Cognitive Control Beyond Single-Item Tasks:
Insights from Pupillometry, Gaze, and Behavioral Measures

Laoura Ziaka and Athanassios Protopapas

Department of Special Needs Education, University of Oslo, Norway

In Press, *Journal of Experimental Psychology: Human Perception and Performance*

Author note

Laoura Ziaka https://orcid.org/0000-0002-0559-7331

Athanassios Protopapas https://orcid.org/0000-0002-7285-8845

Experimental materials as well as data and analysis scripts are available at https://osf.io/fxgkd/

We thank SR Research support specialist William Brixius for his assistance in developing eye-tracking experiments and analyzing the corresponding data.

Correspondence should be addressed to Laoura Ziaka, Department of Special Needs Education, University of Oslo, P.O. box 1140 Blindern, 0318 Oslo, Norway. E-mail: laoura.ziaka@isp.uio.no

Word count: 13771

Abstract

Cognitive control has been typically examined using single-item tasks. This has implications for the generalizability of theories of control implementation. Previous studies have revealed that different control demands are posed by tasks depending on whether they present stimuli individually (i.e., single-item) or simultaneously in array format (i.e., multi-item). In the present study we tracked within-task performance in single-item and multi-item Stroop tasks using simultaneous pupillometry, gaze, and behavioral response measures, aiming to explore the implications of format differences for cognitive control. The results indicated within-task performance decline in the multi-item version of the Stroop task, accompanied by pupil constriction and dwell time increase, in both the incongruent and the neutral condition. In contrast, no performance decline or dwell time increase was observed in the course of the single-item version of the task. We interpret these findings in terms of capacity constraints on cognitive control, with implications for cognitive control research, and highlight the need for better understanding the cognitive demands of multi-item tasks.

Public Significance Statement— Although cognitive control is an essential part of our everyday functioning with control failures leading to maladaptive behaviors, laboratory tasks used for the investigation of cognitive control are often limited to single-item presentation posing presumably minimal control requirements on the individuals. Having as starting point the observation that real-life situations are more complex and often demand the parallel execution of different tasks (i.e., multi-tasking), we evaluated and compared within-task performance in otherwise similar tasks differing only on stimuli presentation (one-by-one or simultaneously), while at the same time we took advantage of pupillometry and gaze measures as additional

indexes of processing demands. Altogether, our results show that cognitive control implementation depends highly on the task, with more complex tasks leading to performance deterioration, and suggest that the use of primarily simple, laboratory tasks might restrict our understanding of cognitive control implementation in our everyday life.

Keywords: cognitive control; conflict monitoring; eye-tracking; multi-tasking; pupillometry; Stroop task

Cognitive Control Beyond Single-Item Tasks: Insights from Pupillometry, Gaze, and Behavioral Measures

Cognitive control has been defined in terms of "optimal parameterization of task processing" (Botvinick & Cohen, 2014, p. 1249). In more general terms, it refers to the ability to identify conflicting tasks and adapt to them (Botvinick et al., 2004; Shenhav et al., 2013). Refraining from an automatic behavior or highly habitual response is considered an act of control (Botvinick et al., 2004; Muraven & Baumeister, 2000) requiring cognitive control implementation and allocation. Theories of cognitive control aim to elucidate how the cognitive system determines how much control is needed to accomplish a task.

Conflict Monitoring and Expected Value of Control

To address allocation and implementation of cognitive control, Botvinick et al. (2001) proposed the conflict monitoring hypothesis (CMH). They assumed that control is required in conflicting situations (e.g., the incongruent condition of the Stroop task) and posited a conflict monitoring system that is responsible for detecting occasions of conflict and effecting on-line adjustments. More recently, Shenhav and colleagues (2013) proposed the Expected Value of Control (EVC) theory, arguing that during conflicting tasks a cost-benefit analysis optimizes control allocation by increasing control while diminishing the costs of its implementation. This account is an extension of the CMH aiming to approach control evaluation and allocation in greater detail.

Both CMH and EVC have been primarily tested with the use of single-item interference tasks (e.g., Stroop task, Eriksen flanker task, Simon task; for a detailed discussion see Ziaka & Protopapas, 2022). That is, tasks in which each stimulus appears individually on the screen, usually for a fixed period of time or until the participant's response, with a temporal gap between trials, namely the interstimulus interval (MacLeod, 2005). Indeed, in their introductory article in 2013, Shenhav et al. (2013) stated that their focus is on single-control demanding tasks as the "simplest and most common circumstance" (p. 220).

On this background, the present work is motivated by evidence suggesting that adopting only one paradigm for examining control issues does not warrant generalization of conclusions to other paradigms, highlighting task specificity (Rey-Mermet et al., 2018). In particular, we intend to explore how cognitive control implementation can be applied beyond single-item tasks.

Single-Item and Multi-Item Stroop Task

The Stroop task is an ideal candidate to examine this question, for two main reasons: First, much experimental evidence for the CMH and EVC has been based on the Stroop task (Botvinick et al., 2001; Shenhav et al., 2013; Frömer et al., 2021). Second, and in contrast to other tasks often used in the cognitive control literature[1] (e.g., Eriksen flanker task or Simon task; Nee et al., 2007), the Stroop task is not restricted by its nature to single-item presentation.

In particular, two versions of the color-word Stroop task exist: The original, or close to the original, multi-item version, in which all stimuli are presented simultaneously on a sheet of paper, and the more modern, single-item (i.e., computerized) version, in which each item appears individually on a screen. Standardized neuropsychological assessment and clinical practice makes use of the former for identifying inhibition, control, and attentional deficits (e.g., Björngrim et

[1] Cognitive control as measured by interference tasks (i.e., Stroop task, flanker tasks, and Simon task among others) is also referred to in the literature as attention control (e.g., Draheim et al., 2021), inhibition (e.g., Rey-Mermet, 2018), or executive control (e.g., Rey-Mermet, 2019). Here we will only use the term "cognitive control", for consistency, given that relevant studies often use "cognitive control" as an additional keyword.

al., 2019; Bezdicek et al., 2015; Penner et al., 2012; Periáñez et al., 2021; Rabin et al., 2005; Salo et al, 2001; Scarpina & Tagini, 2017; for a review see Strauss et al., 2006), while research on the corresponding cognitive processes uses the latter (MacLeod, 2005; Salo et al., 2001). The choice in both cases is due to convenience; namely, easy, effortless administration in clinical practice (Rabin et al., 2005), vs. access to error-free data and resistance to speed-accuracy trade-offs in research (MacLeod, 2005; but cf. Draheim et al., 2021, for an opposite view regarding speed-accuracy trade-offs in single-item tasks).

Despite a presumed near equivalence of the two versions, studies have shown that the choice of administration and presentation is not a simple, secondary matter of convenience (Salo et al, 2001; MacLeod, 2005; Ziaka & Protopapas, 2022) because it has implications for the observed Stroop effect. In particular, interference estimates are larger in the multi-item version (MacLeod, 2005; Salo et al., 2001) and, in addition, individual differences studies result in divergent patterns of shared variance (Ziaka et al., 2022), indicating that at least partially different underlying cognitive mechanisms may be involved in the two versions.

It has long been proposed that the single-item version fails to retain the initial "Stroop-like" (i.e., highly conflicting) nature of the multi-item version (Penner et al., 2012), presumably because in the multi-item task nearby items act as distractors increasing its difficulty (Ludwig et al., 2010). This idea is in line with findings from the Eriksen flanker task which show that spatially adjacent distractors belonging to the target set cause interference (Eriksen & Eriksen, 1974), something that should also hold for the multi-item Stroop task, in which all simultaneously presented items are targets to be named. Moreover, findings from single-item tasks have showed that temporal separation of items modulates performance (Esterman et al., 2013; Glaser &

Glaser, 1982, 1989; for a review see Fortenbaugh, et al., 2017). These findings imply a cost for multi-item tasks, in which temporal separation is not possible. Unavoidable temporal and spatial adjacency can thus jointly contribute to the highly conflicting nature of multi-item tasks.

If that is indeed the case, then two different processes must be involved in the multi-item Stroop task. Specifically, based on Friedman and Miyake (2004) taxonomy, the (single-item) Stroop task requires the inhibition of an automatic response named “prepotent response inhibition”, while flanker tasks involve “resistance to distractors interference” (see Draheim, 2021; Rey-Mermet et al., 2019, for supporting evidence). Bearing in mind that the multi-item Stroop task involves both, that is, an automatic task-irrelevant response as well as flankers (i.e., the adjacent items), we suggest that its successful execution entails two different inhibition processes, concurrently active.

Furthermore, the need to generate eye movements during the multi-item version of the task (Salo et al., 2001) makes the visual input more complex and dynamic (Snell et al., 2018). Eye movement studies of parafoveal processing in other multi-item naming tasks (Henry et al., 2018; Pan et al., 2013; Kuperman et al., 2016) suggest that in the single-item version the attentional and perceptual field is restricted to the currently presented item, whereas for the multi-item version it is expanded in order to process the upcoming target items. This attentional expansion introduces a potential for interference between successive items in the multi-item version, and hence for additional cognitive control demands, which are not present in the single-item version. The magnitude of between-item interference may depend on the condition, that is, incongruent or neutral.

More specifically, eye movement research suggests that in the incongruent condition participants are parafoveally exposed to the upcoming item and need to identify its automatic dimension (i.e., word) as irrelevant. At the same time, processing of the current item requires simultaneous processing of two dimensions, namely, the color dimension—integrated, controlled, and slower— and the word dimension—task-irrelevant but automatic, which must be suppressed. This ultimately leads to at least three simultaneous active responses, two to be identified and filtered out, and one to be articulated. And all this while uttering the response to the preceding item (for eye-voice span in other multi-item naming tasks see Gordon & Hoedemaker, 2016; Huang, 2018; Pan et al., 2013; Silva et al., 2016). In sum, it seems that reading and naming, identification of the upcoming responses, and response selection, planning, and articulation run in parallel in the multi-item Stroop task. Much the same should hold for the neutral condition with the exception of the (automatic) word dimension.

If in fact concurrent tasks and subtasks run in parallel in both conditions of the multi-item Stroop task, this would mean that time sharing among them is required, potentially resulting in additional task interference (Wickens, 2002) and crosstalk (Fischer & Plessow, 2015). Seen in this perspective, the multi-item Stroop task does not consist in a single control-demanding task, as commonly understood. Instead, it can be conceptualized as a “multi-task”, meaning that—although it is presented as a single task—its successful execution entails simultaneous and parallel activation of more than one task sets, thereby satisfying the conditions for multi-tasking (i.e., execution of concurrent tasks demanding the simultaneous maintenance of two or more task sets; Koch et al., 2018; Meyer & Kieras, 1997; Monsell, 1996; for a similar view related to pupillary responses see Mathôt, 2018). Consequently, performance costs and control deterioration could be expected, as typically observed in the multi-tasking literature (Fischer & Plessow, 2015).

Ziaka and Protopapas (2022) adopted within-task performance evaluation as a methodology for comparing single-item and multi-item Stroop tasks, and showed performance decrements in the multi-item version only, justifying the conceptualization of the multi-item Stroop task as a multi-task. They interpreted these findings as suggestive of capacity constraints in control implementation and allocation under conditions requiring parallel execution of multiple cognitive tasks. However, certain alternative interpretations cannot be ruled out, such as for example the impact of incentive components on control allocation. To address these possibilities, within-task performance evaluation must take advantage of different methodologies.

Pupillometry and Gaze Measures

In the EVC theory control implementation is accompanied by a cost because it requires mental effort, which participants seek to minimize due to its aversive nature. Consequently, control implementation must be intense enough, to maximize rewards, but not too intense, so that the cost of control implementation does not rise excessively. If a discrepancy between costs and benefits occurs, it is detected, and control implementation is adjusted accordingly and dynamically (Shenhav et al. 2013). Therefore, an alternative to the capacity constraint interpretation for within-task performance decrements in the multi-item Stroop task might be that participants “decided” to withdraw from the task through cost-benefit analysis because cost was too high and payoffs too low. Put simply, according to this view it was no longer worth investing control in this task; thus, disengagement emerged and performance decrements ensued.

Behavioral response measures such as response times do not allow us to distinguish whether within-task performance variations in the Stroop task should be attributed to capacity constraints of the control system or to withdrawal from the task via a cost-benefit analysis. To disentangle the alternative interpretations, in the present work we turn to a combination of pupillometry, eye movements, and behavioral response measures. Pupillary responses can reveal the presence of conflict (Hershman & Henik, 2019), while eye movements can help distinguish among predictions involving the underlying cognitive processes of control implementation and allocation (Olk, 2013; Valki et al., 2019).

Pupil size

Pupillary responses have long been linked to "within-task processing requirements" (Hess & Polt, 1964, p. 1190) and "second-to-second variations in the load that the mental activity imposes" (Kahneman et al., 1967, p. 218). In the domain of cognitive control, pupil size variations have been studied in reference to updating, switching, and inhibition (for a review see van der Wel & van Steenbergen, 2018), with pupil-size changes, that is, constriction and dilation, being considered as markers of effort and processing load (Laeng et al., 2012; van der Wel & van Steenbergen, 2018).

A typical manipulation for investigating processing and cognitive load in simple demanding tasks via pupillometry involves conditions which vary in difficulty, linking pupil size variations to task demands (Tapper et al., 2021). In the context of the single-item Stroop task, pupil size changes in the incongruent condition are compared to simpler, non-conflicting conditions. As in other single-item conflicting tasks, it is a consistent finding in the single-item Stroop task that the pupil dilates as task demands increase, that is, in the more demanding incongruent condition, compared to non-conflicting conditions (e.g., Brown et al., 2014; Hasshim & Parris, 2015; Hershman et al., 2020; Laeng et al., 2011; but cf. Frömer et al., 2021, for an opposite effect). However, studies using other tasks have revealed that under conditions of disproportionate processing load pupil size reaches a plateau and even constricts (Granholm et al., 1996, 1997; Johnson et al., 2014; Peavler, 1974; Poock, 1973; Zekveld & Kramer, 2014). Thus, while pupil dilation reflects load, pupil constriction reflects overload.

It is particularly relevant for our argument that similar changes in pupil size have been observed under multi-tasking. Specifically, under dual-task conditions, the pupil initially dilates and then constricts, in parallel with behavioral results, that is, performance improvement followed by deterioration (Kahneman et al., 1967; Shiga & Ohkubo, 1978). Nowadays it is considered a stable finding that the pupil constricts in dual-task conditions (compared to a single-task baseline), while performance drops (Häuser et al., 2019; Karatekin, 2004; Recarte & Nunes, 2000; Recarte et al., 2008; Tapper et al., 2021). Two alternative interpretations have been proposed to explain these results: The first one posits that pupil constriction indicates capacity constraints due to cognitive overload (Kahneman, 1973; Tapper et al., 2021; Zekveld & Kramer, 2014; Zekveld et al., 2011). For the second interpretation, motivation plays a crucial role, leading participants to disengage from the task (which is no longer "worth it"), and as a result the pupil constricts (Granholm et al., 1996; Winn et al., 2018).

Although no consensus has been reached yet regarding the origin of pupil constriction under dual-task conditions, what is of primary importance here is that, if the multi-item Stroop task is indeed a multi-task, as we propose, then the pupil should shrink within the task, following the pattern of results observed in dual task conditions. Indeed this would be expected to occur in both conditions, that is, incongruent and neutral. In contrast, for the single-item task, in which task demands are moderate, the pupil should dilate in the incongruent condition, in accordance with previous evidence from single tasks.

Dwell Time

Because pupil constriction can be attributed either to capacity constraints or to disengagement, in the present study we additionally take into account gaze measures. Fixations, that is, time spent looking at a specific area, are thought to reflect the time needed to attentively process an event or item (Eckstein et al., 2017). Dwell time is the sum of all fixation durations within an area, reflecting both initial and later processing (Rayner, 2005). Dwell times can thus index the complexity of the task and the amount of attentional resources needed to accomplish it (Busjahn et al., 2014), and are used as an implicit measure of task-dependent interest, engagement, and decision making (Kellar et al., 2004; Kellough et al., 2008; Liu & Belkin, 2010; Liu et al., 2012; Liversedge et al., 1998; Saastamoinen & Järvelin, 2018; van der Laan, 2015; White & Kelly, 2006; Yagle et al. 2017).

Several studies have used pupillometry to study the emergence and management of conflict in the Stroop task. In comparison, only a few studies have adopted gaze measures. Olk (2013) found that dwell times were longer for the incongruent than for the congruent condition of a modified numerical Stroop task. This finding was interpreted as attention allocation on the incongruent items due to the presence of conflict and the time needed to resolve it in order to identify the correct response. Vakil et al. (2019) obtained similar results using a different version of the single-item color-word Stroop task. Both of these studies measured dwell times in single-item variants of the Stroop task, thus their relevance for the multi-item version is limited.

We are aware of only one study recording eye movements in the multi-item Stroop task. Specifically, Wu et al. (2018) used a three-card variant of the Stroop Color Word Test (SCWT; Golden, 1978; Golden & Freshwater, 2002) to compare

Chinese children with and without developmental dyslexia. They found shortest average fixation durations for word reading of color-neutral words (i.e., color words printed in black ink; Card A), intermediate for color naming of neutral stimuli (e.g., blue circle; Card B), and longest for color naming of incongruent stimuli (e.g., the word red printed in blue; Card C). Card and group did not interact. These results indicate that word reading of color words demands less attentional resources and engagement when compared to color naming of either neutral or incongruent stimuli, with the highest level of attentional demands in the incongruent condition. Wu et al. calculated average fixation duration for entire cards, leaving open the question of within-task variations in dwell time.

Altogether, the literature suggests dwell time as a measure of attentional engagement. Thus, we hypothesized that, if participants decide to disengage from the multi-item task through a cost-benefit analysis, this should be mirrored in all three dependent measures (i.e., behavioral, pupil size, and dwell time) indicating an overall withdrawal. In contrast, if performance and pupil decrease is accompanied by dwell time increase, suggesting amplification of attentional maintenance under an overload state, then the most plausible explanation would be capacity constraints on control due to the multi-tasking demands of the multi-item Stroop task.

The Present Study

The aim of the present study was to examine performance variations within the course of a Stroop task as a function of the format in which the task is presented, that is, single-item or multi-item. To achieve this, behavioral measures were combined with recordings of pupillary responses and eye movements, aiming to distinguish between different alternative interpretations.

Hypothesis 1: The multi-item Stroop task is a single control-demanding task

If the multi-item Stroop task is a single control-demanding task, as suggested by administration and presentation and, thus, equivalent in terms of processing requirements to the single-item version of the task, then control adaptations are expected during the incongruent condition because control is dynamically adjusted, taking payoffs and cost of control implementation into account. No performance changes should be observed in the course of the neutral condition, in which control demands are minimal or even absent, depending on stimulus selection and/or readability of the material (e.g., colored letter strings such as 'XXX' or semantically unrelated colored words such as 'CAT'; Augustinova et al., 2018; Kalanthroff et al. 2013; Levin & Tzelgov, 2014; see also Botvinick et al. 2001, Figure 1).

Consequently, performance and dwell time should remain stable during the task in both conditions (i.e., incongruent and neutral), while the pupil should dilate primarily in the incongruent condition. Crucially, the pattern of within-task variation should be similar between the multi-item and single-item version of the task.

Hypothesis 2: The multi-item Stroop task is a multi-task

A conceptualization of the multi-item Stroop task as a multi-task implies that multiple control demanding tasks *run in parallel*. In this case, three alternative predictions can be made for the incongruent condition.

If the increased control demands remain within manageable bounds, there are two possibilities: First, within-task variation in all measures (i.e., performance, pupil, and dwell time) during the multi-item Stroop task should resemble the corresponding variation of its single-item counterpart; that is, performance and dwell times should remain stable and pupil should dilate primarily in the incongruent condition of the task because control is gradually and adaptively adjusted.

As a second possibility within the purview of CMH/EVC, if the multi-tasking nature of the task leads cost-benefit analysis to indicate that the cost of control implementation is too high and pay-offs too low, then disengagement should follow, indicated by performance drop, pupil constriction, and dwell time *decrease*.

The third alternative prediction would hold if capacity constraints apply (Shenhav et al., 2013; Ziaka & Protopapas, 2022) that are overwhelmed by the control demands of multi-tasks. In this case cognitive overload ensues, while engagement is maintained, leading to performance deterioration, pupil constriction, and dwell time *increase*.

This latter scenario is primarily relevant for the more complex incongruent condition, but it is not limited to that. Indeed, if the so-called “neutral” condition also requires exertion of intensive control, then a pattern similar to the incongruent condition should be evident, namely, decreasing performance as the task progresses. However, this performance decrement should be modest and more gradual, due to the absence of the irrelevant word response, which raises the level of task complexity in the incongruent condition.

To ensure that within-task variations in the multi-item Stroop task are not restricted to (or even due to) specific task features, different multi-item variants are included in the present study. Two components that are highlighted in the literature are taken into account, namely between-item spacing and naming direction. Specifically, reduced between-item spacing is known to impede target recognition (e.g., Eriksen & Eriksen, 1974; Moll & Jones., 2013; Perea & Gomez, 2012; Rayner et al., 1998; Rayner et al., 2013; Sheridan et al., 2013). With respect to naming direction, parafoveal processing for linguistic material is known to occur to a larger extent in the default reading direction of the language (e.g., Snell & Grainger, 2018;

Snell, Mathôt et al., 2018; Snell et al., 2021). The perceptual span of parafoveal processing is shorter in the vertical direction, if the default reading direction is left-to-right (e.g, Ojanpää et al., 2002; Seo & Lee, 2002; Snell et al., 2018), and it can be extended toward the opposite direction for non-linguistic material (i.e., left-oriented for readers of a left-to-right orthography; e.g., Harms & Bundesen, 1983). The perceptual span is particularly important for multi-item tasks under the multi-task conceptualization because it governs the amount of interference by nearby items. For these reasons, in the present study we use multi-item Stroop tasks differing in between-item spacing (i.e., dense vs. sparse) and naming direction (top-to-bottom vs. left-to-right), aiming to test the robustness of the findings.

Method

Participants

Participants were 42 undergraduate students from Panteion University and the University of Athens (aged 19–22; *M* = 19.52, *SD* = 0.91; 37 females) who received class credit for their participation. All were Greek native speakers with normal or corrected-to-normal vision and normal color perception. Written informed consent was obtained from all participants.

Time of data collection and sample-size justification

Data collection was performed between April and May 2018, sampling young adults from the same population as Ziaka & Protopapas (2022) and aiming for a similar sample size.

With respect to response time, we conducted a post-hoc power analysis of Experiment 2 of Ziaka and Protopapas, which shared many similarities with the current study,[2] using library simr v.1.0.6 (Green & MacLeod, 2016). The power to detect a linear trend of response time by column of the multi-item task (the main effect of interest) in the incongruent condition exceeded .95 for 12 participants. The power to detect an interaction of this linear trend with condition was about .60 for 40 participants (full analyses available on OSF).

For the eye-tracking measures, that is, pupil size and dwell time, no previous studies have examined their within-task variation in comparable tasks. Our sample size is at least as large as previous studies of Stroop (and Stroop-like) tasks examining pupil size (e.g., between 22 and 27 participants in Hershman et al. 2019; 26 in Hershman et al., 2020; 33 in Hasshim & Parris, 2015; 40 in Laeng et al., 2011) and dwell time (e.g., 15 in Olk, 2013).

Materials

Experimental Trials: Incongruent and Neutral Conditions

For the incongruent condition, Greek color words were displayed in a non-matching color. We chose the words for red (κόκκινο /kocino/), green (πράσινο /prasino/), and yellow (κίτρινο /citrino/), because they have the same number of letters and syllables, comparable written frequency (33, 34, and 9 per million, respectively, from the IPLR; Protopapas et al., 2012), and begin with voiceless stops, which facilitates response time triggering. The color of the ink was either red (RGB: 255, 0, 0), green (RGB: 0, 168, 0), or yellow (RGB: 255, 255, 0) with the corresponding colors being familiar, easily distinguishable, and frequently used in behavioral and pupillometry Stroop studies (e.g., Hasshim & Parris, 2015; Hershman et al., 2020; Laeng et al., 2011). The color words for red, green and yellow appeared

[2] The multi-item tasks of Ziaka & Protopapas (2022) were split into only 3 columns of 20 items, whereas here the same total number of items (60) is split into 12 blocks of 5. We analyzed Experiment 2 instead of Experiment 1 to minimize concerns related to regression to the mean (winner's curse).

in a non-matching color.

Stimuli for the neutral condition were strings made up of seven repetitions of the letter X (no spaces; i.e., XXXXXXX) in red, green, and yellow color.

Scanning Baseline: Letter-Scanning Conditions

Scanning conditions were inserted before each experimental condition to serve as a reference baseline accounting for position artifacts, that is, apparent pupil-size changes due to position (Gagl et al., 2011; Mathôt, 2018). Scanning conditions in the multi-item tasks matched the corresponding experimental conditions in stimulus position and extent, but instead of color words (or strings of Xs) presented in color, they presented strings of letters displayed in black color. Each string was composed of seven repetitions of the same Greek letters (among the set X, Λ, and B, or X, Λ, and Π[3]). Different letters were used for different strings, rather than repetitions of the same symbol or letter for all stimuli as previously implemented (Gagl et al., 2011), aiming to facilitate keeping track of position and thereby minimize skipping and the ensuing data loss.

For the single-item Stroop task a neutral stimulus (i.e., seven repetitions of the # symbol) appeared at the center of the screen prior to each experimental trial to serve as the reference baseline. This is termed “scanning” in the analyses, for consistency in treatment, although there was no actual visual scanning in the single-item task (as all individual stimuli appeared at the center of the screen).

Tasks

Familiarization

Before each condition (i.e., neutral and incongruent) of a task, exemplar cards

[3] In the DTB condition, which replicated stimulus presentation in Ziaka and Protopapas (2022), the letters X, Λ, and B were used so as to replicate one of the neutral conditions used in that study. In the other multi-item conditions, B was replaced by Π because its higher visual complexity might make it more salient compared to X and Λ.

with the sequence of screens for the entire trial were shown to the participants (i.e., drift check, # screen or letter-scanning condition, and experimental trials) along with an explanation of the procedure and verification of expected responses in the experimental trials. There were no practice trials.

Screenshots and videos exemplifying each task are available on OSF (https://osf.io/fxgkd/).

Single-Item Stroop Task

The single-item task consisted of 60 trials. For each trial, a single stimulus (neutral or incongruent) was displayed at the center of the screen in bold 20-pt Courier New font on a gray background.

Multi-item Stroop tasks

Dense Top-to-Bottom (DTB). Each condition was presented in a single-screen array of three columns of 20 stimuli, for a total of 60 stimuli per condition, displayed in bold 20-pt Arial bold font on gray background. For the incongruent condition there were 20 repetitions of each word and 20 repetitions of each color, counterbalanced in their combinations and equally distributed over columns. Colors and color words were randomly ordered with the constraint that adjacent items within columns were not the same. The vertical spacing (edge to edge) between items was 4.5 mm (approximately 0.28 degrees of visual angle). The Arial font was used in this task in order to match the corresponding condition of Ziaka and Protopapas (2022) as closely as possible.

Sparse Top-to-Bottom (STB). The structure of this task was similar to that of the Dense top-to-bottom (DTB), differing in font and stimulus placement. Specifically, stimuli were displayed in bold 20-pt Courier bold font arranged in a matrix of 6 columns of 10 items each, for a total of 60 stimuli per condition. The vertical spacing (edge to edge) between items was 19.2 mm (approximately 1.22 degrees of visual angle). The monospaced Courier font was used in this and the following task so that between-item spacing could be fixed to allow direct comparisons between top-to-bottom and left-to-right naming directions.

Sparse Left-to-Right (SLR). This variant was identical to the Sparse top-to-bottom (STB) task with the only difference that stimuli were arranged in 6 rows of 10 items each, requiring left-to-right naming. The horizontal spacing (edge to edge) between items was 19.2 mm (approximately 1.22 degrees of visual angle), matching the vertical edge-to-edge spacing of STB.

Task Order and Administration of Conditions

Task order within a session was random. For each task, the neutral condition was administered first, followed by the incongruent condition. (Thus, presentation of conditions in the single-item task was blocked.) Maintenance of a fixed order of conditions is compatible with most of the commonly used Stroop tests in which the non-conflicting (i.e., neutral) condition is administered first (e.g., Victoria version, Golden version; Golden, 1978; Golden & Freshwater, 2002; Strauss et al., 2006).

Procedure

General Procedure

Participants were tested individually in a windowless room with artificial light. Eye movements were recorded using an EyeLink 1000 Plus eye tracker (SR Research, Toronto, ON, Canada) using a 35 mm lens. The sampling rate was 1000 Hz. Stimuli were displayed on a 27″ LED computer screen (2560 × 1440 pixels with a refresh rate of 144 Hz) at a viewing distance of approximately 90 cm. The experiment was implemented in Experiment Builder 2.2.1 (SR Research, 2016). The EB scripts implementing the tasks are available on OSF (https://osf.io/fxgkd/).

A forehead rest stabilized head position. Before each experimental session pupil and corneal reflection thresholds were adjusted. Because the Eyelink 1000 Plus eye tracker records pupil size as the number of thresholded pixels, no further adjustment of thresholds were made during the experimental session. The participants' eye position was calibrated using 13 black dots (custom target) covering the entire horizontal and vertical extent of the screen. Calibration targets were presented individually in random order and subjects were asked to fixate the center of each dot. For increased accuracy, acceptance of each target fixation was done manually by the experimenter. Calibration was immediately followed by a validation routine to determine the stability and accuracy of calibration. Validation was considered acceptable if the worst point error was less than 1.5 degree and the average error was less than 1.0 degree. The initial calibration and validation procedures were binocular in order to define the eye with the least maximum and average error. The selected eye was "locked" after that and was kept constant during the experiment. After the initial calibration and validation and before each experimental condition and task, monocular calibration and validation procedure followed the same procedure as before. Calibration accuracy was checked before each trial using a drift-check point (same as the calibration point) at the center of the screen.

Throughout the entire experimental session the background color (RGB 204, 204, 255) was kept constant during every step (i.e., instructions, calibration, validation, experimental tasks, and "thank you" screen).

Single-item Task

A trial started with a neutral stimulus (i.e., seven repetitions of # in black bold font) appearing for 1500 ms at the center of the screen, constituting the reference "scanning" baseline. This was immediately followed by the experimental stimulus appearing at the same location for 2000 ms. There was no blank screen between scanning and experimental stimulus. Stimuli were presented in random order (different randomization for each participant). Participants were asked to name the color of the ink as fast as possible and try to avoid errors.

Multi-item Tasks

For the multi-item Stroop tasks each trial started with a neutral stimulus (i.e., seven repetitions of # symbols in black bold font) appearing at the center of the screen for 1500 ms. The same string of #s appeared again at the position of the first item to be named for 1500 ms. The exact screen coordinates (horizontal, vertical) in pixels from top left, were (1050, 360) for DTB, (800, 263) for STB, and (487, 470) for SLR.

A letter-scanning trial preceded each experimental trial (i.e., each condition). In the letter-scanning trials, participants were required to simply scan the card in the order required by the current task (i.e., top-to-bottom or left-to-right) without any additional processing. When participants fixated the final fixation dot at the bottom right of the screen, the experimenter initiated presentation of the corresponding experimental trial.

In the experimental trial, the same procedure with the fixation stimuli was followed (i.e., ####### at central location followed by ####### at the location of the first stimulus). When the experimental card appeared participants were required to name the color of the ink, as instructed during the preceding familiarization phase. When all responses had been produced and participants fixated the bottom right fixation dot, the experimenter terminated the trial. The exact same procedure was followed for both conditions (neutral and incongruent) in all multi-item tasks (DTB, STB, and SLR).

Data processing

Preprocessing

Behavioral Measures.

Single-Item Task. Accuracy and naming times were processed offline with CheckFiles (Protopapas, 2007) to mark response times, that is, onset latency of the vocal response. For the measure of accuracy, mispronunciations, substitutions, and self-corrections were considered errors. Response times were inverted and multiplied by 1000 to produce a rate measure of "items per second", which is approximately normally distributed. Graphical and statistical analyses of normality for both the original and transformed times are listed in the online Supplemental Material B (pp. 7–12). In addition, all behavioral analyses reported below have also been conducted using the raw (untransformed) response times; these are also listed in the online Supplementary Material A (section 5) to facilitate comparisons with the Stroop task literature.

For each participant and condition, a mean response rate was calculated for each set of five consecutive trials[4], resulting in a sequence of 12 trial "blocks" to be compared to the five-item blocks of the multi-item versions.

Multi-Item Tasks. The recorded responses were manually processed offline using Praat (Boersma & Weenink, 2012) to determine the accuracy and total naming

[4] Response times to both correct and incorrect responses were retained for analysis, aiming to maximize comparability between the single-item and multi-item versions (Egner & Hisrh, 2005; MacLeod, 2005). That is, we took into account that the multi-item Stroop task contains unavoidable errors that cannot be individually removed and that errors may affect all dependent measures, that is, response time (e..g., Rabbitt, 1966; Kleiter & Schwarzenbacher, 1989), pupil size (e.g., Braem et al., 2015; Maier et al., 2019), and eye-movements (e.g., Inhoff et al., 2016). Because results for the multi-item tasks would necessarily contain erroneous responses, we decided to retain those for the single-item task as well.

duration for each set of five consecutive items, resulting in measures for 12 5-item blocks[5] in each condition and task. There were thus four blocks per column in the DTB task, two per column in the STB task, and two per row in the SLR task.

Naming times for each block were inverted and multiplied by five (i.e., the number of items) to produce a rate scale comparable to the single-item version (i.e., items per second).

Pupil Size and Dwell Time Measures. A separate interest area was defined for each item[6]. For each task and condition, Data Viewer 4.2.1 (SR Research, 2021) was used to generate sample reports for pupil data and interest area reports for gaze data. The interest area report contained dwell time as the sum of durations across all fixations within each interest area.

Pupil data were preprocessed with the gazeR package (Geller at al., 2020) in R 4.1.1 (R Core Team, 2021), including de-blinking (removal of data 100 ms before and after blinks), cubic-spline interpolation of missing data (due to blink removal), and smoothing with an 11-ms Hanning window (Mathôt, 2013; Mathôt et al., 2018).

[5] The division into 5-item blocks was dictated by constraints on material construction. Specifically, we aimed to replicate Ziaka and Protopapas (2022) in the DTB task, while also fitting sparser arrays of the same number of stimuli within the usable portion of the screen (avoiding edges, to minimize data loss) in the other tasks. Taking into account findings showing that disruption of temporal overlap in simultaneous processing is beneficial for performance (e.g., Fischer & Plessow, 2015) and the effect of return sweeps on eye-movements (e.g., Slattery and Parker, 2019), the division of items into blocks of five aimed to ensure comparable data between tasks, minimally affected by breaks between transition blocks and/or return sweeps.

[6] To minimize data loss, the height of the interest areas was made as large as possible depending on stimulus density in each task. For the single-item task, interest areas were defined by adding 30 pixels to the top and bottom of the items. For the DTB multi-item task, non-edge interest areas were 38 pixels high, whereas the top and bottom interest areas in each column were 97 pixels high, to account for drift associated with column transitions. Finally, interest areas were 80 pixels high for the STB task and 96 pixels for SLR.

To facilitate interpretation, raw pupil size data were converted from arbitrary units to millimeters based on recording an artificial pupil (following instructions from SR Research) and calculating the mean of 60 items.

Finally, sets of five consecutive interest areas (i.e., items) for dwell time and pupil size data were grouped into blocks by averaging, resulting in a total of 12 block-level values for each measure (dwell time and pupil size) for each condition and task. Note that, because items in the single-item task appeared centrally for a fixed amount of time and did not require eye movements, dwell times for the single-item task exhibit substantial negative skewness (Figure S.3 in Supplemental Material B). Accordingly, within-task differences in dwell time in the single-item tasks indicate fixations away from the centrally presented items. Supplemental analysis on cubic root-transformed dwell times for the single-item task showed the same pattern of results as the untransformed data (see online Supplemental Material A; section 6).

Statistical analysis

To examine the effect of condition and block on response rate and dwell time we tested the two-way interaction of condition and block using function lmer of the lme4 package v. 1.1-27 (Bates et al., 2015). Random effects for participants included intercepts and slopes for condition. In models for this as well as all other dependent variables, condition was deviation-coded using function contr.sum, while block was treatment-coded, so that the effects of block would be evaluated at the average of the two conditions, while the effects of condition would be evaluated at the first block. Interactions would then concern differences from the first block. In R notation the model formula was specified as

```
~ condition * block + (1 + condition | subject)
```

For pupil size, an additional independent variable was included to account for position artifacts, namely “scanning”, with two levels: baseline and experimental. Level “baseline” comprised letter-scanning trials preceding the experimental trials, while level “experimental” comprised the experimental neutral and incongruent trials. This variable was treatment-coded so that its effects would correspond to the difference between baseline and experimental. In this way interactions with this variable allow us to test the effects (or interactions) of other variables controlling for position artifacts. A “scanning” variable (similarly coded, based on the string of #s preceding each experimental stimulus) was also used to analyze pupil size data from the single-item task, even though no position artifacts are possible, for consistency in analysis. Random effects for participant in models for pupil size for every task included random intercepts as well as random slopes for scanning, condition, and their interaction.

For error rate, the two-way interaction between condition and block for each task was tested with Bayesian generalized mixed-effects models for binomial distributions instead of generalized mixed-effects models because the latter failed to converge due to complete separation. For these models we used function bglmer of the blme package v. 1.0-5 (Chung et al., 2013) with default priors on the fixed effects and appropriate choice of optimizers.

Because we were interested to ascertain whether within-task variations in our dependent measures followed a specific pattern, that is, whether the levels of block corresponded to increasing or decreasing values of the dependent variables (or to a U-shape pattern) and whether such polynomial effects differed across levels of condition, we performed additional trend analyses for each task by using polynomial coding for the block variable in models otherwise identical to those described above.

Finally, we directly compared tasks differing along a single dimension of interest (i.e., Dense vs. Sparse top-to-bottom; and Sparse top-to-bottom vs. left-to-right) to examine whether between-item spacing and naming direction affected the slopes of the trends by blocks.

For all analyses, simpler models were used if full random structure resulted in nonconvergence.

Transparency and openness

All materials, raw data, and analysis scripts, including graphical and statistical analysis of normality, complete model output as well as analysis-of-variance tables for linear and generalized linear mixed-effects models for the different dependent measures and tasks are available on OSF (https://osf.io/fxgkd/) and in the online Supplemental Material A and B.

Results

Two participants were excluded from all tasks due to poor calibration. Three additional participants were excluded from the DTB task only, due to software failure or poor calibration.

Results for all measures and tasks are depicted in Figure 1 for behavioral measures and Figure 2 for eye-tracking measures. Within-participant confidence intervals were calculated with function summarySE of the Rmisc package (Hope, 2013). The middle column in Figure 2 displays the difference in absolute pupil size between the experimental and corresponding scanning conditions, that is, relative pupil size adjusted for position artifacts. This is only an illustration to facilitate interpretation of the statistical analyses; the adjusted sizes were not directly modeled. Instead, the analyses employed interactions with the “scanning” variable to control for position artifacts while properly modeling variance across conditions.

Table 1 presents analysis-of-variance tables for linear and generalized linear mixed-effects models for each dependent measure and task. A significant main effect of scanning was evident in every task, consistent with pupil dilation in the experimental conditions relative to baseline, as expected. This effect appears considerably larger in the multi-item tasks than in the single-item task (see Figure 2). Tables 2 and 3 present the results for the corresponding polynomial effects for behavioral and eye-tracking measures, respectively.

Single-item task

As shown in Table 1, for response rate and errors only the main effect of condition reached significance, with the incongruent condition being overall slower and more error prone than the neutral condition. This is Stroop interference, as expected. There was no difference by block and no significant linear or quadratic term for the two behavioral measures (Table 2), suggesting that performance as reflected in response rate and errors was overall stable within this task.

A different pattern of results emerged for pupil size (see Table 1). Specifically, the significant interaction of condition with scanning indicates a Stroop effect (i.e., dilated pupil for the incongruent relative to the neutral condition; see Figure 2, middle panel in top row). The significant interaction of block with scanning indicates that pupil size (controlling for position artifacts) did not remain the same as in the first block. The effect of block was also evident in the subsequent trend analysis (Table 3), which showed a significant—although weak—linear trend, consistent with decreasing pupil size during the course of the task. There was no significant three-way interaction in either analysis, indicating that the Stroop effect on pupil size remained constant throughout the task.

For dwell time block and condition interacted significantly (Table 1), indicating that the size of the Stroop effect increased during the course of the task (Figure 2, top right). The significant linear and quadratic terms for block (Table 3) are consistent with a decelerating decrease in dwell-time, and the interaction of the linear term with condition indicates that the decrease was greater for the neutral than the incongruent condition.

Dense Multi-Item Top-to-Bottom Task

For response rate in the DTB task, the main effects of condition and block were significant (Table 1), consistent with a Stroop effect and a timecourse effect, respectively. Trend analysis (Table 2) produced significant linear, quadratic, and cubic effects for block, consistent with a decelerating decrease in response rate during the course of the task (indeed, an apparent asymptote; see second row on the left in Figure 1). Their interaction was not significant, consistent with no difference between the time course of response rate between the two conditions: Response rates in both the neutral and the incongruent condition decreased rapidly during the first column and remained approximately stable thereafter.

For errors, neither the main effects of block and condition nor their interaction reached significance (Table 1). No consistent polynomial trend emerged either (Table 2).

For pupil size, the interaction of block with scanning was significant (Table 1), consistent with a difference from the first block (controlling for position artifacts). The nonsignificant interaction of block with condition (and nonsignificant triple interaction with scanning) indicated that the effect of block on pupil size was similar in both conditions. Trend analysis confirmed a linear trend consistent with decreasing pupil size during the course of the task, not interacting with condition (Table 3 and Figure 2, second row left and middle).

The results for dwell time were similar, that is, significant main effects of block and condition in the absence of a significant interaction between them (Table 1), indicating that dwell time was similarly affected for both experimental conditions. More importantly, however, and in stark contrast to the single-item task, trend analysis showed a substantial linear *increase* in dwell time within the course of the task, not decelerating and not interacting with condition (Table 3 and Figure 2, second row right). The significant cubic effect is consistent with fluctuations in association with column changes (continuous vertical grey lines in the graph).

Sparse Multi-Item Top-to-Bottom task

For response rate in the STB task the interaction of block and condition was significant (Table 1), indicating that the substantial Stroop effect (significant main effect of condition) was not entirely stable across blocks. Trend analysis revealed significant linear and quadratic terms (Table 2), not interacting with condition, consistent with a decelerating decrease in response rate across conditions (Figure 1, third row left).

Turning to errors, neither the main effects of block and condition nor their interaction reached significance (Table 1). However, trend analysis indicated an overall linear increase within the task, which was steeper in the incongruent condition, evidenced by the significant interaction of condition with the linear trend of block (Table 2). Taken together, the findings for response rate and errors suggest that during the course of the STB task participants slowed down as their error rate increased (Figure 1, third row).

For pupil size, both condition and block interacted significantly with scanning, consistent with a Stroop effect and a timecourse effect, respectively (Table 1). Trend analysis confirmed a linear decrease in pupil size across conditions (Table 3 and Figure 2, third row left and middle).

For dwell time, there were again significant main effects of condition and block, and no interaction, suggesting a stable Stroop effect throughout the course of the task. Trend analysis confirmed a linear increase of dwell time during the task across conditions (Table 3 and Figure 2, third row right).

Sparse Multi-Item Left-to-Right Task

For response rate in the SLR task, the main effects of block and condition, as well as their interaction, were significant (Table 1), indicating a Stroop effect and a timecourse effect that was modulated by condition (as depicted in Figure 1, bottom left). Trend analysis confirmed a decelerating decrease in response rate across conditions during the task (significant linear and quadratic terms of block not interacting with conditions; Table 2).

The same pattern of results was found for errors, namely significant main effects and interaction of block and condition (Table 1). This amounted to a significant linear increase in error proportion that was greater for the incongruent condition, evidenced by the significant interaction of condition with the linear trend of block (Table 2 and Figure 1, bottom right). In sum, as for the STB task, the findings for the behavioral measures indicate a clear drop in performance during the task, namely a decrease in response rate accompanied by an in increase in errors.

For pupil size, there were significant interactions of condition and block with scanning, and no triple interaction, following the pattern observed in the STB multi-item task (Table 1). Trend analysis also confirmed a linear decrease of pupil size within the task, as in the previous tasks, except this time there was also a smaller significant cubic trend (Table 3), possibly due to an transient increase in the beginning of the task (Figure 2, bottom middle).

Finally, for dwell time, the SLR followed the same pattern as the other multi-item tasks, that is, significant main effects of block and condition in the absence of a significant interaction, consistent with a stable Stroop effect and an even timecourse effect across conditions (Table 1). Trend analysis confirmed a linear increase, with a steeper slope in the incongruent condition, evidenced by the significant interaction of condition with the linear trend of block (Table 3 and Figure 2, bottom right).

Figure 1

Results of behavioral measures. Response rate (items per second) and accuracy (error proportion) in each block, condition, and task. Error bars show within-participant 95% confidence intervals.

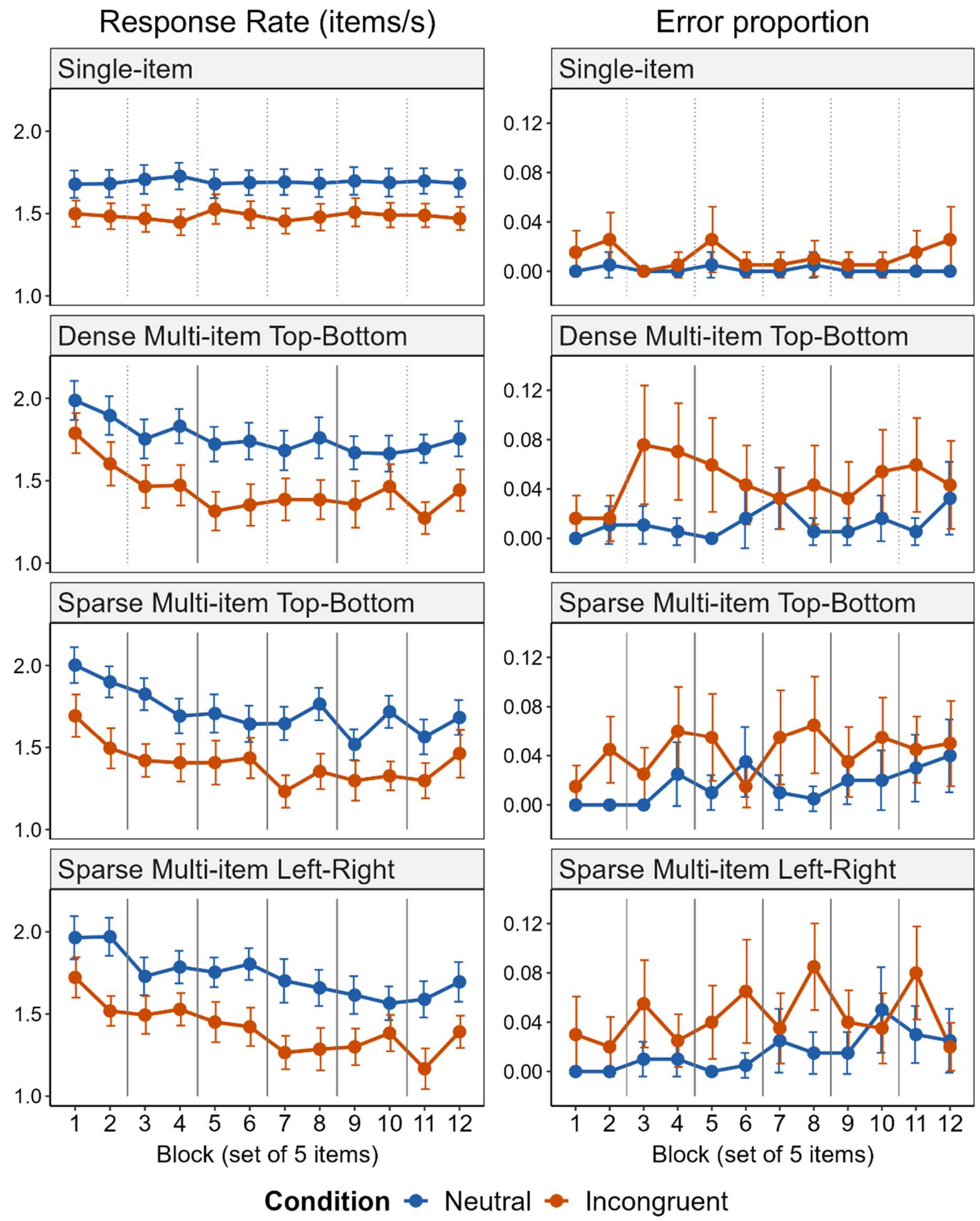

Figure 2

Results of eye-tracking measures. Raw pupil size (mm), adjusted pupil size (mm), and dwell time (ms) in each block, condition, and task. Adjusted pupil size is calculated as the difference between each experimental condition and the corresponding scanning baseline. Error bars show within-participant 95% confidence intervals.

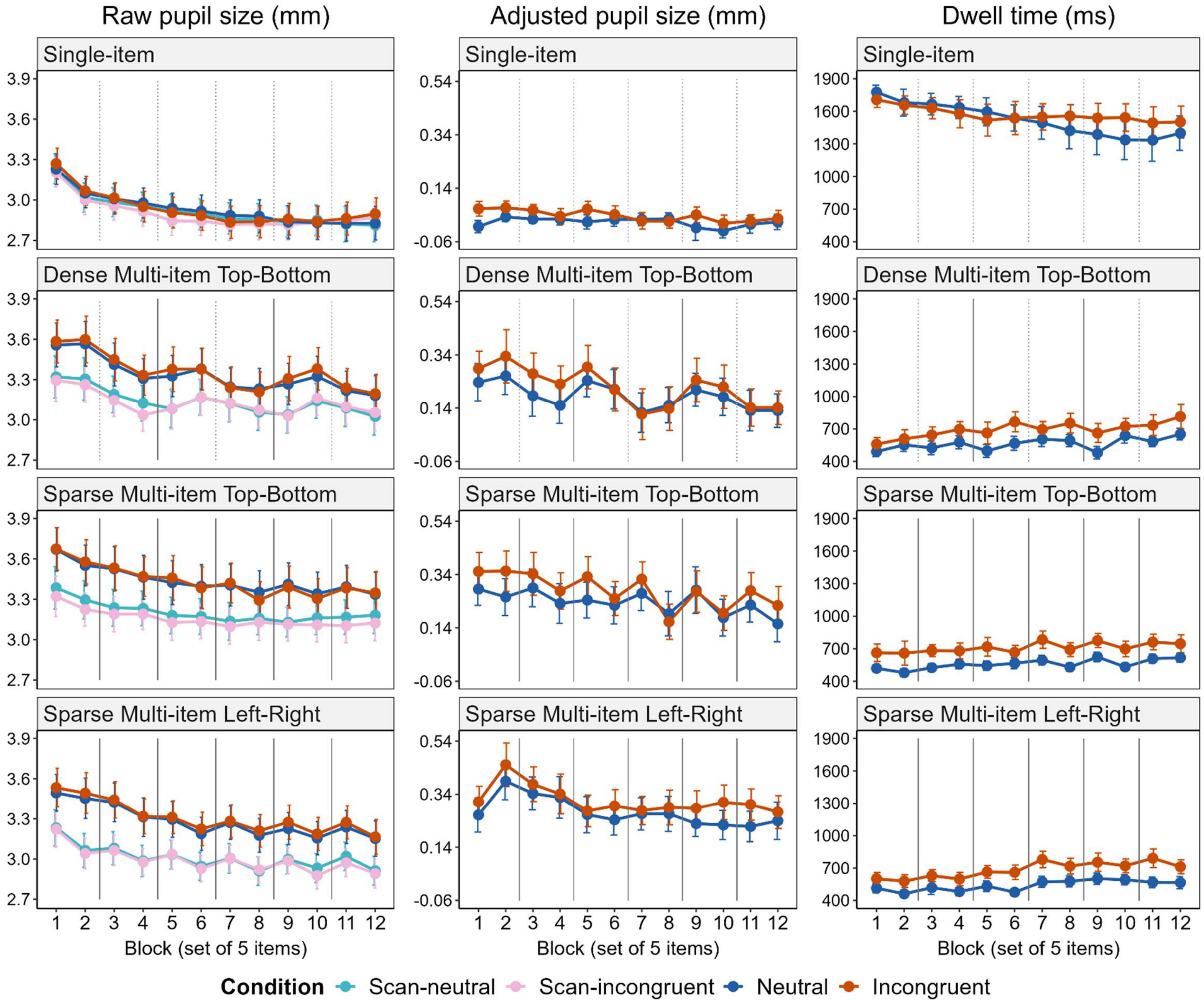

Table 1

ANOVA results for linear and generalized linear mixed-effects models for each dependent measure and task.

		Response Rate		Errors		Pupil Size		Dwell Time	
Task	*Fixed effect*	*F*	*p*	χ^2	*p*	*F*	*p*	*F*	*p*
SNI	condition	**59.45**	**< .001**	**4.09**	**.042**	0.68	.413	0.26	.606
	block	0.42	.945	8.80	.640	**203.21**	**< .001**	**10.84**	**< .001**
	scanning	–	–	–	–	**13.86**	**< .001**	–	–
	block × condition	1.30	.217	2.47	.996	**4.26**	**< .001**	**2.22**	**.011**
	condition × scanning	–	–	–	–	**6.41**	**.011**	–	–
	block × scanning	–	–	–	–	0.78	.658	–	–
	condition × block × scanning	–	–	–	–	0.45	.929	–	–
DTB	condition	**98.10**	**< .001**	3.59	.058	0.06	.793	**58.06**	**< .001**
	block	**17.54**	**< .001**	13.72	.248	**120.46**	**< .001**	**6.91**	**< .001**
	scanning	–	–	–	–	**56.41**	**< .001**	–	–
	block × condition	1.76	.056	15.64	.155	0.77	.664	0.74	.691
	condition × scanning	–	–	–	–	2.53	.120	–	–
	block × scanning	–	–	–	–	**9.82**	**< .001**	–	–
	condition × block × scanning	–	–	–	–	0.81	.628	–	–
STB	condition	**114.31**	**< .001**	3.02	.082	2.70	.108	**57.32**	**< .001**

	block	**21.00**	**< .001**	17.42	.095	**98.46**	**< .001**	**7.12**	**< .001**
	scanning	–	–	–	–	**93.79**	**< .001**	–	–
	block × condition	**2.38**	**.006**	18.25	.075	0.41	.948	0.60	.822
	condition × scanning	–	–	–	–	**6.61**	**.014**	–	–
	block × scanning	–	–	–	–	**8.69**	**< .001**	–	–
	condition × block × scanning	–	–	–	–	1.15	.314	–	–
SLR	condition	**161.88**	**< .001**	**5.09**	**.024**	0.23	.633	**214.56**	**< .001**
	block	**25.97**	**< .001**	**20.45**	**.039**	**142.32**	**< .001**	**11.36**	**< .001**
	scanning	–	–	–	–	**150.60**	**< .001**	–	–
	block × condition	**2.83**	**.001**	**20.64**	**.037**	0.36	.968	1.39	.170
	condition × scanning	–	–	–	–	**8.29**	**.006**	–	–
	block × scanning	–	–	–	–	**8.73**	**< .001**	–	–
	condition × block × scanning	–	–	–	–	0.56	.857	–	–

Note. Statistically significant effects indicated in boldface. SNI: Single-item Stroop task; DTB: Dense top-to-bottom multi-item Stroop task; STB: Sparse top-to-bottom multi-item Stroop task; SLR: Sparse left-to-right multi-item Stroop task.

Table 2

Polynomial effects of block and their interaction with condition for each behavioral measure in each task

		Response Rate						Errors					
		block			block × condition			block			block × condition		
Task	*Effect*	*β*	*t*	*p*	*β*	*z*	*p*	*β*	*t*	*p*	*β*	*t*	*p*
SNI	Linear	< .01	–0.12	.901	< .01	–0.08	.933	–0.23	–0.17	.862	–0.71	–0.53	.591
	Quadratic	< .01	–0.41	.675	< .01	–0.41	.677	0.29	0.22	.821	–1.04	–0.81	.416
	Cubic	< .01	–0.29	.767	.02	–1.42	.153	0.03	0.02	.981	0.07	0.05	.957
DTB	Linear	**–0.26**	**–9.81**	**< .001**	–0.02	–1.04	.294	1.26	1.91	.055	0.62	0.95	.342
	Quadratic	**0.22**	**8.15**	**< .001**	–0.05	–1.93	.052	–0.49	–0.69	.488	0.18	0.26	.791
	Cubic	**–0.07**	**–2.69**	**.007**	0.05	1.92	.054	1.00	1.43	.152	0.05	0.08	.931
STB	Linear	**–0.28**	**–10.93**	**< .001**	–0.04	–1.63	.102	**2.39**	**3.05**	**.002**	**1.61**	**2.06**	**.039**
	Quadratic	**0.22**	**8.45**	**< .001**	–0.02	–0.75	.448	–0.90	–1.31	.188	–0.50	–0.73	.459
	Cubic	–0.02	–1.02	.305	0.02	–1.06	.288	0.73	1.09	.272	0.38	0.57	.566
SLR	Linear	**–0.37**	**–14.57**	**< .001**	0.01	0.46	.638	**2.22**	**3.00**	**.002**	**1.81**	**2.45**	**.014**
	Quadratic	**0.15**	**5.94**	**< .001**	–0.02	–0.99	.320	–0.74	–1.01	.312	0.02	0.03	.969
	Cubic	0.03	1.30	.192	< .01	0.03	.975	–0.28	–0.42	.672	0.16	0.24	.807

Note. Statistically significant effects indicated in boldface. SNI: Single-item Stroop task; DTB: Dense top-to-bottom multi-item Stroop task; STB: Sparse top-to-bottom multi-item Stroop task; SLR: Sparse left-to-right multi-item Stroop task.

Table 3

Polynomial effects of block and their interaction with condition for each eye tracking measure in each task

		Pupil Size						Dwell Time					
		block × scan			block × condition × scan			block			block × condition		
Task	*Effects*	*β*	*t*	*p*	*β*	*z*	*p*	*β*	*t*	*p*	*β*	*t*	*p*
SNI	Linear	**–0.03**	**–2.26**	**.023**	0.01	1.08	.279	**–303.74**	**–10.52**	**< .001**	**–124.01**	**–4.29**	**< .001**
	Quadratic	< .01	–0.03	.974	–0.01	–0.86	.388	**72.47**	**2.51**	**.012**	–5.17	–0.17	.857
	Cubic	0.02	1.23	.215	< .01	0.50	.612	–10.51	–0.36	.715	42.27	1.46	.143
DTB	Linear	**–0.14**	**–5.97**	**< .001**	0.03	1.44	.147	**144.28**	**6.58**	**< .001**	–29.10	–1.32	.184
	Quadratic	0.02	0.92	.353	–0.01	–0.53	.596	–9.40	–0.42	.667	15.38	0.70	.482
	Cubic	–0.01	–0.53	.592	–0.01	–0.60	.543	**64.92**	**2.96**	**.003**	–19.78	–0.90	.366
STB	Linear	**–0.12**	**–7.11**	**< .001**	0.02	1.45	.146	**105.86**	**6.54**	**< .001**	3.01	0.18	.852
	Quadratic	0.01	0.58	.557	–0.02	–1.62	.104	–13.94	–0.86	.388	4.26	0.26	.792
	Cubic	< .01	–0.24	.809	–0.03	–1.74	.081	3.64	0.22	.821	12.05	0.74	.456
SLR	Linear	**–0.11**	**–6.76**	**< .001**	–0.01	–0.71	.474	**160.57**	**9.34**	**< .001**	**–41.17**	**–2.39**	**.016**
	Quadratic	0.02	1.34	.180	–0.01	–0.98	.326	–30.63	–1.78	.074	20.45	1.19	.234
	Cubic	**0.04**	**2.57**	**.010**	0.01	1.10	.270	**–64.68**	**–3.76**	**< .001**	–0.21	–0.01	.989

Note. For pupil size the polynomial effects of block refer to the interaction with scanning, to account for position artifacts (see text for variables and contrast coding). Statistically significant effects indicated in boldface. SNI: Single-item Stroop task; DTB: Dense top-to-bottom multi-item Stroop task; STB: Sparse top-to-bottom multi-item Stroop task; SLR: Sparse left-to-right multi-item Stroop task.

Summary of Results

To sum up and as shown in Table 4, for the single-item task response rate and errors were stable throughout the task, while pupil size and dwell time decreased. In contrast, for all multi-item tasks, response rate decreased, pupil size decreased, and dwell time increased; the pattern for errors was not consistent as it was stable in the dense task and increased in the two sparse tasks depending, however, on condition.

Table 4

Summary of linear trends of block for each task and measure

	Response Rate		Errors		Pupil Size		Dwell Time	
Task	Increase	Decrease	Increase	Decrease	Increase	Decrease	Increase	Decrease
SNI						✓		✓
DTB		✓				✓	✓	
STB		✓	✓			✓	✓	
SLR		✓	✓			✓	✓	

Note. SNI: Single-item Stroop task; DTB: Dense top-to-bottom multi-item Stroop task; STB: Sparse top-to-bottom multi-item Stroop task; SLR: Sparse left-to-right multi-item Stroop task.

Between-task Comparisons

Table 5 lists the results of comparing the polynomial trends of block between pairs of multi-item tasks, by testing the interaction of task with block. For the effects on pupil size, the three-way-interaction between task, block, and scanning is presented instead, to account for position artifacts. Condition (neutral vs. incongruent) was deviation-coded in the models, thus the results in Table 5 apply to the mean of the two conditions. Random effects for participant in models for pupil size for each comparison included random intercepts as well as random slopes for scanning, condition, task, and their interaction. For models that did not converge, a simpler random effects structure without the interaction was used (see online supplementary material). The selected pairs of tasks highlight the effect of item density (DTB vs. STB) and naming direction (STB vs. SLR).

Dense Versus Sparse Top-to-Bottom

There were no significant interactions of task with linear, quadratic, or cubic terms of block for response rate, suggesting that within-task performance was declining equally in both tasks. No significant interactions were observed for error proportions and pupil size. For dwell time, there was no significant interaction of task with the linear or quadratic trend of block, but the interaction with the cubic term was significant, indicating more fluctuations in the task with sparser items.

Sparse Top-to-Bottom versus Left-to-Right

The significant interaction of task with the linear trend of block for response rate indicates that the slope for the left-to-right task (SLR) was steeper, that is, faster decline in performance, than in the top-to-bottom task (STB). There was no difference between tasks in the trends for error proportions or pupil size. For dwell time, a significant interaction of task with the linear and cubic trend of block was observed, indicating a steeper and less even increase for the left-to-right task.

Table 5

Interactions between task and polynomial effects of block for each dependent measure

Task pair	*Effect*	Response Rate			Errors			Pupil Size			Dwell Time		
		β	*t*	*p*	*β*	*z*	*p*	*β*	*t*	*p*	*β*	*t*	*p*
DTB vs STB	Linear	–0.02	–0.55	.577	1.36	1.27	.201	0.02	0.59	.551	–38.38	–1.43	.151
	Quadratic	< .01	0.02	.976	–0.56	–0.55	.579	< .01	–0.25	.799	–4.66	–0.17	.861
	Cubic	0.04	1.22	.222	0.11	1.12	.902	< .01	0.13	.891	**–61.29**	**–2.29**	**.022**
STB vs SLR	Linear	**–0.08**	**–2.40**	**.016**	0.35	0.30	.759	< .01	0.24	.803	**54.70**	**2.30**	**.021**
	Quadratic	–0.06	–1.87	.061	–0.21	–0.19	.844	0.01	0.50	.614	–16.68	–0.70	.482
	Cubic	0.06	1.64	.100	–0.86	–0.89	.372	0.04	1.88	.059	**–68.33**	**–2.87**	**.004**

Note. Statistically significant effects indicated in boldface. For pupil size, the three-way-interaction between task (reference level listed first), block (polynomial contrast), and scanning (reference level was scanning) is presented; for the other dependent measures, results refer to the two-way interaction between task and block. DTB: Dense top-to-bottom multi-item Stroop task; STB: Sparse top-to-bottom multi-item Stroop task; SLR: Sparse left-to-right multi-item Stroop task.

Discussion

The starting point of the present study was the observation that the empirical foundation of popular cognitive control theories is largely based on single-item tasks, which may pose too low demands to challenge the capacity of the system and expose its limits. Consequently, we considered predictions derived from the conflict monitoring hypothesis and the expected value of control theory (CMH/EVC) for different versions of the Stroop task, that is, a single-item task and multi-item variants of the task, focusing on within-task evaluation of behavioral and eye-tracking measures.

Our results showed clear performance decrements within all multi-item variants, contrasting with stable performance in the single-item task. Notably, this was the case in both the incongruent and the neutral condition, irrespective of between-items distance and naming direction. These findings indicate that there are considerable differences between task formats in control demands and are consistent with conceptualization of the multi-item task as a "multi-task".

This is not the first time such results have been observed. Similar findings have been reported by Klein et al. (1997), who examined the effect of test duration on Stroop task performance, and Amtmann et al. (2007) in row-by-row analysis of multi-item naming tasks. The same pattern was also obtained by Ziaka and Protopapas (2022) applying within-task performance evaluation in Stroop tasks in three different experiments and in two developmental stages, that is, in adults and Grade 4–5 children. Ziaka and Protopapas proposed capacity constraints on control as a plausible explanation, arguing against attributing the findings to withdrawal from the task (i.e., disengagement due to high control demands). However, the available measures did not permit them to conclusively rule out alternative interpretations.

Aiming to shed further light on the issue, in the present study we augmented accuracy and response rate with pupillometry and gaze measures. Our results confirmed the expected patterns of performance for the single-item Stroop task. In contrast, performance in the multi-item Stroop tasks did not follow the same pattern. Specifically, within-task performance decreased, accompanied by pupil constriction and—crucially—dwell-time *increase*. We interpret these results as indicative of capacity constraints on control rather than disengagement from the task, consistent with the third alternative hypothesis.

The single-item Stroop task

It may seem puzzling that gradual pupil constriction was also observed in the course of the single-item task. However, apart from the quantitative difference (much less constriction in the single-item task), we submit that the origin of constriction is different between the two Stroop task versions. Specifically, in the single-item task performance was quite stable throughout, consistent with participants efficiently managing the task requirements. In contrast, in all multi-item tasks pupil constriction was accompanied by substantial performance deterioration. This pattern is similar to findings of studies examining pupil size and performance under dual-task conditions (Häuser et al., 2019; Kahneman et al.,1967; Karatekin, 2004; Recarte & Nunes, 2000; Recarte et al., 2008; Tapper et al., 2021).

Moreover, a clear decrease in dwell time was observed during the course of the single-item task, indeed greater in the neutral condition. This indicates a diminishing need to intensely attend to the items for successful performance. In contrast, in the multi-item task dwell time increased throughout the task, suggesting a need to intensify processing. Overall, the differential patterns of results between the single-item and multi-item version lend themselves to qualitatively different interpretations, namely practice effects for the single-item version versus capacity constraints for the multi-item version. Hershman and Henik (2019) have recently commented that “pupil changes could indicate conflict even in the absence of behavioral indications for the conflict" (p. 1899). By analogy, we submit that pupil changes could also indicate practice effects even in the absence of behavioral indications for these practice effects.

This argument is in line with CMH/EVC, which assume dynamical adaptations of control within the task depending on efficacy (Frömer et al., 2021). Here efficacy is understood as “the likelihood that a goal will be reached with a given investment of control”, leading to “changes in behavioral and neural signatures of control allocation” (Frömer et al., 2021, p. 2). Pupil size is an index of dynamic brain activity (Kahneman, 1973; Beatty, 1982; Beatty & Lucero-Wagoner, 2000); and the experience of efficacy due to stable performance within the single-item task was reflected in pupil size as (slight) gradual constriction. Indeed, similar results were obtained by Frömer et al. as they investigated the role of efficacy and reward in control implementation in a modified Stroop task: They also found that the pupil constricted, rather than dilate, when efficacy was perceived as high.

All things considered, the pattern observed in the single-item task is consistent with previous findings examining control implementation and allocation in single-item tasks. This was not the case for the multi-item tasks. Hence, our findings suggest that the use of multi-item tasks can help expand the scope of the study of control implementation beyond the limited demands of single-item laboratory tasks, which may differ substantially from the demands of multi-item tasks and, possibly, from the demands of more ecologically realistic tasks of everyday life (see also Draheim et al., 2021; Martin et al., 2020; Schuch et al., 2019).

In a related approach, Martin et al. (2020) set out to investigate the extent to which well-known measures of control could predict real-word (i.e., multi-tasking) job performance. They found that the single-item Stroop task and the flanker task failed to significantly predict multi-tasking performance, concluding that "the degree to which attention control provides unique prediction … depends on the way in which attention control is measured" (p. 332). By this they referred to the need for accuracy-based measures of attentional control, rather than the reaction time difference scores commonly employed in the field (see Draheim et al., 2021, for a detailed discussion of this topic). We believe that our findings point to an additional reason contributing to the null effects of Martin et al., that is, the use of single-item tasks for predicting complex everyday behaviors that are inherently more akin to multi-tasking.

The impact of capacity constraints on multi-item processing

EVC theorists acknowledge capacity constraints on control (Shenhav et al., 2013). Performance decrements in the model have been operationalized in terms of incentives, that is, overt disengagement due to task difficulty (Agrawal et al., 2022). Here we propose that performance decrements may also be operationalized in terms of strategic (moment-to-moment) control adaptation leading participants under cognitive overload to actively adapt to the task, rather than withdraw from it.

More specifically, we take our findings to show that in the multi-item Stroop task the need for control implementation goes well beyond the "normal" requirements of a single task, because of parallel execution of diverse subtasks necessitated by the simultaneous presence of multiple items to be named and the consequent within- and between-items interference.

The excessive control demands of the multi-item tasks has led to a pattern of findings—consistently observed across tasks—over the four dependent variables that suggests capacity limitations and control constraints. The effect of these limitations is quite substantial, consistent with the idea that if capacity limits are reached in the course of carrying out a task, cognitive overload results in within-task performance deterioration. This view is in line with findings from the multi-tasking literature showing that simultaneous processing implemented through short stimulus-onset-asynchronies (SOAs) has a detrimental effect on performance, and that greater performance decrements are associated with increased temporal overlap (Fischer & Plessow, 2015).

However, it is important that performance merely decreases; it does not collapse. This suggests that an overwhelmed control system can resort to mitigation strategies, declining in efficiency but retaining control. What is the cognitive mechanism that allows the task to be carried out, mostly successfully, despite excessive demands on cognitive control? Ziaka and Protopapas (2022) speculated that within-task performance decrements might be a consequence of shifting from parallel to serial processing once the limits of the available resources are reached.

Specifically, Ziaka and Protopapas (2022) argued that in tasks involving simultaneous presentation of multiple items, parallel processing is an emergent default behavior, in part thanks to parafoveal processing of upcoming items concurrent with processing of the foveated item, and in part because an upcoming item can be processed while the response to a preceding item is uttered. At the beginning of the task participants apply their default parallel strategy, entrenched as a result of years of reading experience; simultaneous processing of the current and nearby items allows one to proceed faster without sacrificing performance. However, as the task proceeds and overload emerges, conflict monitoring indicates the need to adapt control. This leads the goal of current-item processing to be prioritized over the processing of upcoming items in order to maintain successful performance, known as task shielding in the multi-tasking literature (Berger et al., 2019; Fischer & Hommel, 2012; Fischer & Plessow, 2015). In other words, serial processing (i.e., item-by-item processing; Fischer & Plessow, 2015) replaces the default parallel procedure. Attention is restricted to the currently processed item by adopting "lockout scheduling" (Meyer & Kieras, 1997, p. 20), meaning that subsequent items are excluded from processing until response planning of the current item has been completed (Roelofs, 2007).

We submit that lockout scheduling can be applied to the interpretation of the findings of the present study. Specifically, our results indicate that at the very beginning of the task participants were in general faster in the multi-item tasks, compared to the single-item (Figure 1, right), indicating a "serial advantage". The serial advantage has been previously attributed to parallel processing of successive items, such that one item is processed while the previous one is named and the next one is viewed (Altani et al., 2020; Protopapas et al., 2013, 2018). As the task progressed, performance started to drop and the serial advantage was abolished, especially in the case of the incongruent condition (see online Supplemental Material A, section 4.3). Importantly, pupil also began to constrict, indicating the emergence of an overload state. At the same time, dwell time showed an increase, meaning that participants spend more time looking at each item for response selection and suggesting that items captured more attentional resources. In other words, processing of the next item was postponed until the current response was free from between-items interference (i.e., "lockout scheduling"; Roelofs, 2007). Notably, the temporal separation of item processing that leads to decreased response rate in multi-item tasks is not posed by experimental manipulation, as it is in single-item or dual tasks (e.g., Esterman et al., 2013; Fischer et al., 2007; Glasser & Glasser, 1982, 1989; McCann & Johnston, 1992), but by participants' self-regulation, suggesting moment-to-moment adaptation.

This interpretation is in line with the proposal of Roelofs (2007), who examined eye movements in a picture-word Stroop task under dual task conditions and showed that the time participants "decided" to shift their gaze to the secondary task was strategically adjusted based on task requirements. In particular, gaze shift was delayed if the primary and the secondary task required the same response mapping (here, oral responding), ultimately resulting in a shift from parallel to serial processing. Taking into account that oral responses are required for all items of our multi-item tasks, implementation of a serial strategy seems justifiable—if not inevitable. Taking this one step further, in our experiment dwell time increase was accompanied by pupil size decrease, extending Roelofs' proposal by suggesting that it is not (only) a strategic decision, but (also) an emergent necessity due to the overload state in which participants find themselves.

Our interpretation receives further support from studies investigating attentional breadth as an index of shifts of covert attention (Brocher et al., 2018). More specifically, attentional breadth is based on the focus of attention, which is narrowed when the focal point of attention is centered, and broadens when attention is peripherally expanded (Mathôt, 2020). Importantly, shifts of attentional breadth can be reflected on pupil size, with narrowing of attention causing constriction and expansion leading to dilation (Brocher et al., 2018; Daniels et al., 2012; Mathôt, 2020). Moreover, attentional breadth is related to exploration and exploitation, with exploration causing pupil dilation, as attention is less focused, while exploitation is associated with smaller pupil size because of the narrowing of attention. Interestingly, exploitation and, hence, pupil constriction, is considered to emerge when the focus is on a single task, whereas exploration emerges when the focus of attention is expanded (e.g., in task switching), causing behavior to be prone to distraction and resulting in larger pupils (Mathôt, 2018).

Taking into account that parafoveal processing and perceptual span are all about expansion of attention (i.e., a case of attentional breadth in reading and naming tasks), studies investigating attentional breadth are highly relevant in the context of the present study. More specifically, they are consistent with our interpretation of a shift from a more parallel to a more serial processing via "lockout scheduling" on the assumption that attention is dynamically adjusted (i.e., narrowed), causing the pupil to shrink. This adjustment of attention has the effect of decelerating participants, due to more centered processing, while at the same time making them less prone to interference from nearby items (i.e., distraction). In other words, we submit that pupil constriction reflects an overload state because of the narrowing of attention; capacity constraints lead to a centered focused attention by blocking parallel processing, resulting in the well-documented pupil constriction under dual-task or excessive load conditions (e.g., Häuser et al., 2019; Karatekin, 2004; Recarte & Nunes, 2000; Recarte et al., 2008; Tapper et al., 2021).

This interpretation is also supported by the correlations between behavioral and eye-tracking measures in our study. Specifically, for the multi-item tasks, response rate correlated positively with pupil size but negatively with dwell time (while pupil size and dwell time were negatively correlated; see Tables S.10–S.13 in Supplemental Material B). Thus, it appears that pupil size decreased as participants were slowed down and the time looking at each item accordingly increased.

Finally, the comparison between the two sparse tasks, which showed that naming direction modulates the timecourse effect enhances our interpretation. Specifically, the task comparison indicated a steeper decrease of response rate and a steeper increase in dwell time for the left-to-right task, compared to top-to-bottom. This suggests that participants locked out nearby-items processing more intensively in the left-to-right task. This finding can be related to evidence for the effect of naming direction on parafoveal processing and the perceptual span. Specifically, previous studies have shown that parafoveal processing is larger in the default reading direction than in a vertical direction (Snell et al., 2018). As for the perceptual span, it can exceed 10–15 character spaces to the right of the fixation in alphabetical orthographies (Rayner, 1998), whereas it is limited to 4–5 character spaces in the vertical direction (Ojanpää et al., 2002). Taken together, we interpret these findings as suggesting that nearby-items interference is stronger in the left-to-right task, forcing participants to adopt a stricter serial strategy, compared to the top-to-bottom task.

Implications for clinical settings

Although it is known that Stroop interference appears inflated in the multi-item version of the task when compared to its single-item counterpart (MacLeod, 2005; Salo et al., 2001), the multi-item Stroop task is used in standardized neuropsychological assessment and clinical practice for identifying inhibition, control and attentional deficits (e.g., Björngrim et al., 2019; Bezdicek et al., 2015; Penner et al., 2012; Periáñez et al., 2021; Rabin et al., 2005; Salo et al., 2001; Scarpina & Tagini, 2017). The observed difference in interference between the two Stroop versions has been partly attributed to the neutral condition, which seems faster in the multi-item Stroop task than in the single-item task (Salo et al., 2001), presumably due to simultaneous presentation of items speeding up responses (MacLeod, 2005). In contrast, the incongruent condition seems to produce either the same or similar response times across versions. Thus, interference, which results from a subtraction of the two, appears larger in the multi-item version (Salo et al., 2001).

Focusing in finer detail on the time course of each condition, our results show that in the beginning of the task both conditions of the multi-item task produce faster responses than the corresponding conditions of the single-item task (see Supplemental Material A, section 4.3). Further into the task, however, this advantage vanishes, especially in the incongruent condition (see Figure 1 left). Previous studies seem to have missed this pattern and its impact on interference estimation because response times were averaged throughout the entire task, as per common practice. It is the continuous tracking of within-task performance that has allowed us to identify the origin of the differences. This finding seems particularly important in light of the popularity of the multi-item Stroop task in clinical practice and neuropsychological assessment (e.g., Rabin et al., 2005).

Finally, our results point to a significant role of naming direction. In particular, the steeper performance decrements observed in the left-to-right task (Table 5) suggest that naming in a left-to-right direction may be more challenging than naming top-to-bottom. This finding stands in contrast to the only past study on this issue we have been able to identify, namely McCown and Arnoult (1981), who found no significant effect of naming direction (i.e., vertical or horizontal) on response times for the whole task in four different variants of the multi-item Stroop task. They concluded that different Stroop forms can be used without distorting their conflicting nature. If replicated, our finding will be of relevance for clinical practice, bearing in mind that naming direction in popular standardized Stroop tasks varies greatly, with some of them using left-to-right naming (e.g., Victoria version) and others top-to-bottom (e.g., Golden version; for a review see Strauss et al., 2006), all striving to identify the same types of deficits.

Alternative interpretations

Any discussion of sequential effects on Stroop task performance should acknowledge the presence of “negative priming” items and their possible impact on performance. Negative priming refers to the increase of response time in the incongruent condition when the incorrect word-response of the preceding item corresponds to the correct color-response of the currently named item (e.g., the word “red” printed in green followed by an item printed in red color; Dalrymple-Alford & Budayr, 1966; Neill, 1977). Negative priming items were indeed present in all of the multi-item tasks used in the present study, with varying proportions among five-item blocks (see online task material). However, the consistent results of our dependent measures suggests that this effect cannot have been the primary source of the findings. More importantly, the fact that negative priming is by definition absent in the neutral condition of the task, which however showed a similar pattern of results, excludes negative priming as a valid interpretation for the overall pattern of our findings.

A potential alternative approach to our interpretation of within-task differences in the multi-item version might focus on speed-accuracy tradeoffs, defined as the switch between the tendency to respond slowly and more accurately and the tendency to respond faster and be less accurate (Zimmerman, 2011). However, there was no systematic decrease in error rate associated with increased naming time, therefore our findings cannot be attributed to a speed-accuracy tradeoff. The dissociation is further highlighted by the observation that color naming times showed a decrease in both conditions even when error rate was not significantly affected (e.g., Dense multi-item task) or when error rate showed a steeper increase (e.g., incongruent conditions of sparse tasks when compared to the corresponding neutral conditions; Table 2).

Finally, post-error slowing does not appear to be a plausible alternative interpretation either. Post-error slowing refers to the tendency of participants to slow down and be more conservative after errors (Carter & van Veen, 2007). In the present context, post-error slowing could account for an increase in color naming times between blocks if increased naming times were systematically accompanied by increased numbers of errors across conditions and multi-item tasks. However, our results showed that errors varied widely among conditions and tasks, in contrast to the robust increase of color-naming time.

Limitations and future directions

In the present study, participants were instructed to name the color "as fast as possible and try to avoid errors". The instruction thus put emphasis on speed, as usually done in the context of the Stroop task and other interference tasks (MacLeod, 1991, 2005; Draheim et al., 2021). However, previous studies have shown the modulating role of instructions (Fischer & Hommel, 2012; Lehle & Hübner 2009; Lehle et al., 2009). Hence, by emphasizing speed we may have indirectly promoted parallel processing for speeding up responses and, consequently, nearby-items interference. Future studies can shed more light on possible effects of instructions on within-task performance.

Furthermore, we examined within-task variations in only one of the tasks used in the cognitive control literature (i.e., the Stroop task). However, as discussed in the introduction, in the multi-item Stroop task two different inhibition processes are active, that is, prepotent response inhibition and resistance to distractor inhibition (Friedman & Miyake, 2004). It remains unclear if the same pattern of results would be obtained in tasks where only one type of inhibition is required, as for example in a modified multi-item Eriksen flanker task composed by visual symbols. Further experiments using different tasks and materials are needed before final conclusions about the robustness of our findings and the task specificity of the relevant cognitive theories can be confidently reached.

Finally, all participants of our study were young adults (*M* = 19.52). Age may modulate the effect of within- and between-items interference in the multi-item Stroop task. The difference between adults and children reported by Ziaka and Protopapas (2022) strengthens this possibility. Future research including eye-tracking measures is needed to examine potential differences in capacity limits between different developmental changes and populations and their impact on interference. Such studies are also important for the purpose of guiding clinical assessment.

It should be clarified that none of our findings are meant to be taken as invalidating or contradicting the CMH/EVC framework. Indeed EVC theory is sufficiently flexible in the conceptualization of costs and benefits that it may be difficult to pin down specific predictions. Although performance enhancement due to control adaptation would normally be predicted by the CMH in the course of a control-requiring task (Botvinick et al., 2001), the EVC can also accommodate performance decrements as a function of incentives, that is, disengagement (Shenhav et al., 2013). Instead, we take our findings to indicate that in multi-item tasks (and situations posing complex multi-tasking demands more generally) capacity constraints may be seen as an alternative explanation for the occurrence of performance decrements; and we proposed a way in which control adaptation under capacity constraints can be operationalized in terms of attentional shifts.

Other recent work has also led to a call for reconsideration of performance decrements during demanding tasks beyond incentives. Specifically, Agrawal et al. (2022) proposed that voluntary rests, conceptualized as opportunities for offline computations, can lead to performance decrements. Because rests subserve learning, they can be rewarding for future behavior despite current negative effects on performance and can be seen as indexes of covert engagement—not disengagement—related to exploitation, consistent with our interpretation of pupil constriction as an index of exploitation. Future studies can directly address whether participants in our study may have pursued rests for the purpose of learning during the multi-item tasks or whether, as we propose, it was capacity constraints and the concomitant attentional shifts that caused performance to drop and pupil size to shrink. In our view, the continuous nature of multi-item tasks makes rests unlikely. Moreover, performance in our multi-item tasks indicated consistent deterioration, in contrast to the purported rewarding nature of rests due to learning effects. Further research examining silent intervals between articulations and comparing multi-item tasks with item lists of varying length can help illuminate the underlying source of performance decrements.

Finally, it should be noted that the proposed shift from more parallel to more serial processing via attentional lockout scheduling is a post-hoc interpretation. Future research can examine within-task changes in spatial eye-voice span (i.e., the number of items between the currently fixated item and the currently named item) and temporal eye-voice span (i.e., the elapsed time between first fixation on an item and articulation onset of that item). The effects of shifting to a more serial processing mode should be evident as a decrease of spatial eye-voice span in conjunction with an increase in temporal eye-voice span during the course of the task.

Conclusion

The present study showed that different variants of the same task may involve qualitatively distinct control requirements and mechanisms leading to successful performance. The combination of pupillometry, eye tracking, and behavioral response measures allows us to preclude withdrawal from the task as a plausible explanation for the observed patterns during the course of carrying out the tasks. Instead, we propose that capacity constraints under concurrent control-demanding tasks result in a shift from a more parallel to a more serial processing. It remains to be seen whether—and how—our proposal can be incorporated into current theories of cognitive control.

References

Agrawal, M., Mattar, M. G., Cohen, J. D., & Daw, N. D. (2022). The temporal dynamics of opportunity costs: A normative account of cognitive fatigue and boredom. *Psychological Review, 129*(3), 564–585. https://doi.org/10.1037/rev0000309

Altani, A., Protopapas, A., Katopodi, K., & Georgiou, G. K. (2020). Tracking the serial advantage in the naming rate of multiple over isolated stimulus displays. *Reading and Writing*, *33*(2), 349–375. https://doi.org/10.1007/s11145-019-09962-7

Amtmann, D., Abbott, R. D., & Berninger, V. W. (2007). Mixture growth models of RAN and RAS row by row: Insight into the reading system at work over time. *Reading and Writing*, *20*(8), 785–813. https://doi.org/10.1007/s11145-006-9041-y

Augustinova, M., Silvert, L., Spatola, N., & Ferrand, L. (2018). Further investigation of distinct components of Stroop interference and of their reduction by short response-stimulus intervals. *Acta Psychologica*, *189*, 5462. https://doi.org/10.1016/j.actpsy.2017.03.009

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss. v067.i01

Beatty J. (1982). Task-evoked pupillary responses, processing load, and the structure of processing resources. *Psychological bulletin*, *91*(2), 276–292.

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In J. T. Cacioppo, L. G. Tassinary, & G. Berntson (Eds.), *Handbook of psychophysiology* (pp. 142–162). Cambridge, MA: Cambridge University Press

Berger, A., Fischer, R., & Dreisbach, G. (2019). It’s more than just conflict: The functional role of congruency in the sequential control adaptation. *Acta Psychologica*, *197*, 64–72. https://doi.org/10.1016/j.actpsy.2019.04.016

Bezdicek, O., Lukavsky, J., Stepankova, H., Nikolai, T., Axelrod, B. N., Michalec, J., Růžička, E., & Kopecek, M. (2015). The Prague Stroop Test: Normative standards in older Czech adults and discriminative validity for mild cognitive impairment in Parkinson's disease. *Journal of clinical and experimental neuropsychology*, *37*(8), 794–807. https://doi.org/10.1080/13803395.2015.1057106

Björngrim, S., van den Hurk, W., Betancort, M., Machado, A., & Lindau, M. (2019). Comparing traditional and digitized cognitive tests used in standard clinical evaluation—A study of the digital application minnemera. *Frontiers in Psychology*, *10*, 2327. https://doi.org/10.3389/fpsyg.2019.02327

Boersma, P., & Weenink, D. (2012). *Praat: Doing phonetics by computer* (Version 5.3.17). [Computer program]. http://www.praat.org.

Botvinick, M. M., Braver, T. S., Barch, D. M., Carter, C. S., & Cohen, J. D. (2001). Conflict monitoring and cognitive control. *Psychological Review*, *108*(3), 624–652. https://doi.org/10.1037/0033-295X.108.3.624

Botvinick, M. M., & Cohen, J. D. (2014). The computational and neural basis of cognitive control: Charted territory and new frontiers. *Cognitive Science*, *38*(6), 1249–1285. https://doi.org/10.1111/cogs.12126

Botvinick, M. M., Cohen, J. D., & Carter, C. S. (2004). Conflict monitoring and anterior cingulate cortex: An update. *Trends in Cognitive Sciences*, *8*, 539–546. https://doi.org/10.1016/j.tics.2004.10.003

Braem, S., Coenen, E., Bombeke, K., van Bochove, M. E., & Notebaert, W. (2015). Open your eyes for prediction errors. *Cognitive, Affective, & Behavioral Neuroscience*, *15(2)*, 374–380. https://doi.org/10.3758/s13415-014-0333-4

Brocher, A., Harbecke, R., Graf, T., Memmert, D., & Hüttermann, S. (2018). Using task effort and pupil size to track covert shifts of visual attention independently of a pupillary light reflex. *Behavior Research Methods*, *50(6)*, 2551–2567. https://doi.org/10.3758/s13428-018-1033-8

Brown, S. B., van Steenbergen, H., Kedar, T., & Nieuwenhuis, S. (2014). Effects of arousal on cognitive control: empirical tests of the conflict-modulated Hebbian-learning hypothesis. *Frontiers in Human Neuroscience*, *8*, 23. https://doi.org/10.3389/fnhum.2014.00023

Busjahn, T., Bednarik, R., & Schulte, C. (2014). What influences dwell time during source code reading? Analysis of element type and frequency as factors. *Eye Tracking Research and Applications Symposium (ETRA)*. 335–338. https://doi.org/10.1145/2578153.2578211

Carter, C. S., & van Veen, V. (2007). Anterior cingulate cortex and conflict detection: An update of theory and data. *Cognitive*, *Affective*, & *Behavioral Neuroscience*, *7,* 367– 379. https://doi.org/10.3758/CABN.7.4.367

Chung, Y., Rabe-Hesketh, S., Dorie, V., Gelman, A., & Liu, J. (2013). A nondegenerate penalized likelihood estimator for variance parameters in multilevel models. *Psychometrika*, *78*(4), 685–709. https://doi.org/10.1007/s11336-013-9328-2

Dalrymple-Alford, E. C., & Budayr, B. (1966). Examination of some aspects of the Stroop Color-Word Test. *Perceptual and Motor Skills, 23*(3, PT. 2), 1211–1214. https://doi.org/10.2466/pms.1966.23.3f.1211

Daniels, L. B., Nichols, D. F., Seifert, M. S., & Hock, H. S. (2012). Changes in pupil diameter entrained by cortically initiated changes in attention. *Visual Neuroscience*, *29(2)*, 131–142. https://doi.org/10.1017/S0952523812000077

Draheim, C., Tsukahara, J. S., Martin, J. D., Mashburn, C. A., & Engle, R. W. (2021). A toolbox approach to improving the measurement of attention control. *Journal of Experimental Psychology: General, 150*(2), 242–275. https://doi.org/10.1037/xge0000783

Eckstein, M. K., Guerra-Carrillo, B., Miller Singley, A. T., & Bunge, S. A. (2017). Beyond eye gaze: What else can eyetracking reveal about cognition and cognitive development?. *Developmental Cognitive Neuroscience*, *25*, 69–91. https://doi.org/10.1016/j.dcn.2016.11.001

Egner, T., & Hirsch, J. (2005). The neural correlates and functional integration of cognitive control in a Stroop task. *NeuroImage*, *24*(2), 539–547. https://doi.org/10.1016/j.neuroimage.2004.09.007

Eriksen, B. A., & Eriksen, C. W. (1974). Effects of noise letters upon the identification of a target letter in a nonsearch task. *Perception* & *Psychophysics*, *16*(1), 143–149. https://doi.org/10.3758/BF03203267

Esterman, M., Noonan, S. K., Rosenberg, M., & Degutis, J. (2013). In the zone or zoning out? Tracking behavioral and neural fluctuations during sustained attention. *Cerebral cortex*, *23*(11), 2712–2723. https://doi.org/10.1093/cercor/bhs261

Fischer, R., & Hommel, B. (2012). Deep thinking increases task-set shielding and reduces shifting flexibility in dual-task performance. *Cognition*, *123*(2), 303–307. https://doi.org/10.1016/j.cognition.2011.11.015

Fischer, R., Miller, J., & Shubert, T. (2007). Evidence for parallel semantic memory retrieval in dual tasks. *Memory & Cognition*, *35*(7), 1685–1699. https://doi.org/10.3758/bf03193502

Fischer, R., & Plessow, F. (2015). Efficient multitasking: Parallel versus serial processing of multiple tasks. *Frontiers in Psychology*, *6*. https://doi.org/10.3389/fpsyg.2015.01366

Fortenbaugh, F. C., DeGutis, J., & Esterman, M. (2017). Recent theoretical, neural, and clinical advances in sustained attention research. *Annals of the New York Academy of Sciences*, *1396*(1), 70–91. https://doi.org/10.1111/nyas.13318

Friedman, N. P., & Miyake, A. (2004). The Relations among inhibition and interference control functions: A latent-variable analysis. *Journal of Experimental Psychology: General*, *133(1)*, 101–135. https://doi.org/10.1037/0096-3445.133.1.101

Frömer, R., Lin, H., Dean Wolf, C. K., Inzlicht, M., & Shenhav, A. (2021). Expectations of reward and efficacy guide cognitive control allocation. *Nature Communications*, *12*(1). https://doi.org/10.1038/s41467-021-21315-z

Gagl, B., Hawelka, S., & Hutzler, F. (2011). Systematic influence of gaze position on pupil size measurement: Analysis and correction. *Behavior Research Methods*, *43*(4), 1171–1181. https://doi.org/10.3758/s13428-011-0109-5

Geller, J., Winn, M. B., Mahr, T., & Mirman, D. (2020). GazeR: A Package for Processing Gaze Position and Pupil Size Data. *Behavior Research Methods*, *52*(5), 2232–2255. https://doi.org/10.3758/s13428-020-01374-8

Glaser, M. O., & Glaser, W. R. (1982). Time course analysis of the Stroop phenomenon. *Journal of Experimental Psychology: Human Perception and Performance, 8*(6)*,* 875–894. https://doi.org/10.1037/0096-1523.8.6.875

Glaser, W. R., & Glaser, M. O. (1989). Context effects in Stroop-like word and picture processing. *Journal of Experimental Psychology: General*, *118*(1), 13–42. https://doi.org/10.1037/0096-3445.118.1.13

Golden, C. J. (1978). *Stroop Color and Word Test: A manual for clinical and experimental uses*. Chicago, IL: Stoelting Co.

Golden, C. J., & Freshwater, S. M. (2002). *Stroop Color and Word Test: Revised examiner's manual.* Wood Dale, IL: Stoelting Co.

Gordon, P. C., & Hoedemaker, R. S. (2016). Effective scheduling of looking and talking during rapid automatized naming. *Journal of Experimental Psychology: Human Perception and Performance*, *42*(5), 742–760. https://doi.org/10.1037/xhp0000171

Granholm, E., Asarnow, R. F., Sarkin, A. J., & Dykes, K. L. (1996). Pupillary responses index cognitive resource limitations. *Psychophysiology*, *33*(4), 457–461. https://doi.org/10.1111/j.1469-8986.1996.tb01071.x

Granholm, E., Morris, S. K., Sarkin, A. J., Asarnow, R. F., & Jeste, D. V. (1997). Pupillary responses index overload of working memory resources in schizophrenia. *Journal of Abnormal Psychology, 106*(3), 458–467. https://doi.org/10.1037/0021-843X.106.3.458

Green P., & MacLeod, C. J. (2016). simr: an R package for power analysis of generalised linear mixed models by simulation. *Methods in Ecology and Evolution*, 7(4), 493–498. https://CRAN.R-project.org/package=simr

Harms, L., & Bundesen, C. (1983). Color segregation and selective attention in a nonsearch task. *Perception & Psychophysics*, *33*(1), 11–19. https://doi.org/10.3758/bf03205861

Hasshim, N., & Parris, B. A. (2015). Assessing stimulus-stimulus (semantic) conflict in the Stroop task using saccadic two-to-one color response mapping and preresponse pupillary measures. *Attention, Perception, & Psychophysics*, *77*(8), 2601–2610. https://doi.org/10.3758/s13414-015-0971-9

Häuser, K. I., Demberg, V., & Kray, J. (2019). Effects of aging and dual-task demands on the comprehension of less expected sentence continuations: Evidence from pupillometry. *Frontiers in Psychology*, *10*, 709. https://doi.org/10.3389/fpsyg.2019.00709

Henry, R., van Dyke, J.A., & Kuperman, V. (2018). Oculomotor planning in RAN and reading: a strong test of the visual scanning hypothesis. *Reading and Writing*, *31*, 1619–1643. https://doi.org/10.1007/s11145-018-9856-3

Hershman, R., & Henik, A. (2019). Dissociation between reaction time and pupil dilation in the Stroop task. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *45*(10), 1899–1909. https://doi.org/10.1037/xlm0000690

Hershman, R., Levin, Y., Tzelgov, J., & Henik, A. (2020). Neutral stimuli and pupillometric task conflict. *Psychological Research*, *85*(3), 1084–1092. https://doi.org/10.1007/s00426-020-01311-6

Hess, E. H., & Polt, J. M. (1964). Pupil Size in relation to mental activity during simple problem-solving. *Science (New York, N.Y.)*, *143*(3611), 1190–1192. https://doi.org/10.1126/science.143.3611.1190

Hope, R. M. (2013). Rmisc: Ryan Miscellaneous. R package version 1.5. https://CRAN.R-project.org/package=Rmisc

Huang, Y. T. (2018). Real-time coordination of visual and linguistic processes in novice readers. *Journal of Experimental Child Psychology*, *173*, 388–396. https://doi.org/10.1016/j.jecp.2018.02.010

Inhoff, A. W., Gregg, J., & Radach, R. (2018). Eye movement programming and reading accuracy. *Quarterly Journal of Experimental Psychology*, *71*(1), 3–10. https://doi.org/10.1080/17470218.2016.1226907

Johnson, E. L., Miller Singley, A. T., Peckham, A. D., Johnson, S. L., & Bunge, S. A. (2014). Task-evoked pupillometry provides a window into the development of short-term memory capacity. *Frontiers in Psychology*, *5*, 218. https://doi.org/10.3389/fpsyg.2014.00218

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

Kahneman, D., Beatty, J., & Pollack, I. (1967). Perceptual deficit during a mental task. *Science, 157*(3785), 218–219. https://doi.org/10.1126/science.157.3785.218

Kalanthroff, E., Goldfarb, L., Usher, M., & Henik, A. (2013). Stop interfering: Stroop task conflict independence from informational conflict and interference. *Quarterly Journal of Experimental Psychology*, *66*(7), 1356–1367. https://doi.org/10.1080/17470218.2012.741606

Karatekin, C. (2004). Development of attentional allocation in the dual task paradigm. *International Journal of Psychophysiology, 52*(1), 7–21. https://doi.org/10.1016/j.ijpsycho.2003.12.002

Kellar, M., Watters, C., Duffy, J., & Shepherd, M. (2004). Effect of task on time spent reading as an implicit measure of interest. *Proceedings of the American Society for Information Science and Technology*, *41*(1), 168–175.

Kellough, J. L., Beevers, C. G., Ellis, A. J., & Wells, T. T. (2008). Time course of selective attention in clinically depressed young adults: an eye tracking study. *Behaviour Research and Therapy*, *46*(11), 1238–1243. https://doi.org/10.1016/j.brat.2008.07.004

Klein, M., Ponds, R. W., Houx, P. J., & Jolles, J. (1997). Effect of test duration on age-related differences in Stroop interference. *Journal of Clinical and Experimental neuropsychology*, *19*(1), 77–82. https://doi.org/10.1080/01688639708403838

Kleiter, G. D., & Schwarzenbacher, K. (1989). Beyond the answer: Post-error processes. *Cognition*, *32*(3), 255–277. https://doi.org/10.1016/0010-0277(89)90037-1

Koch, I., Poljac, E., Müller, H., & Kiesel, A. (2018). Cognitive structure, flexibility, and plasticity in human multitasking—An integrative review of dual-task and task-switching research. *Psychological Bulletin*, *144*(6), 557–583. https://doi.org/10.1037/bul0000144

Kuperman, V., Van Dyke, J. A. and Henry, R. (2016). Eye movement control in RAN and reading. *Scientific Studies of Reading*, *20*(2), 173–188. https://doi.org/10.1080/10888438.2015.1128435

Laeng, B., Ørbo, M., Holmlund, T., & Miozzo, M. (2011). Pupillary Stroop effects. *Cognitive Processing*, *12*(1), 13–21. https://doi.org/10.1007/s10339-010-0370-z

Laeng, B., Sirois, S., & Gredebäck, G. (2012). Pupillometry: A window to the preconscious? *Perspectives on Psychological Science*, *7*(1), 18–27. https://doi.org/10.1177/1745691611427305

Lehle, C., & Hübner, R. (2009). Strategic capacity sharing between two tasks: Evidence from tasks with the same and with different task sets. *Psychological Research*, *73*(5), 707–726. https://doi.org/10.1007/s00426-008-0162-6

Lehle, C., Steinhauser, M., & Hübner, R. (2009). Serial or parallel processing in dual tasks: What is more effortful? *Psychophysiology*, *46*(3), 502–509. https://doi.org/10.1111/j.1469-8986.2009.00806.x

Levin, Y., & Tzelgov, J. (2014). Conflict components of the Stroop effect and their "control". *Frontiers in Psychology*, *5*. https://doi.org/10.3389/fpsyg.2014.00463

Liu, J. & Belkin, N.J. (2010). Personalizing information retrieval for multi-session tasks: The roles of task stage and task type. In Proceedings of the 33rd Annual International ACM SIGIR Conference on Research & Development on Information Retrieval (SIGIR '10). Geneva, Switzland, July 19-23, 2010.

Liversedge, S. P., Paterson, K. B., & Pickering, M. J. (1998). Eye movements and measures of reading time. In G. Underwoon (Ed.), *Eye guidance in reading and scene perception* (pp. 55–75). Elsevier Science. https://doi.org/10.1016/ B978-0-08-043361-5.X5000-7

Ludwig, C., Borella, E., Tettamanti, M., & de Ribaupierre, A. (2010). Adult age differences in the Color Stroop Test: A comparison between an item-by-item and a blocked version. *Archives of Gerontology and Geriatrics*, *51*(2), 135–142. https://doi.org/10.1016/j.archger.2009.09.040

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: An integrative review. *Psychological Bulletin*, *109*(2), 163–203. https://doi.org/10.1037/0033-2909.109.2.163

MacLeod, C.M. (2005). The Stroop task in cognitive research. In A. Wenzel & D. S. Rubin (Eds.) *Cognitive methods and their application to clinical research* (pp. 17–40). Washington, DC: American Psychological Association.

Maier, M. E., Ernst, B., & Steinhauser, M. (2019). Error-related pupil dilation is sensitive to the evaluation of different error types. *Biological Psychology*, *141*, 25–34. https://doi.org/10.1016/j.biopsycho.2018.12.013

Martin, J., Mashburn, C. A., & Engle, R. W. (2020). Improving the validity of the Armed Service Vocational Aptitude battery with measures of attention control. *Journal of Applied Research in Memory and Cognition*, *9(3)*, 323-335. https://doi.org/10.1016/j.jarmac.2020.04.002

Mathôt, S. (2013). A Simple Way to Reconstruct Pupil Size During Eye Blinks. Retrieved from doi: https://doi.org/10.6084/m9.figshare. 688001

Mathôt, S. (2018). Pupillometry: Psychology, Physiology, and Function. *Journal of Cognition*, *1(1)*, 16. https://doi.org/10.5334/joc.18

Mathôt S. (2020). Tuning the senses: How the pupil shapes vision at the earliest stage. *Annual Review of Vision Science*, *6*, 433–451. https://doi.org/10.1146/annurev-vision-030320-062352

Mathôt, S., Fabius, J., Van Heusden, E., & Van der Stigchel, S. (2018). Safe and sensible preprocessing and baseline correction of pupil-size data. *Behavior Research Methods*, *50*(1), 94–106. https://doi.org/10.3758/s13428-017-1007-2

McCann, R. S., & Johnston, J. C. (1992). Locus of the single-channel bottleneck in dual-task interference. *Journal of Experimental Psychology: Human Perception and Performance, 18*(2), 471–484. https://doi.org/10.1037/0096-1523.18.2.471

McCown, D. A. & Arnoult, M.D. (1981). Interference produced by modified Stroop stimuli. *Bulletin of Psychonomic Society, 17*, 5–7. https://doi.org/10.3758/BF03333649

Meyer, D. E., & Kieras, D. E. (1997). A computational theory of executive cognitive processes and multiple-task performance: Part I. Basic mechanisms. *Psychological Review*, *104*(1), 3–65. https://doi.org/10.1037/0033-295x.104.1.3

Moll, K., & Jones, M. (2013). Naming fluency in dyslexic and nondyslexic readers: Differential effects of visual crowding in foveal, parafoveal, and peripheral vision. *Quarterly Journal of Experimental Psychology*, *66*(11), 2085–2091. https://doi.org/10.1080/17470218.2013.840852

Monsell, S. (1996). Control of mental processes. In V. Bruce (Ed.), *Unsolved mysteries of the mind* (pp. 93–148). Hove, England: Erlbaum.

Muraven, M., & Baumeister, R. F. (2000). Self-regulation and depletion of limited resources: Does self-control resemble a muscle? *Psychological Bulletin*, *126*(2), 247–259. https://doi.org/10.1037/0033-2909.126.2.247

Nee, D. E., Wager, T. D., & Jonides, J. (2007). Interference resolution: Insights from a meta-analysis of neuroimaging tasks. *Cognitive, Affective, & Behavioral Neuroscience*, *7*(1), 1–17. https://doi.org/10.3758/cabn.7.1.1

Neill, W. T. (1977). Inhibitory and facilitatory processes in selective attention. *Journal of Experimental Psychology: Human Perception and Performance, 3*(3), 444–450. https://doi.org/10.1037/0096-1523.3.3.444

Ojanpää, H., Näsänen, R., & Kojo, I. (2002). Eye movements in the visual search of word lists. *Vision Research*, *42(12)*, 1499–1512. https://doi.org/10.1016/s0042-6989(02)00077-9

Olk, B. (2013). Measuring the allocation of attention in the Stroop task: evidence from eye movement patterns. *Psychological research*, *77*(2), 106–115. https://doi.org/10.1007/s00426-011-0405-9

Pan, J., Yan, M., Laubrock, J., Shu, H., & Kliegl, R. (2013). Eye–voice span during rapid automatized naming of digits and dice in Chinese normal and dyslexic children. *Developmental Science*, *16*(6), 967–979. https://doi.org/10.1111/desc.12075

Peavler, W. S. (1974). Pupil size, information overload, and performance differences. *Psychophysiology*, *11*(5), 559–566. https://doi.org/10.1111/j.1469-8986.1974.tb01114.x

Penner, I., Kobel, M., Stocklin, M., Weber, P., Opwis, K., & Calabrese, P. (2012). The Stroop task: Comparison between the original paradigm and computerized versions in children and adults. *The Clinical Neuropsychologist*, *26*(7), 1142–1153. https://doi.org/10.1080/13854046.2012.713513

Perea, M., & Gomez, P. (2012). Increasing interletter spacing facilitates encoding of words. *Psychonomic Bulletin & Review*, *19*(2), 332–338. https://doi.org/10.3758/s13423-011-0214-6

Periáñez, J. A., Lubrini, G., García-Gutiérrez, A., & Ríos-Lago, M. (2021). Construct validity of the Stroop Color-Word Test: Influence of speed of visual Search, verbal fluency, working memory, cognitive flexibility, and conflict monitoring. *Archives of Clinical Neuropsychology: The Official Journal of the National Academy of Neuropsychologists*, *36*(1), 99–111. https://doi.org/10.1093/arclin/acaa034

Poock, G. K. (1973). Information processing vs pupil diameter. *Perceptual and Motor Skills*, *37*(3), 1000–1002. https://doi.org/10.2466/pms.1973.37.3.1000

Protopapas, A. (2007). CheckVocal: A program to facilitate checking the accuracy and response time of vocal responses from DMDX. *Behavior Research Methods*, *39*, 859–862. https://doi.org/10.3758/BF03192979

Protopapas, A., Tzakosta, M., Chalamandaris, A., & Tsiakoulis, P. (2012). IPLR: An online resource for Greek word-level and sublexical information. *Language Resources & Evaluation*, *46*, 449–459. https://doi.org/10.1007/s10579-010-9130-z

Protopapas, A., Altani, A., & Georgiou, G. K. (2013). Development of serial processing in reading and rapid naming. *Journal of Experimental Child Psychology*, *116(4)*, 914–929. https://doi.org/10.1016/j.jecp.2013 .08.004

Protopapas, A., Katopodi, K., Altani, A., & Georgiou, G. K. (2018). Word reading fluency as a serial naming task. *Scientific Studies of Reading*, *22*(3), 248–263. https://doi.org/10.1080/10888438.2018.1430804

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://www.Rproject.org/

Rabbitt, P. M. (1966). Errors and error correction in choice-response tasks. *Journal of Experimental Psychology*, *71*(2), 264–272. https://doi.org/10.1037/h0022853

Rabin, L. A., Barr, W. B., & Burton, L. A. (2005). Assessment practices of clinical neuropsychologists in the United States and Canada: a survey of INS, NAN, and APA Division 40 members. *Archives of Clinical Neuropsychology*, *20*(1), 33–65. https://doi.org/10.1016/j.acn.2004.02.005

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin*, *124*(3), 372–422. https://doi.org/10.1037/0033-2909.124.3.372

Rayner, K., Fischer, M. H., & Pollatsek, A. (1998). Unspaced text interferes with both word identification and eye movement control. *Vision Research*, *38*(8), 1129–1144. https://doi.org/10.1016/s0042-6989(97)00274-5

Rayner, K., Juhasz, B. J., & Pollatsek, A. (2005). Eye movements during reading. In M. J. Snowling and C. Hulme (Eds.) *The Science of reading: A handbook* (pp. 79–97). Blackwell Publishing Ltd.

Rayner, K., Yang, J., Schuett, S., & Slattery, T. J. (2013). Eye movements of older and younger readers when reading unspaced text. *Experimental Psychology*, *60*(5), 354–361. https://doi.org/10.1027/1618-3169/a000207

Recarte, M. A., & Nunes, L. M. (2000). Effects of verbal and spatial-imagery tasks on eye fixations while driving. *Journal of Experimental Psychology: Applied, 6*(1), 31–43. https://doi.org/10.1037/1076-898X.6.1.31

Recarte, M., Pérez, E., Conchillo, Á, & Nunes, L. (2008). Mental workload and visual impairment: Differences between pupil, blink, and subjective rating. *The Spanish Journal of Psychology, 11*(2), 374-385. doi:10.1017/S1138741600004406

Rey-Mermet, A., Gade, M., & Oberauer, K. (2018). Should we stop thinking about inhibition? Searching for individual and age differences in inhibition ability. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 44*(4), 501–526. https://doi.org/10.1037/xlm0000450

Rey-Mermet, A., Gade, M., Souza, A. S., von Bastian, C. C., & Oberauer, K. (2019). Is executive control related to working memory capacity and fluid intelligence? *Journal of Experimental Psychology: General, 148*(8), 1335–1372. https://doi.org/10.1037/xge0000593

Roelofs, A. (2007). Attention and gaze control in picture naming, word reading, and word categorizing. *Journal of Memory and Language*, *57*(2), 232–251. https://doi.org/10.1016/j.jml.2006.10.001

Saastamoinen, M., & Järvelin, K. (2018). Relationships between work task types, complexity and dwell time of information resources. *Journal of Information Science*, *44*(2), 265–284. https://doi.org/10.1177/0165551516687726

Salo, R., Henik, A., & Robertson, L. C. (2001). Interpreting Stroop interference: An analysis of differences between task versions. *Neuropsychology*, *15*(4), 462–471. https://doi.org/10.1037/0894-4105.15.4.462

Scarpina, F., & Tagini, S. (2017). The Stroop color and word test. *Frontiers in Psychology*, *8*, 557. https://doi.org/10.3389/fpsyg.2017.00557

Schuch, S., Dignath, D., Steinhauser, M., & Janczyk, M. (2019). Monitoring and control in multitasking. *Psychonomic Bulletin & Review*, *26*(1), 222–240. https://doi.org/10.3758/s13423-018-1512-z

Seo, H., & Lee, C. (2002). Head-free reading of horizontally and vertically arranged texts. *Vision Research*, *42*(10), 1325–1337. https://doi.org/10.1016/s0042-6989(02)00063-9

Shenhav, A., Botvinick, M. M., & Cohen, J. D. (2013). The expected value of control: an integrative theory of anterior cingulate cortex function. *Neuron*, *79*, 210–240. https://doi.org/10.1016/j.neuron.2013.07.007

Sheridan, H., Rayner, K., & Reingold, E. M. (2013). Unsegmented text delays word identification: Evidence from a survival analysis of fixation durations. *Visual Cognition, 21*(1), 38–60. https://doi.org/10.1080/13506285.2013.767296

Shiga, N., & Ohkubo, Y. (1979). Pupillary responses to auditory stimuli—A study about the change of the pattern of the pupillary reflex dilation. *Tohoku Psychologica Folia*, *38*, 57-65.

Silva, S., Reis, A., Casaca, L., Petersson, K. M., & Faísca, L. (2016). When the eyes no longer lead: Familiarity and length effects on eye-voice span. *Frontiers in Psychology*, *7*, 1720. https://doi.org/10.3389/fpsyg.2016.01720

Slattery, T. J., & Parker, A. J. (2019). Return sweeps in reading: Processing implications of undersweep-fixations. *Psychonomic Bulletin & Review*, *26*(6), 1948–1957. https://doi.org/10.3758/s13423-019-01636-3

Snell, J., & Grainger, J. (2018). Parallel word processing in the flanker paradigm has a rightward bias. *Attention, Perception, & Psychophysics*, *80*(6), 1512–1519. https://doi.org/10.3758/s13414-018-1547-2

Snell, J., Cauchi, C., Grainger, J., & Lété, B. (2021). Attention extends beyond single words in beginning readers. *Attention, Perception, & Psychophysics*, *83*(1), 238–246. https://doi.org/10.3758/s13414-020-02184-y

Snell, J., Mathôt, S., Mirault, J., & Grainger, J. (2018). Parallel graded attention in reading: A pupillometric study. *Scientific Reports*, *8*(1), 3743. https://doi.org/10.1038/s41598-018-22138-7

Speech Analyzer (Version 3.0.1) [Computer Software]. Dallas: SIL International.

SR Research. (2016). *SR Research Experiment Builder* (Version 2.2.1). [Computer software]. Mississauga, Ontario, Canada: SR Research Ltd.

SR Research. (2021). *Eyelink Data Viewer* (Version 4.2.1). [Computer software]. Mississauga, Ontario, Canada: SR Research Ltd.

Strauss, E., Sherman, E. M. S., & Spreen, O. (2006). *A Compendium of neuropsychological tests: Administration, norms, and commentary* (3rd ed.). New York: Oxford University Press.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology, 18*(6), 643–662. https://doi.org/10.1037/h0054651

Tapper, A., Gonzalez, D., Nouredanesh, M., & Niechwiej-Szwedo, E. (2021). Pupillometry provides a psychophysiological index of arousal level and cognitive effort during the performance of a visual-auditory dual-task in individuals with a history of concussion. *Vision Research*, *184*, 43–51. https://doi.org/10.1016/j.visres.2021.03.011

Vakil, E., Mass, M., & Schiff, R. (2019). Eye Movement performance on the Stroop test in adults with ADHD. *Journal of Attention Disorders*, *23(10)*, 1160–1169. https://doi.org/10.1177/1087054716642904

van der Laan, L. N., Hooge, I. T. C., de Ridder, D. T. D., Viergever, M. A., & Smeets, P. A. M. (2015). Do you like what you see? The role of first fixation and total fixation duration in consumer choice. *Food Quality and Preference, 39,* 46–55. https://doi.org/10.1016/j.foodqual.2014.06.015

van der Wel, P., & van Steenbergen, H. (2018). Pupil dilation as an index of effort in cognitive control tasks: A review. *Psychonomic Bulletin & Review*, *25(6)*, 2005–2015. https://doi.org/10.3758/s13423-018-1432-y

White, R. W. & Kelly, D. (2006). A study on the effects of personalization and task information on implicit feedback performance. In Proceedings of the 15th ACM international conference on Information and knowledge management (pp. 297-306). Arlington, Virginia, USA.

Wickens, C. D. (2002). Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159–177. https://doi.org/10.1080/14639220210123806

Winn, M.B., Wendt, D., Koelewijn, T., Kuchinsky, S.E. (2018). Best practices and advice for using pupillometry to measure listening effort: An introduction for those who want to get started. *Trends in Hearing*, *22*, 233121651880086.. doi:10.1177/2331216518800869

Wu, Y. J., Yang, W. H., Wang, Q. X., Yang, S., Hu, X. Y., Jing, J., & Li, X. H. (2018). Eye-movement patterns of Chinese children with developmental dyslexia during the Stroop Test. *Biomedical and Environmental Sciences: BES*, *31*(9), 677–685. https://doi.org/10.3967/bes2018.092

Yagle, K., Richards, T., Askren, K., Mestre, Z., Beers, S., Abbott, R., Nagy, W., Boord, P., & Berninger, V. (2017). Relationships between eye movements during sentence reading comprehension, word spelling and reading, and DTI and fmri connectivity in students with and without dysgraphia or dyslexia. *Journal of Systems and Integrative Neuroscience*, *3*(1), 10.15761/JSIN.1000150. https://doi.org/10.15761/JSIN.1000150

Zekveld, A. A., & Kramer, S. E. (2014). Cognitive processing load across a wide range of listening conditions: insights from pupillometry. *Psychophysiology*, *51*(3), 277–284. https://doi.org/10.1111/psyp.12151

Zekveld, A. A., Kramer, S. E., & Festen, J. M. (2011). Cognitive load during speech perception in noise: the influence of age, hearing loss, and cognition on the pupil response. *Ear and Hearing*, *32*(4), 498–510. https://doi.org/10.1097/AUD.0b013e31820512bb

Ziaka, L., & Protopapas, A. (2022). Conflict monitoring or multi-tasking? Tracking within-task performance in single-item and multi-item Stroop tasks. *Acta Psychologica*, *226*, 103583. https://doi.org/10.1016/j.actpsy.2022.103583

Ziaka, L., Skoteinou, D., & Protopapas, A. (2022). Task format modulates the relationship between reading ability and Stroop interference. *Journal of*

Experimental Psychology: Human Perception and Performance, *48(4)*, 275–288. https://doi.org/10.1037/xhp0000964

Zimmerman, M. E. (2011). Speed-Accuracy Tradeoff. In J. S. Kreutzer, J. DeLuca, & B. Caplan (Eds.), *Encyclopedia of clinical neuropsychology* (p. 2344). New York, NY: Springer. https://doi.org/10.1007/978-0-387-79948-3_1247