeable color differences (JNDs) in CIELAB ($\Delta\ E^*_{ab} = 2.3$), the highlighted regions were selected to have mean values that differed by two to four JNDs (mean $\Delta\ E^*_{ab} \in [5.3, 7.8]$), and then were layered on the sRGB basemaps in Photoshop. We also added a legend specifying a constant elevation range of [75, 510] meters to each stimulus, as the actual data values are not important to the experimental tasks. The relative perceived differences were our primary interest, which we modeled and control for via JNDs.

Half of the regions were chosen to ensure that merely selecting the region with the highest point did not lead to correct responses. These *trick trials* controlled for participants who were not doing the full mental aggregation across regions but were instead using a simpler strategy of selecting the region with the single highest point, as observed in Padilla et al. [63]. To generate additional trials, we rotated (0°and 90°) and reflected (no reflection and reflection over the vertical center line) the ten basemaps, which resulted in a total of 40 trials for the easy condition and 40 trials for the hard condition.

3.3 Task and Design

Participants were randomly assigned to either the single or dual-task group. After signing a consent form, participants in the single-task group received instructions for the task and then one practice trial. For each trial, participants were randomly shown either an easy (i.e., two regions) or hard (i.e., four regions) stimulus for five seconds. After five seconds the experiment would progress to a response screen that included only the red squares in the same location as on the stimuli but on a white background (See Figure 2). Participants were instructed to click as quickly and accurately as possible in the red squares that previously contained the highest average elevation. Click speed and accuracy were recorded. Participants completed 40 trials (20 easy and 20 hard, randomly ordered) and then took a mandatory two-minute break and completed the other 40 trials (20 easy and 20 hard, randomly ordered).

The participants in the dual-task group completed the same primary task as those in the single-task group but before viewing each stimuli they were shown a different randomly selected seven-digit non-repeating number for five seconds. Participants were instructed to remember each corresponding seven-digit number while completing

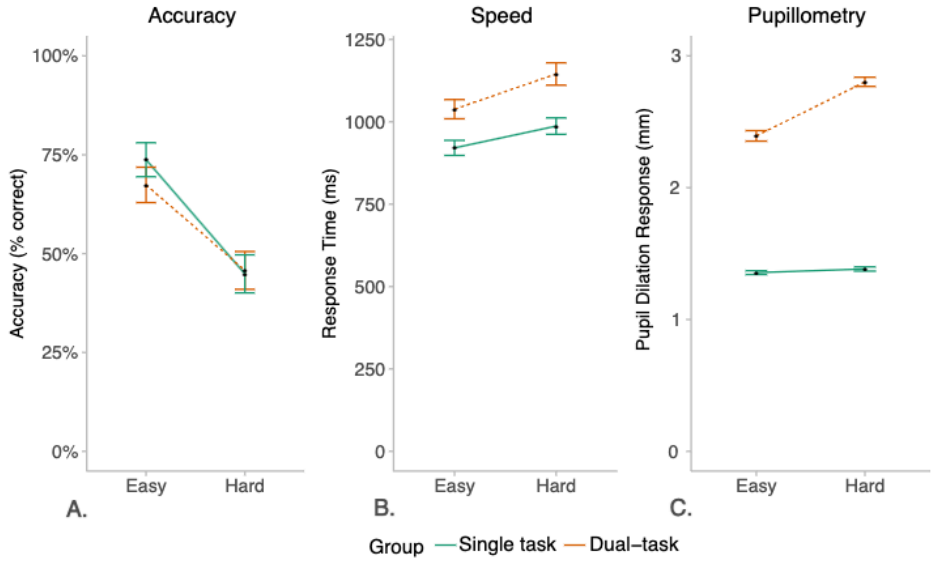

Fig. 4: Each plot shows the effects of the experimental group and task difficulty on accuracy (A), task completion time (B), and pupil dilation (C). Error bars show +/- 95% confidence intervals. Error bars for pupilometry were created by taking the average pupil dilation per trial.

each trial of the primary task. After indicating the region with the highest average elevation, they then entered the number they remembered in a text box. The accuracy of the number that they entered in the text box was recorded.

All participants completed the experiment on the same laptop in the same location in a room one at a time. The visual angle of the laptop screen was 24.08°vertical and 41.34°horizontal. The eye tracker was placed at the bottom of the screen and angled at 30°. The Seeing Machines Fovio Eye Tracker and software was calibrated using a four-point calibration procedure that was conducted twice, once at the beginning of the experiment and again after the two-minute break. The resolution of the laptop display was set at 1920 x 1080.

3.4 Analyses

Multilevel logistic regression models using the lme4 package [10] in R [71] were employed to account for variance in user accuracy, speed, and pupil dilation with task difficulty (easy and hard trials) and experimental group (single and dual-task groups) as predictors. Multilevel models were utilized due to their appropriateness for nested data structures (in this case, repeated measurements within persons) and binary outcome data. The following sections detail the specific findings from the accuracy, speed, and pupil dilation analyses. In each of the models, we included the covariate interactions of *trick trials * experimental groups* and *trick trials * task difficulty*. These covariates allowed the analysis to control for cases where individuals used the strategy of selecting the region with the highest single point rather than comparing the average elevation per region as instructed. The strategy of selecting the region with the highest single point was observed in Padilla et al. [63] and does significantly predict variance in accuracy, speed, and pupil dilation. However, as this effect is specific to the task and not relevant to the goal of illustrating a dual-task experimental design with pupillometry, we did not detail the results of the trick trial covariate terms in the body of the paper. The full output of our models including covariates is available in the supplemental materials, along with the data and executable code to generate the analyses. The statistics in the following sections control for the effect of trick trials. In other words, all effects reported are after removing any effect of trick trial on accuracy, speed, and pupil dilation.

3.5 Accuracy

Each trial was considered either correct (1) or incorrect (0). We checked to determine if dual-task participants were completing the secondary task prior to analysis. Overall, individuals in the dual-task condition answered the correct number in the secondary task 64.6% of the time (SD = 47.8%). Individual's mean correct responses on the secondary task ranged from 18.8% correct to 98.8% correct, giving us confidence that we had selected a sample with a representative spread of working-memory capacity in the general population. Prior to analyses, experimental group (single-task = -0.5, dual-task = 0.5) and task difficulty (easy = -0.5, difficult = 0.5) were contrast coded to test group differences, and trick trials were contrast coded (non-trick = -0.5, trick = 0.5) as a control variable. The fixed factors included in the multilevel model consisted of the interaction between *experimental group * difficulty*, interactions with trick trials, the average difference in L* between regions within a given image, and the lower order terms. The average L* difference was added to control for any effects that the limited variation in JND averages across different stimuli regions might have had on task difficulty. Participants were included as random effects. Note that the accuracy effects (β) are reported using log odds in the text, but converted to accuracy odds-ratios for easy interpretation in the figures. Odds are defined as the ratio of correct responses to incorrect responses, and an *odds-ratio* is calculated as the ratio of the odds of two groups. An odds-ratio of 1 indicates the same odds of answering correctly between the groups, an odds-ratio below one indicates that the group coded 0.5 has lesser odds of answering correctly, and an odds-ratio above one indicates that the group coded 0.5 has greater odds of answering correctly.

In examination of **H1**, more difficult trials (with four regions) elicited significantly worse accuracy on average than easy trials (with two regions) (see Figure 4 A; β = -1.25, SE = 0.08, *Odds-ratio* = 0.29, $p < .001$, *95% CI* [-1.41, -1.09]). As we deliberately selected tasks where chance performance for two regions was 50% and 25% for four regions, the finding that performance was significantly worse for the harder trials confirms our manipulation.

For **H2**, our analysis did not provide evidence that individuals in the dual-task condition differed in accuracy from individuals in the single-task condition (β = -0.14, SE = 0.14, p = 0.32, *95% CI* [-0.42, 0.14]). Although there was not a main effect of experimental group, an interaction between experimental group and task difficulty was observed for accuracy.

To evaluate **H3**, we examined the interaction between experimental group and task difficulty. The results of our analysis suggest that dual-task costs are influenced by the difficulty of the task but not in the specific way we predicted (β = 0.35, SE = 0.16, p = 0.03, *95% CI* [0.03, 0.66]).

To examine the significant interaction between experimental group and task difficulty, we conducted two post hoc multilevel models in which we recoded the task difficulty variable. One model included

task difficulty recoded such that the easy trials were the referent (easy = 0, hard = 1) and the other with the hard trials was the referent (hard = 1, easy = 0). In each of these models, the interaction between *difficulty and experimental group* was used to predict the variance in accuracy along with the lower order terms, and the average *L between regions for each stimulus was included as a covariate. This analysis revealed that there was a main effect of experimental group for the easy trials (β = -0.30, *SE* = 0.15, *Odds-ratio* = 0.74, $p = 0.049$, *95% CI* [-0.61, -0.0003]) but not for the hard trials ($\beta = 0.03$, *SE* = 0.15, *Odds-ratio* = 1.036, $p = 0.81$, *95% CI* [-0.25, 0.32]). These results suggest that for easy trials, individuals who completed only the primary task demonstrated significantly better performance than those who completed both tasks, and that there was no difference in performance for the hard tasks between groups. These findings do not support our predictions in H3.

Key findings from the analysis of accuracy:

- Individuals in the single-task group had greater accuracy with the easier trials compared to individuals in the dual-task group.

The findings of this analysis are in line with prior work that revealed inconclusive evaluations of visualizations using only measures of accuracy in a dual-task paradigm [88]. We propose that measures of accuracy, particularly binary judgments (e.g., correct or incorrect), may not be fine-grained enough to pick up on important differences in working memory between visualizations. Further, we advocate for a converging methods approach, where several evaluation techniques are utilized, including task completion times and pupil dilation, as detailed in the following sections.

3.6 Task Completion Time

In addition to accuracy, the speed by which users clicked in their selected region was recorded. When participants accidentally clicked multiple times or changed their selection, the click speed associated with their final selection was used in the following analysis. In addition, any trial where an individual's mean click time was greater than three standard deviations within their group was removed prior to analysis. In total, 51 out of 3200 trials (1.6%) were removed as outliers. We used the same model as in the accuracy analysis but with task completion time as the outcome variable and a Gaussian distribution was fitted.

Our results revealed that when individuals completed the difficult trials, they reacted 83 ms slower on average than when they did the easy trials (see Figure 4 B; $\beta = 83.42$, *SE* = 10.16, $p < .001$, *95% CI* [63.5, 103.3]), which supports **H1**. In line with prior work [90], this result provides converging evidence that working memory demands increase along with task difficulty.

In evaluation of **H2**, we found that individuals in the dual-task group reacted 141 ms slower on average across all trials than individuals in the single-task group ($\beta = 141.04$, *SE* = 56.71, $p < .012$, *95% CI* [30, 252]). This finding supports H2 and provides behavioral evidence that more working memory is required to do two tasks at the same time than one task in a visualization context. The differences in response times between individuals in the dual-task and single-task groups can be seen in Figure 4 B.

For **H3**, the results of this analysis did not reveal a significant interaction between experimental group and task difficulty, meaning that individuals in both groups showed the same relative changes in response times for easy and hard trials.

Key findings from the analysis of task completion time:

- Increased task difficulty resulted in significantly longer response times.
- Individuals who simultaneously completed both the primary and secondary tasks responded significantly slower than individuals who completed only the visualization task.
- Task completion time was not sensitive enough to measure dual-task costs.

The response time findings expand our understanding of the working memory associated with difficulty and dual-tasks, beyond an analysis of accuracy alone. However, measuring only the swiftness of a viewer's response may not be sensitive enough to provide a full picture of the working memory utilized in these tasks. For this reason, we go further by also considering how pupil dilation can enrich our ability to evaluate relative working memory differences required by visualizations and tasks.

3.7 Pupil Dilation Response

The pupil dilation response refers to a stimulus-reactive response of the pupil to dilate when a person puts effort toward a goal-directed behavior. The pupil dilation response makes this physiological measure useful for indicating the inherent effort of certain activities, such as making decisions with visualizations. In an examination of H4-H6, the viewers' baseline pupil diameter was compared to the pupil diameter during each trial. For H4, we examined the change in pupil dilation between easy and hard trials. To test H5, we compared the change in pupil diameter for participants in the single and dual-task groups, and for H6, we examined the influence of the interaction between *experimental group * task difficulty* on pupil dilation. Prior to analysis, the eye tracking data was cleaned using a trackloss procedure [24], in which changes of greater than .5 units were removed, indicating blinks and temporary dropped calibration of the fovea [53]. Minimum baseline pupil dilation was collected at the beginning of the study for 40 seconds while the participants viewed the start screen of the study. The baseline pupil dilation was calculated for each participant and then subtracted from the participant's pupil dilation during the study to compute the pupil dilation response.

A similar multilevel model in R was used as in the prior analyses; however, the average lightness of each image, the average lightness of the regions in a given image, and the average relative difference in lightness between the regions per image were added as fixed covariates to help control for the influence of luminance on pupil dilation. One participant from the single-task group was removed from this analysis due to the eyetracking software's failure to record. The results in the following paragraphs account for the variance in pupil dilation over and beyond the significant effects of the perceived luminance variation across images. The full output of the model, including effects of lightness, can be found in the supplemental materials.

The results of the pupil dilation analysis revealed a main effect of task difficulty such that individuals' pupil diameters were significantly larger during harder tasks compared to easier tasks (see Figure 4 C; β = 0.009, *SE* = 0.001, $p < .00$, *95% CI* [0.007, 0.01]), which provides support for **H4**. Together with the response speed data, converging evidence indicates that more working memory is required to complete the hard spatial aggregation task compared to the easy task. For **H5**, there was also a main effect of experimental group ($\beta = 0.15$, *SE* = 0.06, $p < .00$, *95% CI* [0.05, 0.26]), suggesting that individuals in the dual-task group had significantly larger pupils compared to those in the single-task group. Further, there was a significant interaction between *experimental group * task difficulty*, which provides evidence for **H6** ($\beta = 0.036$, *SE* = 0.001, $p < .00$, *95% CI* [0.032, 0.039]). This interaction can be observed in Figure 4 C.

To break down this interaction, we conducted post hoc linear regression analyses on the dual-task and single-task groups separately, where task difficulty was used to predict pupil dilation. Then we compared the slopes of the two models, which revealed that individuals in the dual-task group had a larger increase in pupil diameter from easy to hard trials ($\beta = 0.04$, *SE* = 0.002, $p < .001$), compared to those in the single-task group ($\beta = .003$, *SE* = .001, $p = .005$)($t = 17.06$, $p < .001$), which is evident in Figure 4 C. This finding illustrates the classic relationship between dual-task cost and task difficulty, in which tasks that are more difficult have greater dual-task cost. The pupillometry results provide additional converging evidence that individuals who completed both tasks were under greater working memory load than those who completed one task and that dual-taskers' load increased during difficult trials.

Key findings from the analysis of pupil dilation:

- Pupillometry revealed the classic dual-task cost, where individuals with increased working memory load from a demanding secondary task show larger changes in pupil dilation from easy to hard tasks compared to individuals under less working memory load.
- These findings illustrate how pupillometry can be more sensitive to changes in working memory load compared to measures of speed and accuracy.

In line with the large body of work that finds that pupil dilation is a consistent measure of mental effort [90], this analysis suggests that pupillometry can be used as an objective measure of working memory and thereby as a relatively stable evaluation metric for visualizations.

4 Discussion of Results

The selection of an appropriate evaluation technique requires careful consideration in visualization research. A vibrant body of research proposes many user experience goals (e.g., memorability, engagement, and enjoyment; [75]), which we agree can be essential metrics of visualization quality. Here we suggest that in order to achieve user experience goals, minimum usability criteria need to be met. Accurately measuring the relative capacity of visualizations to clearly and effectively visually communicate data in itself is challenging. One approach that we detail here involves the examination of working memory during visualization tasks as a measure of visualization effort. In a case study, we illustrated the use of both a dual-task paradigm and pupillometry to test the working memory load associated with easy and hard geospatial aggregation tasks. The results of the case study reveal that when completing more difficult visualization tasks, participants have longer task completion times than for easier tasks. Additionally, response times are sensitive enough to pick up on the influence of dual-tasking; individuals who completed a secondary working memory demanding task had significantly longer response times than those who completed only one task. Although measures of speed and accuracy can reveal some differences associated with working memory, they were not sensitive enough to expose the full impact of working memory load in this experiment. Our findings illustrate that pupillometry was able to show the lower dual-task costs associated with easier visualization tasks compared to the higher dual-task costs of more difficult tasks. We find that pupillometry is highly sensitive to differences in working memory, making pupillometry a viable physiological evaluation option for visualization practitioners.

Although this discussion of visualization evaluation focuses on measuring working memory fluctuations within task, another approach is to measure users' working memory capacity [49, 67, 78, 92], known as *individual differences measures*. For example, Zhu and Watts [98] found that individuals with low working memory capacity had difficulty using certain types of network diagrams. This approach is noteworthy because the findings reflect real differences in users' abilities [27] that should be considered in visualization research. Finding that one type of visualization is easier for people with low working memory capacity to use would suggest that the visualization requires less working memory. An individual differences approach represents a perspective to research that should be highlighted as it focuses on removing barriers for people with different abilities and in doing so improves the visualization experience for all users [49, 67]. An individual differences approach also has drawbacks, such as requiring large numbers of participants to examine the difference between groups. Further, it can be hard to find a group of participants with a sufficiently wide range of differences in working memory capacity or other individual differences. Finally, dual-tasking and individual differences paradigms are not mutually exclusive.

In addition to the method of evaluation, researchers should consider the ability of the visualization task to reveal differences in working memory. In the current case study, we employed a simple choice task that produced a binary (correct or incorrect) measure. Binary measures might not be sensitive enough to pick up on important differences in working memory between visualizations. An alternative approach is to have participants perform a task that produces a continuous outcome measure. For example, in the prior work of Padilla et al. [63], participants in one task reported the average elevation of a region rather than comparing the average elevation in multiple regions. Considering construct validity (the capacity of a metric to measure what it claims to measure [63]) can help to formalize evaluation of the pros and cons of various outcome metrics.

Beyond the considerations for the primary task, researchers interested in dual-task paradigms should consider the influence of different types of secondary tasks on the primary visualization task. Here we used one of the most conservative types of working memory demanding secondary tasks. However, researchers who have examined simple visual-spatial working memory tasks find large dual-task costs when presenting viewers with two visual tasks [29, 30]. Researchers might find a greater dual-task costs with a secondary task that requires spatial working memory, but that hypothesis remains untested for a complex visualization task. This approach may be more sensitive to subtle differences between visualizations, but may also cause an overload, resulting in failure to complete either task or ignoring one.

A variety of limitations exist in the presented case study that we would suggest considering in future instantiations of dual-task experiments and pupillometry. Researchers might want to consider using a fully within-subjects design rather than the mixed within- and between-design used here. We used a mixed design because we wanted to illustrate a comparison between groups, which are commonly used in visualizations studies that compare visualization techniques. Comparing dual-task cost between groups is not as straightforward as comparing dual-task costs within a subject. Further, we thought that there would be a learning effect, where users would be faster and more accurate in the second block. Although this might be the case, it is likely that more variability in responses would be observed by comparing judgments from participants in different groups than the variability produced from learning. Further, learning effects are systematic and easier to account for in analysis procedures than non-systematic variability across individuals.

Additionally, we suggest that researchers consider controlling for the luminance differences within the stimuli [9]. Our approach was to normalize the lightness within the DEMs, but we did not control for the average luminance in an image. In the pupillometry analysis, we attempted to account for differences in luminance by using lightness measures as covariates in our model and found that the lightness variation accounted for a significant proportion of variance in pupil dilation. We also controlled for the average lightness in the regions and the average difference in lightness in the regions for each image. Since lightness represents our perception of luminance value rather than the actual physical luminance value, it may not be the optimal choice for controlling for luminance effects on physiological responses, but it may be a sufficient proxy in certain situations (e.g., when looking at average values). An alternative approach to deal with luminance issues would be to use the model propose by Bastian et al. [9], which accounted for 70% of luminance differences when tested in six luminance conditions.

5 Conclusions and Contributions

In this paper, we provided a critical discussion of working memory evaluation techniques and described essential concepts in working memory theory. We detailed empirically validated tests of working memory and the selection of an appropriate working memory demanding task. To illustrate the utility of these approaches, we provided a case study using several converging methods for measuring working memory in a visualization task. We propose that researchers interested in the relative differences in working memory between visualizations should consider a converging methods approach, where multiple tests of working memory are employed to generate a rich evaluation of visualization quality.

References

[1] D. Alnæs, M. H. Sneve, T. Espeseth, T. Endestad, S. H. P. van de Pavert, and B. Laeng. Pupil size signals mental effort deployed during multiple

object tracking and predicts brain activity in the dorsal attention network and the locus coeruleus. *Journal of vision*, 14(4):1–1, 2014.

[2] R. Amar, J. Eagan, and J. Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization, 2005. INFOVIS 2005.*, pages 111–117. IEEE, 2005.

[3] E. W. Anderson. Evaluating scientific visualization using cognitive measures. *ACM BELIV*, 12, 2012.

[4] E. W. Anderson, K. C. Potter, L. E. Matzen, J. F. Shepherd, G. A. Preston, and C. T. Silva. A user study of visualization effectiveness using eeg and cognitive load. In *Computer graphics forum*, volume 30, pages 791–800. Wiley Online Library, 2011.

[5] S. Attfield, G. Kazai, M. Lalmas, and B. Piwowarski. Towards a science of user engagement (position paper). In *WSDM workshop on user modelling for Web applications*, pages 9–12, 2011.

[6] A. D. Baddeley. Working memory oxford. *England: Oxford Uni*, 1986.

[7] A. D. Baddeley and G. Hitch. Working memory. *Psychology of learning and motivation*, 8:47–89, 1974.

[8] A. Bandlow, L. E. Matzen, K. S. Cole, C. C. Dornburg, C. J. Geiseler, J. A. Greenfield, L. A. McNamara, and S. M. Stevens-Adams. Evaluating information visualizations with working memory metrics. *HCI International 2011–Posters' Extended Abstracts*, pages 265–269, 2011.

[9] P. Bastian, D. Fekety, A. Schmidt, and A. Kun. A model relating pupil diameter to mental workload and lighting conditions. In *In Proceedings of the 2016 CHI conference on human factors in computing systems)*, pages 5776–5788. ACM, 2016.

[10] D. Bates, M. Mächler, B. Bolker, and S. Walker. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48, 2015.

[11] R. Borgo, A. Abdul-Rahman, F. Mohamed, P. W. Grant, I. Reppa, L. Floridi, and M. Chen. An empirical study on using visual embellishments in visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2759–2768, 2012.

[12] M. M. Botvinick, T. S. Braver, D. M. Barch, C. S. Carter, and J. D. Cohen. Conflict monitoring and cognitive control. *Psychological review*, 108(3):624, 2001.

[13] M. Brehmer and T. Munzner. A multi-level typology of abstract visualization tasks. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2376–2385, 2013.

[14] S. Castro. How handheld mobile device size and hand location may affect divided attention. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 61, pages 1370–1374. SAGE Publications Sage CA: Los Angeles, CA, 2017.

[15] S. Castro, J. Cooper, and D. Strayer. Validating two assessment strategies for visual and cognitive load in a simulated driving task. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 60, pages 1899–1903. SAGE Publications Sage CA: Los Angeles, CA, 2016.

[16] S. Castro, D. Strayer, D. Matzke, and A. Heathcote. Cognitive workload measurement and modeling under divided attention. *Journal of Experimental Psychology: General*, 2018.

[17] M. Chekaf, N. Cowan, and F. Mathy. Chunk formation in immediate memory and how it relates to data compression. *Cognition*, 155:96–107, 2016.

[18] A. R. Conway, M. J. Kane, M. F. Bunting, D. Z. Hambrick, O. Wilhelm, and R. W. Engle. Working memory span tasks: A methodological review and user's guide. *Psychonomic bulletin & review*, 12(5):769–786, 2005.

[19] N. Cowan. Evolving conceptions of memory storage, selective attention, and their mutual constraints within the human information-processing system. *Psychological bulletin*, 104(2):163, 1988.

[20] N. Cowan. The many faces of working memory and short-term storage. *Psychonomic bulletin & review*, 24(4):1158–1170, 2017.

[21] N. Cowan, E. M. Elliott, J. S. Saults, C. C. Morey, S. Mattox, A. Hismjatullina, and A. R. Conway. On the capacity of attention: Its estimation and its role in working memory and cognitive aptitudes. *Cognitive psychology*, 51(1):42–100, 2005.

[22] N. Cowan, J. S. Saults, and L. D. Nugent. The role of absolute and relative amounts of time in forgetting within immediate memory: The case of tone-pitch comparisons. *Psychonomic Bulletin & Review*, 4(3):393–397, 1997.

[23] M. Daneman and P. A. Carpenter. Individual differences in working memory and reading. *Journal of verbal learning and verbal behavior*, 19(4):450–466, 1980.

[24] J. Dink and B. Ferguson. *eyetrackingR*, 2018. R package version 0.1.8.

[25] J.-L. Doumont. Magical numbers: The seven-plus-or-minus-two myth. *IEEE Transactions on Professional Communication*, 45(2):123–127, 2002.

[26] T. Drew, S. E. Boettcher, and J. M. Wolfe. Searching while loaded: Visual working memory does not interfere with hybrid search efficiency but hybrid search uses working memory capacity. *Psychonomic bulletin & review*, 23(1):201–212, 2016.

[27] R. W. Engle, M. J. Kane, and S. W. Tuholski. Individual differences in working memory capacity and what they tell us about controlled attention, general fluid intelligence, and functions of the prefrontal cortex. 1999.

[28] I. EyeTracking. Eyeworks. `http://www.eyetracking.com`, year = 2011, location = Solana Beach, CA.

[29] D. Fougnie and R. Marois. Distinct capacity limits for attention and working memory: Evidence from attentive tracking and visual working memory paradigms. *Psychological Science*, 17(6):526–534, 2006.

[30] D. Fougnie and R. Marois. Dual-task interference in visual working memory: A limitation in storage capacity but not in encoding or retrieval. *Attention, Perception, & Psychophysics*, 71(8):1831–1841, 2009.

[31] S. Haroz, R. Kosara, and S. L. Franconeri. Isotype visualization: Working memory, performance, and engagement with pictographs. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 1191–1200. ACM, 2015.

[32] P. Herath, T. Klingberg, J. Young, K. Amunts, and P. Roland. Neural correlates of dual task interference can be dissociated from those of divided attention: an fmri study. *Cerebral cortex*, 11(9):796–805, 2001.

[33] W. Huang, P. Eades, and S.-H. Hong. Measuring effectiveness of graph visualizations: A cognitive load perspective. *Information Visualization*, 8(3):139–152, 2009.

[34] J. Hullman, E. Adar, and P. Shah. Benefitting infovis with visual difficulties. *Visualization and Computer Graphics, IEEE Transactions on*, 17(12):2213–2222, 2011.

[35] M. P. Janisse. *Pupillometry: The psychology of the pupillary response*. Halsted Press, 1977.

[36] Z. Johannes, P. Ulrike, and H. Reiterer. Measuring cognitive load using eye tracking technology in visual computing. *ACM BELIV*, 14, 2016.

[37] S. Joshi, Y. Li, R. M. Kalwani, and J. I. Gold. Relationships between pupil diameter and neuronal activity in the locus coeruleus, colliculi, and cingulate cortex. *Neuron*, 89(1):221–234, 2016.

[38] M. A. Just, P. A. Carpenter, and A. Miyake. Neuroindices of cognitive workload: Neuroimaging, pupillometric and event-related potential studies of brain work. *Theoretical Issues in Ergonomics Science*, 4(1-2):56–88, 2003.

[39] D. Kahneman. *Attention and effort*, volume 1063. Citeseer, 1973.

[40] D. Kahneman and J. Beatty. Pupil diameter and load on memory. *Science*, 154(3756):1583–1585, 1966.

[41] D. Kahnemann and J. Beatty. Pupillary responses in a pitch-discrimination task. *Perception & Psychophysics*, 2(3):101–105, 1967.

[42] M. J. Kane, M. K. Bleckley, A. R. Conway, and R. W. Engle. A controlled-attention view of working-memory capacity. *Journal of experimental psychology: General*, 130(2):169, 2001.

[43] I. Koch, E. Poljac, H. Müller, and A. Kiesel. Cognitive structure, flexibility, and plasticity in human multitasking—an integrative review of dual-task and task-switching research. *Psychological bulletin*, 144(6):557, 2018.

[44] P. Kyllonen and J. Zu. Use of response time for measuring cognitive ability. *Journal of Intelligence*, 4(4):14, 2016.

[45] J. E. Laird. Extending the soar cognitive architecture. *Frontiers in Artificial Intelligence and Applications*, 171:224, 2008.

[46] Z. Liu and J. Stasko. Mental Models, Visual Reasoning and Interaction in Information Visualization: A Top-down Perspective. *IEEE Transactions on Visualization and Computer Graphics*, 16(6):999–1008, Nov. 2010.

[47] S. Lo and S. Andrews. To transform or not to transform: Using generalized linear mixed models to analyse reaction time data. *Frontiers in Psychology*, 6:1171, 2015.

[48] M. Lohani, B. R. Payne, and D. L. Strayer. A review of psychophysiological measures to assess cognitive states in real-world driving. *Frontiers in Human Neuroscience*, 13:57, 2019.

[49] G. L. Lohse. The role of working memory on graphical information processing. *Behaviour & Information Technology*, 16(6):297–308, 1997.

[50] R. D. Luce et al. *Response times: Their role in inferring elementary mental organization*. Number 8. Oxford University Press on Demand, 1986.

[51] M. Mahy, L. Van Eycken, and A. Oosterlinck. Evaluation of uniform

color spaces developed after the adoption of cielab and cieluv. *Color Research & Application Application*, 19(2):105–121, 1994.

[52] S. Mathôt. Pupillometry: Psychology, physiology, and function. *Journal of Cognition*, 1(1), 2018.

[53] S. Mathôt, J. Fabius, E. Van Heusden, and S. Van der Stigchel. Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior research methods*, 50(1):94–106, 2018.

[54] R. D. McKendrick and E. Cherry. A deeper look at the nasa tlx and where it falls short. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 62, pages 44–48. SAGE Publications Sage CA: Los Angeles, CA, 2018.

[55] C. Mélan and N. Cascino. A multidisciplinary approach of workload assessment in real-job situations: investigation in the field of aerospace activities. *Frontiers in psychology*, 5:964, 2014.

[56] G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 63(2):81, 1956.

[57] G. A. Miller. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological review*, 101(2):343, 1994.

[58] A. F. Monk, D. Jackson, D. Nielsen, E. Jefferies, and P. Olivier. N-backer: An auditory n-back task with automatic scoring of spoken responses. *Behavior research methods*, 43(3):888, 2011.

[59] P. R. Murphy, R. G. O'connell, M. O'sullivan, I. H. Robertson, and J. H. Balsters. Pupil diameter covaries with bold activity in human locus coeruleus. *Human brain mapping*, 35(8):4140–4154, 2014.

[60] F. Paas and J. Sweller. Implications of cognitive load theory for multimedia learning. *The Cambridge handbook of multimedia learning*, 27:27–42, 2014.

[61] L. Padilla. A case for cognitive models in visualization research. In *In Proceedings of the Seventh Workshop on Beyond Time and Err33rd Annual ACM Conference on Human Factors in Novel Evaluation Methods for Visualization Computing Systems*, pages 143–1511469–1478. ACM, 20185.

[62] L. Padilla, S. H. Creem-Regehr, M. Hegarty, and J. K. Stefanucci. Decision making with visualizations: A cognitive framework across disciplines. *Cognitive research: principles and implications*, 3(1):29, 2018.

[63] L. Padilla, P. S. Quinan, M. Meyer, and S. H. Creem-Regehr. Evaluating the impact of binning 2d scalar fields. *IEEE transactions on visualization and computer graphics*, 23(1):431–440, 2017.

[64] A. V. Pandey, K. Rall, M. L. Satterthwaite, O. Nov, and E. Bertini. How deceptive are deceptive visualizations?: An empirical analysis of common distortion techniques. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, pages 1469–1478. ACM, 2015.

[65] H. Pashler. Dual-task interference in simple tasks: data and theory. *Psychological bulletin*, 116(2):220, 1994.

[66] R. E. Patterson, L. M. Blaha, G. G. Grinstein, K. K. Liggett, D. E. Kaveney, K. C. Sheldon, P. R. Havig, and J. A. Moore. A human cognition framework for information visualization. *Computers & Graphics*, 42:42–58, 2014.

[67] E. M. Peck, B. F. Yuksel, L. Harrison, A. Ottley, and R. Chang. Icd3: Towards a 3-dimensional model of individual cognitive differences. *ACM BELIV*, 12, 2012.

[68] E. M. M. Peck, B. F. Yuksel, A. Ottley, R. J. Jacob, and R. Chang. Using fnirs brain sensing to evaluate information visualization interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 473–482. ACM, 2013.

[69] G. L. Pellecchia. Dual-task training reduces impact of cognitive task on postural sway. *Journal of motor behavior*, 37(3):239–246, 2005.

[70] S. Pinker. A theory of graph comprehension. *Artificial intelligence and the future of testing*, pages 73–126, 1990.

[71] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018.

[72] R. Ratcliff. Methods for dealing with reaction time outliers. *Psychological bulletin*, 114(3):510, 1993.

[73] L. M. Reder, X. L. Liu, A. Keinath, and V. Popov. Building knowledge requires bricks, not sand: The critical role of familiar constituents in learning. *Psychonomic bulletin & review*, 23(1):271–277, 2016.

[74] L. J. Rips, E. J. Shoben, and E. E. Smith. Semantic distance and the verification of semantic relations. *Journal of verbal learning and verbal behavior*, 12(1):1–20, 1973.

[75] B. Saket, A. Endert, and J. Stasko. Beyond usability and performance: A review of user experience-focused evaluations in visualization. In *Proceedings of the Sixth Workshop on Beyond Time and Errors on Novel Evaluation Methods for Visualization*, pages 133–142. ACM, 2016.

[76] B. Saket, C. Scheidegger, and S. Kobourov. Comparing node-link and node-link-group visualizations from an enjoyment perspective. In *Computer Graphics Forum*, volume 35, pages 41–50. Wiley Online Library, 2016.

[77] M. Schonlau and E. Peters. Comprehension of graphs and tables depend on the task: empirical evidence from two web-based studies. *Statistics, Politics, and Policy*, 3(2), 2012.

[78] P. Shah and A. Miyake. The separability of working memory resources for spatial thinking and language processing: An individual differences approach. *Journal of experimental psychology: General*, 125(1):4, 1996.

[79] A. Shenhav, M. M. Botvinick, and J. D. Cohen. The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, 79(2):217–240, 2013.

[80] A. Shenhav, S. Musslick, F. Lieder, W. Kool, T. L. Griffiths, J. D. Cohen, and M. M. Botvinick. Toward a rational and mechanistic account of mental effort. *Annual review of neuroscience*, 40:99–124, 2017.

[81] Z. Shipstead, T. L. Harrison, and R. W. Engle. Working memory capacity and the scope and control of attention. *Attention, Perception, & Psychophysics*, 77(6):1863–1880, 2015.

[82] S. Sirois and J. Brisson. Pupillometry. *Wiley Interdisciplinary Reviews: Cognitive Science*, 5(6):679–692, 2014.

[83] J. B. Smelcer and E. Carmel. The effectiveness of different representations for managerial problem solving: comparing tables and maps. *Decision Sciences*, 28(2):391–420, 1997.

[84] L. D. Smith, L. A. Best, D. A. Stubbs, A. B. Archibald, and R. Roberson-Nay. Constructing knowledge: The role of graphs and tables in hard and soft psychology. *American Psychologist*, 57(10):749, 2002.

[85] D. L. Strayer, J. M. Cooper, J. Turrill, J. Coleman, N. Medeiros-Ward, and F. Biondi. Measuring cognitive distraction in the automobile. 2013.

[86] D. A. Szafir. The Good, the Bad, and the Biased: Five Ways Visualizations Can Mislead (and How to Fix Them). *Interactions*, 25(4):26–33, June 2018.

[87] K. Thomas, M. Hassib, P. Wozniak, D. Buschek, and A. Florian. Your eyes tell: Leveraging smooth pursuit for assessing cognitive workload. In *In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 436. ACM, 2018.

[88] N. Tintarev and J. Masthoff. Effects of individual differences in working memory on plan presentational choices. *Frontiers in psychology*, 7:1793, 2016.

[89] U.S. Geological Survey. USGS NED 1/3 arc-second n39w079 1x1 degree IMG 2017. Reston, VA: U.S. Geological Survey, 2017. Accessed on: Jan 19, 2018. [Online]. Available: `https://catalog.data.gov/dataset/usgs-ned-1-3-arc-second-n39w079-1-x-1-degree-img-2015c267c`.

[90] P. van der Wel and H. van Steenbergen. Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic bulletin & review*, pages 1–11, 2018.

[91] A. Vandierendonck. A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior research methods*, 49(2):653–673, 2017.

[92] A. Vandierendonck, E. Kemps, M. C. Fastame, and A. Szmalec. Working memory components of the corsi blocks task. *British journal of psychology*, 95(1):57–79, 2004.

[93] I. Vessey. Cognitive fit: A theory-based analysis of the graphs versus tables literature. *Decision Sciences*, 22(2):219–240, 1991.

[94] R. Whelan. Effective analysis of reaction time data. *The Psychological Record*, 58(3):475–482, 2008.

[95] C. D. Wickens. Multiple resources and mental workload. *Human factors*, 50(3):449–455, 2008.

[96] O. Yakobi. Determinants of association and dissociation between subjective and objective measures of workload. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 62, pages 222–226. SAGE Publications Sage CA: Los Angeles, CA, 2018.

[97] J. Yucheng, B. Cardoso, and K. Verbert. How do different levels of user control affect cognitive load and acceptance of recommendations. In *In Proceedings of the 4th Joint Workshop on Interfaces and Human Decision Making (RecSys 2017)*, pages 35–42. CEUR-WS, 2017.

[98] B. Zhu and S. A. Watts. Visualization of network concepts: The impact of working memory capacity differences. *Information Systems Research*, 21(2):327–344, 2010.