Article:
Strzelczyk, D., Clayson, P.E., Sigurdardottir, H.M. et al. (2026) Contralateral delay activity as a marker of visual working memory capacity: a multi-site registered replication. Cortex. ISSN: 0010-9452

https://doi.org/10.1016/j.cortex.2026.04.006

Journal Pre-proof

Contralateral delay activity as a marker of visual working memory capacity: a multi-site registered replication

Dawid Strzelczyk, Peter E. Clayson, Heida Maria Sigurdardottir, Faisal Mushtaq, Yuri G. Pavlov, Hélène Devillez, Anton Lukashevich, Harold A. Rocha, Yong Hoon Chung, Kevin M. Ortego, Viola S. Störmer, José C. García Alanis, Christoph Löffler, Anna-Lena Schubert, Anna Lena Biel, Samuel A. Birkholz, Emily M. Johnson, Jeffrey S. Johnson, Zitong Lu, Yong Min Choi, Eva Lout, Julie D. Golomb, Shuangke Jiang, Myles Jones, Eda Mizrak, Claudia C. von Bastian, Niko A. Busch, Charline Peylo, Larissa Behnke, Yannik Hilla, Maro G. Machizawa, William X.Q. Ngiam, Edward K. Vogel, Nicolas Langer

PII: S0010-9452(26)00113-9

DOI: https://doi.org/10.1016/j.cortex.2026.04.006

Reference: CORTEX 4340

To appear in: *Cortex*

Received Date: 24 February 2026

Revised Date: 8 April 2026

Accepted Date: 8 April 2026

Please cite this article as: Strzelczyk D, Clayson PE, Sigurdardottir HM, Mushtaq F, Pavlov YG, Devillez H, Lukashevich A, Rocha HA, Chung YH, Ortego KM, Störmer VS, García Alanis JC, Löffler C, Schubert A-L, Biel AL, Birkholz SA, Johnson EM, Johnson JS, Lu Z, Choi YM, Lout E, Golomb JD, Jiang S, Jones M, Mizrak E, von Bastian CC, Busch NA, Peylo C, Behnke L, Hilla Y, Machizawa MG, Ngiam WXQ, Vogel EK, Langer N, Contralateral delay activity as a marker of visual working memory capacity: a multi-site registered replication, *Cortex*, https://doi.org/10.1016/j.cortex.2026.04.006.

Contralateral delay activity as a marker of visual working memory capacity: a multi-site registered replication

Dawid Strzelczyk[1,2], Peter E. Clayson[3], Heida Maria Sigurdardottir[4], Faisal Mushtaq[5], Yuri G. Pavlov[6], Hélène Devillez[4], Anton Lukashevich[4], Harold A. Rocha[3], Yong Hoon Chung[7], Kevin M. Ortego[7], Viola S. Störmer[7], José C. García Alanis[8], Christoph Löffler[8], Anna-Lena Schubert[8], Anna Lena Biel[9], Samuel A. Birkholz[10], Emily M. Johnson[10], Jeffrey S. Johnson[10], Zitong Lu[11], Yong Min Choi[11], Eva Lout[11], Julie D. Golomb[11], Shuangke Jiang[12,20], Myles Jones[12], Eda Mizrak[12,14], Claudia C. von Bastian[12], Niko A. Busch[9], Charline Peylo[13], Larissa Behnke[13], Yannik Hilla[13], Maro G. Machizawa[15], William X. Q. Ngiam[16,17], Edward K. Vogel[16,17], Nicolas Langer[1,2,18*]

[1] Methods of Plasticity Research, Department of Psychology, University of Zurich, Zurich, Switzerland

[2] Neuroscience Center Zurich (ZNZ), University of Zurich and ETH Zurich, Zurich, Switzerland

[3] University of South Florida, Tampa, Florida, USA

[4] University of Iceland, Reykjavik, Iceland

[5] University of Leeds, Leeds, UK

[6] University of Tuebingen, Tuebingen, Germany

[7] Dartmouth College, Hanover, New Hampshire, USA

[8] University of Mainz, Germany

[9] Institute of Psychology, University of Münster, Germany

[10] Department of Psychology, North Dakota State University, Fargo, ND, USA

[11] The Ohio State University, Columbus, Ohio, USA

[12] University of Sheffield, Sheffield, UK

[13] Neuropsychology and Cognitive Neuroscience, Department of Psychology, University of Zurich, Zurich, Switzerland

[14] University of Oxford, Oxford, UK

[15] Digital Brain Science Laboratory, Xiberlinc Inc., Tokyo, Japan

[16] Department of Psychology, University of Chicago, Illinois, USA

[17] Institute of Mind and Biology, University of Chicago, Illinois, USA

[18] Center of Reproducible Science, University of Zurich, Zurich, Switzerland

[19] Cognitive Psychology, Department of Psychology, University of Zurich, Zurich, Switzerland

* n.langer@psychologie.uzh.ch

Abstract

The contralateral delay activity (CDA) is a widely used electrophysiological marker of visual working memory (VWM), yet recent work has questioned whether typical sample sizes in CDA studies are sufficient to robustly detect set size effects and brain-behavior correlations. As part of the #EEGManyLabs initiative, the present multi-site replication study aimed to rigorously test replicability of the key findings of Vogel and Machizawa (2004) using a large sample of 304 participants across 10 laboratories and a preregistered analysis plan. We replicated the expected contralateral-ipsilateral asymmetry and observed increases in CDA amplitude from set size 2 to 4 and from set size 2 to 6. In contrast, the hypothesized positive correlation between the CDA increase from set size 2 to 4 and individual VWM capacity was not replicated in the preregistered meta-analytic correlation. Across different pipelines and statistical analyses, the meta-analytic correlation estimate was small (r = 0.15) and substantially attenuated relative to the original effect size in Vogel and Machizawa (2004) study (r = 0.78). To contextualize these findings, we applied a funnel-plot diagnostic combining published effects with the #EEGManyLabs data, indicating small-study inflation and publication bias. Taken together, our results indicate that reports of strong correlations between CDA amplitude and VWM capacity may have been overestimated, in part because statistically significant findings were selectively reported. Our results highlight the importance of open science practices, including well-powered, preregistered studies with transparent data and analysis pipelines, in order to characterize the magnitude and robustness of individual-difference associations in psychophysiology.

Introduction

Visual working memory (VWM) is a temporary storage system that holds information that can be accessed and manipulated by higher cognitive functions (Luck & Vogel, 2013). Visual working memory is considered a central construct in cognitive neuroscience and is a putative intermediary for information transfer (Atkinson & Shiffrin, 1968; A. Baddeley, 2003; A. D. Baddeley, 1986; A. D. Baddeley & Hitch, 1974; Cowan et al., 2005; Cowan & Morey, 2006; Liesefeld & Müller, 2019; Olivers, 2008), thereby facilitating various cognitive functions including reading comprehension (Caplan & Waters, 1999; Daneman & Carpenter, 1980; Lotfi et al., 2022; Martin & Romani, 1994; Wang et al., 2022), planning and problem-solving (Cowan et al., 2005; Miyake & Shah, 1999; Naveh-Benjamin & Cowan, 2023), and learning new skills (A. Baddeley et al., 1988; A. D. Baddeley et al., 2017; Cowan, 2014; Gathercole & Baddeley, 1989; Jongbloed-Pereboom et al., 2019; von Bastian et al., 2022). Researchers have been using electroencephalography (EEG) to understand the neural correlates of VWM in real-time. A commonly used EEG measure of VWM is an event-related potential (ERP) called the contralateral delay activity (CDA). This EEG signal has also been referred to in other studies as Contralateral Negative Slow Wave (CNSW) by Klaver et al. (1999), Sustained Posterior Contralateral Negativity (SPCN) by Brisson and Jolicoeur (2007), Perron et al. (2009), and Contralateral Search Activity (CSA) by Emrich et al. (2009). These different terms all refer to the same visual working memory correlate. Hence, we will maintain the use of the term CDA throughout the remainder of the paper.

The CDA is a difference wave constructed by subtracting ipsilateral from contralateral activity related to the to-be-remembered items. Since items in studies analyzing experimental effects on the CDA are generally shown bilaterally, while only those on one side of the screen are supposed to be memorized, the idea of the subtraction is to eliminate any activity related to early perceptual and low-level processing by assuming that they equally affect ipsilateral and contralateral ERPs. Activity over the contralateral hemisphere tends to be more negative than ipsilateral activity during VWM retention (Luria et al., 2016; Ngiam et al., 2021; Vogel & Machizawa, 2004). Thus, it has been suggested that the CDA reflects the neural activity related to the maintenance of information in VWM, and studies have shown that the amplitude and duration of the CDA are linked to the amount of information stored in working memory.

In a seminal paper from Vogel and Machizawa (2004), the authors demonstrated that the CDA amplitude increases with the number of items stored in VWM and plateaus at around 3 to 4 items, consistent with the typical adult working memory capacity (Forsberg et al., 2023). More

importantly, this study showed that the increase in the CDA amplitude with greater memory load correlated with individual VWM performance (Vogel & Machizawa, 2004). Specifically, individuals with high VWM capacity exhibited a larger increase in the CDA when attempting to memorize 4 compared to 2 items. In this study, the CDA was elicited using a color change detection task (Vogel & Machizawa, 2004). The task involves presenting participants with a central arrow cue that indicates whether participants need to memorize items on the left or right of the screen center. The cue is followed by a bilateral stimulus array with equal numbers of colored squares shown on each side (set size 1 to 10). After a short retention phase, participants are presented with a second array and asked to indicate whether any of the squares on the cued side changed color (Figure 2). The lateralized color change detection task is now a widely used paradigm to examine visual working memory processes (Feldmann-Wüstefeld, 2021; Luria et al., 2016) and has been explored in several variations such as different set sizes, including distractions, retro-cueing and using different shapes and colors (Feldmann-Wüstefeld, 2021; Feuerstahler et al., 2019; Roy & Faubert, 2023; Schneider et al., 2017).

The finding that the CDA amplitude is sensitive to how much visual information is to be remembered has been replicated in numerous studies (Asp et al., 2021; Brady et al., 2016; Hakim et al., 2019; Heuer & Schubö, 2016; Quirk et al., 2020; Unsworth et al., 2015). Furthermore, several studies have validated the positive correlation between the CDA amplitude increase and VWM capacity (Adam et al., 2018; Feldmann-Wüstefeld, 2021; Villena-González et al., 2020). In the review paper of Luria et al. (2016), the authors conducted a meta-analysis from 11 previous studies and reported an aggregated correlation of $r = 0.596$. However, a recent study indicated that the typical numbers of subjects and trials for CDA experiments seen in the literature may be underpowered for detecting set size differences (Ngiam et al., 2021).

The insufficient power issue is even more pressing for the correlation between the VWM capacity and the CDA amplitude increase. Critically, Schönbrodt and Perugini (2013) demonstrated that correlation estimates typically stabilize at a sample size of approximately 250 subjects (Schönbrodt & Perugini, 2013). Except for one large study ($N = 171$; Unsworth et al., 2015), the average sample size of previous studies investigating the relationship between VWM capacity and the CDA amplitude was 32 subjects (range 12-83 subjects for 12 studies; Luria et al, 2016). Finally, the inherent flexibility in EEG analysis, including analysis of the CDA, leaves many decisions up to the researcher. This leaves open the possibility to exploit these researchers' degrees of freedom (i.e., the garden of forking paths; Gelman & Loken, 2013), either intentionally or unintentionally. Such practices can lead to erroneous

inferences and perpetuate replication problems in cognitive neuroscience (Clayson et al., 2019; Luck & Gaspelin, 2017).

To address this issue, the #EEGManyLabs project was initiated (Pavlov et al., 2021). The #EEGManyLabs initiative highlights the importance of replication in science and the need for rigorous research methods to increase confidence in prominent effects. The #EEGManyLabs project aims to replicate pivotal EEG studies which had a critical impact on the cognitive and affective neuroscience community. Importantly, the #EEGManyLabs project is designed to address some of the limitations of previous replication efforts by using a large sample of participants, standardized procedures, and a pre-registered analysis plan (i.e., Registered Report; Pavlov et al., 2021).

As part of the #EEGManyLabs project, the current study aimed to contribute to the existing literature on VWM and the CDA by conducting a robust multi-site, large-scale replication of Vogel and Machizawa's (2004) seminal study. The present study was chosen for replication by a global consortium of EEG specialists owing to its scientific significance (for further information on the selection process, refer to Pavlov et al., 2021). In accordance with the #EEGManyLabs project, this Registered Report closely adhered to the original study design and ensured adequate statistical power with a large sample size. The present study also followed preregistered analysis steps to ensure the integrity of the direct replication and statistical inferences (Paul et al., 2021). To this end, Experiment 3 of the original study was replicated using three set sizes (2, 4, and 6 items per side) to examine whether CDA amplitude varies as a function of memory load. In line with the original study, the following hypotheses were tested:

[H1.1] The CDA amplitude increases from arrays of 2 items per side to arrays of 4 items per side.
[H1.2] The CDA amplitude increases from arrays of 2 items per side to arrays of 6 items per side.
[H1.3] The CDA amplitude for 4 items and 6 items is equivalent.

Additionally, the study examined whether the CDA amplitude is related to performance on the change detection task:

[H2.1] Subjects' VWM capacity (measured behaviorally) is positively correlated with the CDA amplitude increase from 2 to 4 items.

[H2.2] Subjects' VWM capacity (measured behaviorally) is not correlated with the CDA amplitude increase from 4 to 6 items.

Finally, replication success was evaluated separately for each hypothesis. For hypotheses predicting directional effects, replication success was defined as a statistically significant random-effects meta-analytic estimate in the same direction as in the original study, combining results across laboratories. For hypotheses predicting equivalence or the absence of an association, replication success was evaluated using equivalence testing.

2. Methods

The protocol for this replication was developed in consultation with the original authors (co-authors of the present work, EV, MM). The current document is a Stage 2 Registered Report that follows guidelines for open science in psychophysiological research as outlined by Garrett-Ruffin et al. (2021). The Stage-1 Registered Report outlining the preprocessing and analysis steps is available at: https://doi.org/10.31234/osf.io/shdea (Strzelczyk et al., 2023). All raw EEG and behavioral datasets, after marker harmonization, anonymization, and including full datasets that were later excluded from analysis, are openly accessible at: https://gin.g-node.org/EEGManyLabs/EEGManyLabs_Replication_VogelMachizawa2004_Raw and https://gin.g-node.org/EEGManyLabs/EEGManyLabs_Replication_VogelMachizawa2004_Processed. All preprocessing and analysis scripts used in the present report are available on GitHub: https://github.com/ksgfan/EEGManyLabs. Each site has obtained approval from the local ethics committee to conduct the study and share data.

The institution abbreviations used throughout the manuscript are: DART (Dartmouth College, USA), USF (University of South Florida, USA), JGU (Johannes Gutenberg University Mainz, Germany), WWU (University of Münster, Germany), NDSU (North Dakota State University, USA), OSU (The Ohio State University, USA), UI (University of Iceland, Iceland), TUOS (University of Sheffield, UK), UZH-MPR (University of Zurich, Methods of Plasticity Research, Switzerland), and UZH-NCN (University of Zurich, Neuropsychology and Cognitive Neuroscience, Switzerland).

Known differences from the original study

Table 1. Details on original, replication and alternative pipelines. Deviations from the original preprocessing are highlighted in blue in the direct replication pipeline.

Offline Processing Step	**Original and current study parameters for the direct preprocessing pipeline**	**#EEGManyLabs: advanced preprocessing pipeline**
Offline Filter	(1) Hardware online filter: bandpass of 0.01-80 Hz (half-power cutoff, Butterworth filters) (2) 35 Hz LP only for plots	(1) Offline bandpass of 0.01-80 Hz (half-power cutoff, Butterworth filters) (2) 35 Hz LP for plotting
Line noise removal	no line noise removal ZapLine method[1]	ZapLine method
Ocular artifact rejection	(1) Trials containing ocular artifacts were removed (i.e., blinks or eye	If eye-tracker recording is available, we excluded trials with eye movements larger

	movements larger than 1 degree). A heuristic for 1 visual degree was used (25 microvolt bipolar HEOG amplitude threshold; adjusting the threshold for each subject based on visual inspection). A calibration paradigm was used to estimate the subject specific amplitude representing 1 visual degree (2) Blinks: unipolar VEOG >? microvolt Blinks: unipolar VEOG >90 microvolt	than 1 degree. If no eye-tracker is available, we identified ocular artifacts using a tailored subject-specific amplitude threshold for the EOG electrodes, which was obtained from the saccadic calibration task.
Artifact Rejection	(1) peak-to-peak amplitude >200 microvolt (2) Visual inspection was used to identify and exclude trials containing movement artifacts or blocking	(1) Peak-to-peak amplitude >200 microvolt (2) bad trial identification method introduced by Adam, Robison, and Vogel (2018)
Bad Channel Identification	(1) peak-to-peak amplitude >75 microvolt. Bad channels were not interpolated, but artifactual trials were rejected (2) Visual inspection	(1) Correlation below 0.85 with neighboring channels (2) 4 SD or more line noise relative to signal than all other channels (3) Blocking longer than 5 s
Bad Channel Interpolation	NA	Spherical spline interpolation
Reference	Algebraic average of the left and right mastoids	Algebraic average of the left and right mastoids
CDA time interval	300 - 900 ms after memory onset	300 - 900 ms after memory onset
Baseline Interval	-200 - 0 ms	-200 - 0 ms
Region of Interest	left electrode cluster: P3, T5/P7, O1. right electrode cluster: P4, T6/P8, O2	left electrode cluster: P3, T5/P7, O1. right electrode cluster: P4, T6/P8, O2
CDA time interval	retention phase (i.e., 300 - 900 ms after the onset of memory array)	retention phase (i.e., 300 - 900 ms after the onset of memory array)
Set Size	Experiment #3: 2,4,6	2, 4, 6
Visual Memory Capacity	K & d'	K & d'

Note. Deviations from the original study in the direct preprocessing pipeline are shown in blue.

[1] Several labs are recording the task with an eye-tracker, which induces line noise. Therefore, we decided to use ZapLine to reduce the line noise.

Sample size & Inclusion criteria

Participants were recruited from universities or nearby communities. The study only included individuals between 18 and 35 years free from any diagnosed psychiatric or neurological disorders and with intact color vision. We acquired demographics (i.e., age, gender), handedness (Edinburgh Handedness Inventory; Oldfield, 1971) and education level based on International Standard Classification of Education (ISCED; http://uis.unesco.org/en/topic/international-standard-classification-education-isced).

The required sample size was estimated for each hypothesis: For hypothesis #1, we used the CDA power calculator (https://williamngiam.shinyapps.io/CDAPower/; Ngiam et al., 2021) to estimate the required sample size to detect a set-size effect between set size 2 and 4 (which is similar as between set size 2 and 6). With a minimum of 170 clean trials per condition (i.e., excluding subjects with a bad trial rate > 30%) and 90% power, the estimated number of subjects required is 70 (see Figure 1).

The following procedure was conducted to estimate the required sample size to investigate the correlation between VWM capacity and the CDA amplitude difference (i.e., hypothesis #2): In the original study, 36 participants were recruited and the subjects' VWM capacity was correlated with the CDA amplitude increase between 2 and 4 items with a correlation estimate of r = .78 (Vogel & Machizawa, 2004). The power analysis showed that with an alpha level of .02 and an assumed effect size of 50% (i.e., r = .39) of the original study, a sample size of N = 68 is required to achieve 90% power in detecting the effect. For the sample size calculation of hypothesis #2, we used the R package "pwr" (Champely, 2020) (pwr.r.test(r = 0.39, sig.level = 0.02, power = 0.9, alternative = "greater")). However, according to Schönbrodt and Perugini (2013), correlation starts to stabilize at the sample size N = 250. As the #EEGManyLabs is open for any lab to participate, we decided that each participating lab (i.e., N = 10) should recruit 25 participants, resulting in 250 participants in total, which provides sufficient power to investigate both hypotheses.

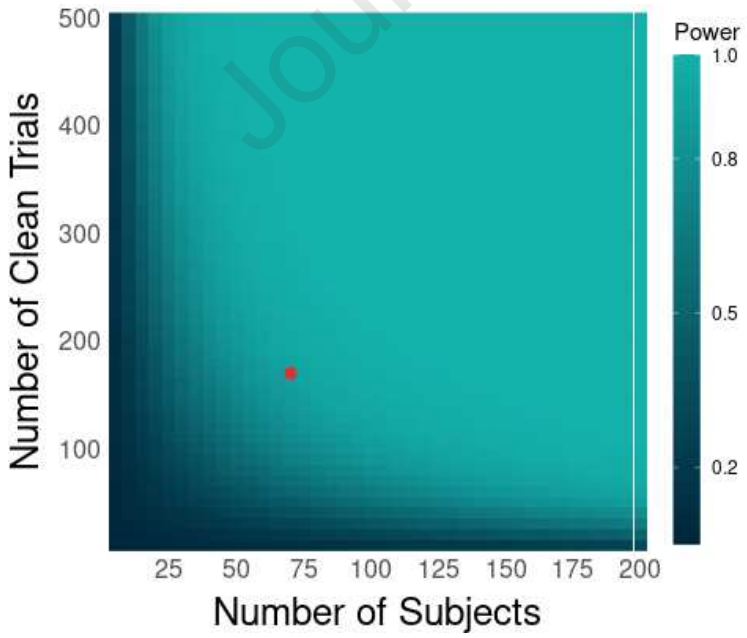

Figure 1. CDA power calculation. We estimated the required sample size for the set size effect between set size 2 and 4, assuming at least 170 trials (i.e., maximum of 30% bad trials) and 90% power. The estimated number of subjects required is 70 (red dot).

Exclusion criteria

The color change detection task requires the participants to discriminate between colored squares, therefore color blindness is a critical exclusion criterion. We tested the color-vision with an online color-vision test (https://colormax.org/color-blind-test/). Participants with scores below 11 out of 12 correct responses were considered to have a color-vision deficiency and were excluded from further analyses.

Exclusion criteria for direct replication

Following the original study, we excluded trials with eye movements, blinks, and blocking (amplifier saturation after drift). To identify eye movements and blinks, horizontal electrooculography (EOG) was concurrently recorded. Contaminated trials were identified by large (>1°) eye movements (Vogel & Machizawa, 2004).

In the original study, the authors used a heuristic for 1° horizontal eye movements and a fixed amplitude threshold for each subject. In this replication study, we deviated from this procedure and calculated the 1° horizontal eye movement amplitude for each participant in order to more accurately estimate an individual's amplitude threshold. To determine the individual participant's exclusion amplitude threshold, which reflects 1° horizontal eye movement, there was a separate horizontal EOG saccade calibration task prior to the main experiment (developed by K.M.Ortego, co-author). This task involves participants making saccades to left and right targets on the screen. Participants started each trial by fixating on the center of the screen. Following a key press there was a jittered interval between 1200~1600 ms and a saccade target (a red disk; 0.6° in size) appeared either 3° or 6° away from the fixation on the left or right side of the screen along the horizontal midline. Participants were instructed to make a saccade to the target location as soon as it appeared and to press a space bar once they had successfully made the saccade. There were 15 trials per condition, resulting in 60 trials total. The data from the saccade calibration paradigm was preprocessed by (1) bandpass filtering the data from 0.1-40 Hz; (2) epoching from -200 to +600 ms with respect to the onset of the saccade target; and (3) baseline correcting using a pre-stimulus baseline interval of -200 to 0 ms. Given previous research showing the saccade onset latency being ~200 ms (Westheimer, 1954a, 1954b), horizontal EOG (i.e., HEOG = HEOGR - HEOGL) channel amplitudes from horizontal saccades were averaged during the 300~400 ms interval across the left and right conditions. The 1° horizontal eye movement amplitude threshold was then

calculated by extrapolating from 3° and 6° eye movements (estimating the linear regression curve using fittype function in MATLAB) as previous reports have shown the HEOG amplitudes and the size of saccades have a consistently linear relationship (Luck, 2014). We did not measure the 1° eye movement directly, as a pilot study demonstrated that estimating the 1° eye movement is more error prone and has too much variability. Furthermore, blinks were detected by using an amplitude threshold (>50 microvolt) in the unipolar VEOG channel. In addition, a segment was marked as bad if any electrodes of interest (i.e., HEOG, P3, T5/P7, O1; P4, T6/P8, O2) had a peak-to-peak amplitude >200 microvolt within one segment. Finally, visual inspection was used to identify bad trials. For an overview of the exclusion criteria and analysis pipeline see the Table 1. If more than 30% of trials (all set-size conditions combined) had to be rejected by these combined criteria, the subjects were excluded from further analysis to assure sufficient number of trials (see sample size calculation).

Alternative analysis pipelines

For the alternative pipelines (Table 1), to identify eye movements and blinks, all labs recorded horizontal and vertical EOG and several labs additionally recorded eye tracking data (see Table 2). Trials contaminated by eye movements larger than 1° (Vogel & Machizawa, 2004) were identified based on eye tracking data if available (i.e., trials containing eye movements larger than 1°), and otherwise based on EOG data as described in the previous section. In addition to the bad trial identification methods described in the direct replication, we also utilized the bad trial identification method introduced by Adam et al. (2018; see below). Again, if more than 30% of trials for a specific set size had to be rejected by these combined criteria, the subjects were excluded from further analysis to assure sufficient number of trials (see sample size calculation).

Procedure

Upon their arrival, participants received a brief overview of the experiment and were asked to give their informed written consent for participating in the study and allowing data sharing. Next, the participants were asked to fill out a short questionnaire regarding their history of psychiatric and neurological disorders, handedness and educational level, and carry out an online color-blind test (https://colormax.org/color-blind-test/). Subsequently, the participants were comfortably seated in a chair. If available, the experiment was conducted in a sound- and electrically shielded Faraday recording cage. Some cages were equipped with a chinrest to minimize head movements. A cap with electrodes was placed on the participant's head and

impedances were checked if provided by the EEG amplifier system and improved if necessary (see Table 2 for details). As this project is part of a wider initiative on replicability in EEG (#EEGManyLabs), several of the laboratories in this replication also collected resting state EEG data together with some personality measures (https://osf.io/sp3ck/, Pavlov et al., 2021). Neither resting EEG nor personality data were analyzed in the current study but were merged across sites as part of a future replication project to be reported elsewhere. Participants first completed an EOG saccade calibration task, after which the color change detection task began. The expected duration of the entire experiment was approximately 120 minutes. Upon completion of the examination, participants received compensation or credit for their participation.

Experimental Paradigm

The color change detection task was identical to the task used in the original study (Vogel & Machizawa, 2004). The paradigm was implemented in MATLAB, using the PsychToolbox extensions (Brainard, 1997; Pelli, 1997). Each trial of the task started with a blank screen presented for 1500 ms. Then, a central arrow appeared for 200 ms, indicating which side of the screen the participant should pay attention to. This was followed by another fixation period of a random time interval between 300 ms and 400 ms. Afterwards, a memory set was presented for 100 ms, which consisted of either 2, 4, or 6 colored squares on each side of the screen. Participants were instructed to only memorize the part of the memory set indicated by the arrow. This was followed by a 900 ms retention interval with a blank screen and a fixation cross (see Figure 2). Finally, a test array was presented for 2000 ms, and the participants were asked to indicate whether the test array was identical to the previous memory array ("no-change" trial) or whether the test array was different by one color ("change" trial).

The participants indicated whether a change occurred by pressing either the A or L button on a keyboard. The button they pressed depended on the instruction they were given, with half of the participants being instructed to press the A button for a change and the L button for no change, while the other half was instructed to do the opposite. Additionally, the participants were instructed to use their left hand to press the A button and their right hand to press the L button. During the task, participants were asked to focus their gaze on the fixation cross in the center of the screen until the probe appeared.

All stimuli were displayed within 2 regions that were 4° x 7.3° in size and were located 3° to the left and right of a central fixation cross on a gray background (8.2 cd m^{-2}). Each memory

array consisted of 2, 4 or 6 colored squares (0.65° x 0.65°) in each visual field. The squares were chosen at random from a set of seven highly distinct colors (red, blue, violet, green, yellow, black and white), and a specific color appeared no more than twice in a single array. In other words, a specific color could be displayed in both hemifields but never twice within a hemifield. The positions of the stimuli were randomized on each trial, with the restriction that the distance between squares within a visual field was at least 2° (center to center). In 50% of the trials, the color of one square in the test array on the cued side was different from the corresponding square in the memory array, while in the remaining trials, the colors of the 2 arrays were identical.

The task was divided into five blocks, each containing 144 trials (i.e., 720 trials per subject and 240 trials per condition and subject). The cue direction (left or right) and set size (2, 4 or 6 items on each side of the screen) were randomly varied throughout the trials to ensure a balanced distribution of all conditions in each block. In line with the original study, no training exercise was conducted prior to the main task.

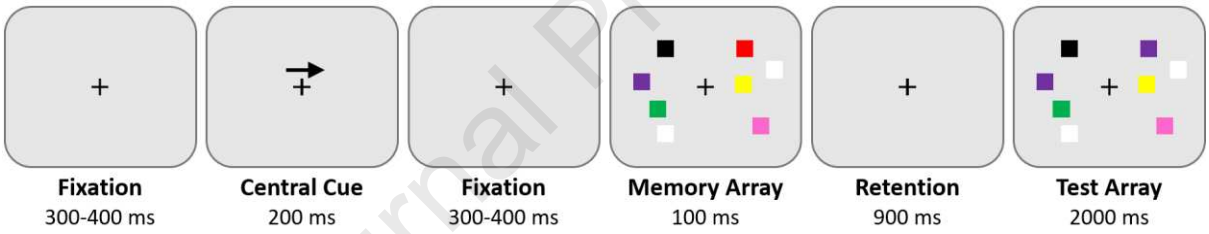

Figure 2. Lateralized color change detection task. The figure is illustrative and not to scale.

Neurophysiological Data Acquisition

The replicating labs used a range of EEG systems and, where available, eye trackers. Acquisition details are provided in Table 2. All labs provided the raw data to Zurich's Lab (UZH-MPR), where it was preprocessed and analyzed.

Table 2. Data acquisition settings for each laboratory.

Lab	Screen type; size; ratio; refresh rate	Stimulus presentation language	Distance between chinrest and monitor	EEG system; number of channels; sampling rate	Reference; grounding	Impedances	Eye tracker; sampling rate	HEOG	Faraday cage	Soundproof or sound attenuated recording room
Dartmouth College	VPixx; 540x300 mm; 1090 x 1080;120 Hz	Psychtoolbox 3.0.18	45 cm	BrainVision; 32 channels; 500 Hz	Right mastoid; Fpz	Kept below 10 KOhm	No	Yes	Yes	Yes
University of South Florida	Dell p2314h, 23" widescreen, 60 Hz	Psychtoolbox 3.0.18	65 cm	Magstim EGI, 128 channels, 500 Hz	Cz; PCz	Kept below 50 KOhm	No	Yes	No	No
University of	Eizo	Psychtoolbox	67 cm	BrainProducts	FCz; Fpz	Kept below 10	No	Yes	Yes	Yes

Mainz	ColorEdge CS2420; 24.1'' diag; 1920x1200; 60 Hz	3.0.18		; 64 channels; 1000 Hz		KOhm				
University of Münster	ViewpixxEEG 1920 x 1080 120Hz	Psychtoolbox 3.0.18	86	Biosemi; 64 + 3 chans; 1024Hz	Reference free; GND adjacent to POz	Not available with Biosemi	Eye Link, 500 Hz	Yes	no	Yes
North Dakota State University	ASUS ROG Strix XG27AQ 27"; 2560 x 1440;	Psychtoolbox 3.0.18	50 cm	Biosemi; 64 + 8 chans; 512 Hz	Reference free; GND adjacent to POz	Not available with Biosemi	Eye Link 1000, 500 Hz	Yes	Yes	Yes
The Ohio State University	BENQ XL2420-B; 1920 x 1080; 120 Hz	Psychtoolbox 3.0.18	80cm	Brain Vision; 32 channels; 1000Hz	Cz, Fpz	Kept below 20 KOhm	EyeLink 1000; 500 Hz	Yes	Yes	Yes
Icelandic Vision Lab, University of Iceland	2560*1440 60 Hz ASUS PG278QR 27"	Psychtoolbox 3.0.18	57cm (nasion to screen distance; *no chinrest for pilot)	BrainVision; 32 channels; 1000 Hz	Fz, Fpz	Kept below 15 KOhm	No	Yes	No	Yes
University of Sheffield	Iiyama G-master GB2488HSU; 531.4 x 298.9mm; 1920 x 1080; 144 Hz	Psychtoolbox 3.0.18	50cm (nasion to screen distance; *no chinrest for pilot)	Biosemi; 64 channel; Recorded at 2048 Hz and then downsampled to 512 Hz	Unreferenced/reference free; CMS/DRL adjacent to POz	Not Available (Only offset within ± 25 mV)	No	Yes	Yes	Yes
University of Zürich (UZH-MPR)	Philips 242E1; 540x414mm; 800x600; 100 Hz	Psychtoolbox 3.0.18	70 cm	ANT Neuro; 128 channels; 500 Hz	CPz; GND adjacent to M1	Kept below 20 KOhm	EyeLink 1000; 500 Hz	Yes	Yes	Yes
University of Zürich (UZH-NCN)	HP Omen 27q; 2560 x 1440; 144 Hz	Psychtoolbox 3.0.18	70 cm; no chin rest	BrainProducts; 64 channels; 1000 Hz	FCz; Fpz	Kept below 15 KOhm; ground and reference electrode below 5 KOhm	No	Yes	Yes	Yes

Artifact Removal and Data Preprocessing

All EEG data were imported into EEGLAB 2025.0.0 (Delorme & Makeig, 2004) and processed using 2 pipelines: a pipeline that follows the original study as closely as possible (see Vogel and Machizawa, 2004), and a recent pipeline optimized for current advances in the field of neuroscience.

Direct replication preprocessing pipeline

Following the Vogel and Machizawa (2004) study, we downsampled the data to 250 Hz and applied a bandpass filter of 0.01-80 Hz (half-power cutoff, Butterworth filters) using the EEGLAB function pop_eegfiltnew (Widmann & Schröger, 2012). Since certain labs measured eye movements using an eye tracker or did not have access to a faraday cage, line noise (50 Hz in Europe, 60 Hz in the US) was introduced as a result. To mitigate this, we used ZapLine Plus, which adaptively identifies and suppresses power-line components and their harmonics.

The algorithm is highly effective at removing power line artifacts while preserving non-artifactual parts of the signal (de Cheveigné, 2020; Klug & Kloosterman, 2022). This deviation from the original study was necessary to ensure accurate measurements. Afterwards, we re-referenced the data to an algebraic average of the left and right mastoids. We segmented the data from -200 to +1200 ms after the presentation of the memory array. The segments with saccadic eye movements (greater than 1° from the fixation cross) were excluded from further analysis using horizontal EOG channel response data from the saccade calibration task (for detailed information please refer to Exclusion Criteria). Furthermore, blinks were detected by using an amplitude threshold (>90 microvolt) in the unipolar VEOG channel. In addition, a segment was marked as bad if any electrodes of interest (i.e., HEOG, P3, T5/P7, O1; P4, T6/P8, O2) had a peak-to-peak amplitude >200 microvolt within one time window (i.e., bad channel criteria). Visual inspection was used to identify bad trials. Finally, a baseline correction was applied using a pre-stimulus interval of -200 to 0 ms.

Advanced preprocessing pipeline

In addition to following the original study's data preprocessing protocol, the data were also processed using recent advancements in neuroscience to assess the robustness of the results. First, error-prone channels were detected by the algorithms implemented in the EEGLAB plugin clean_rawdata (http://sccn.ucsd.edu/wiki/Plugin_list_process) without applying automated subspace removal (ASR). An electrode was defined as error-prone when recorded data from that electrode were correlated at less than 0.85 to an estimate based on neighboring electrodes. Furthermore, an electrode was defined as error-prone if it had more line noise (i.e., 50 Hz in Europe, 60 Hz in USA) relative to its signal than all other electrodes (4 standard deviations). Finally, if an electrode had a longer flat line than 5 s, it was considered error-prone. These error-prone electrodes were automatically removed and later interpolated using a spherical spline interpolation (EEGLAB function eeg_interp.m). Next, data was filtered using a bandpass filter of 0.01-80 Hz (half-power cutoff, Butterworth filters). Again, we used ZapLine Plus to remove line noise. Subsequently, the data was re-referenced to an algebraic average of the left and right mastoids and segmented from -200 to +1200 ms after presentation of the memory array. To determine whether a trial was artifactual, three criteria were applied: First, we excluded trials with blinks and large saccadic eye movements. If eye-tracker recording was available, we excluded trials with blinks and eye movements larger than 1° based on the eye-tracker. If no eye-tracker was available, we identified ocular artifacts using a tailored subject-specific amplitude threshold for the HEOG electrodes, which was obtained from the saccadic calibration task. Second, a sliding time window approach was adopted from

Adam et al. (2018). To identify trials containing blocking artifacts, a sliding time window of 200 ms was shifted across the segments without overlap. If any time window contained 60 ms of flat line activity in any channel (i.e., range of amplitudes <0.1 microvolt), the corresponding segment was marked as bad. Third, to identify trials containing large amplitude artifacts, non-overlapping sliding time windows of 12 ms were used. A segment was marked as bad if any electrode of interest (i.e., HEOG, P3, T5/P7, O1; P4, T6/P8, O2) had a peak-to-peak amplitude >200 microvolt within one time window. To foster scientific transparency and enable exact methodological replications and reproducibility, no visual inspection for bad trials rejection was conducted, because this decision is subjective. Finally, a baseline correction was applied using a pre-stimulus baseline interval of -200 to 0 ms.

CDA extraction

To remove the contribution of any VWM-unspecific, bilateral activity, the CDA was computed as a difference wave on a trial-by-trial basis by subtracting activity ipsilateral to cued items (presented left or right of screen center) from contralateral activity. The CDA amplitude was extracted from a time window of 300-900 ms after the onset of the memory array. We computed the mean CDA amplitude for each participant separately for each set size (i.e., 2, 4 and 6 cued items). For the computation of the CDA, we used posterior parietal, lateral occipital and posterior temporal electrode sites (i.e., left electrode cluster: P3, T5/P7, O1; right electrode cluster: P4, T6/P8, O2). First, the difference was calculated in electrode pairs (P3/P4), (P7/P8), (O1/O2) and then averaged. The CDA was calculated on a trial-by-trial basis for all set sizes and for all trials (i.e., correct and incorrect trials). The final step was to compute the overall average CDA for each set size by averaging the CDA of the right and left cue direction of the respective set size.

Data Quality and Psychometric Internal Consistency

Estimates of data quality and psychometric internal consistency were reported. Data quality estimates characterize the precision of group-level ERP estimates, whereas internal consistency estimates indicate whether scores are measured reliably enough to differentiate between individuals, which is crucial for studying individual differences (Clayson, Brush, et al., 2021; Clayson & Miller, 2017; Luck et al., 2021). These metrics were reported to characterize the obtained data, but data were not excluded based on these metrics to be consistent with the procedures of the original study. Arithmetically derived estimates of the standard error of the mean were used to characterize data quality (Luck et al., 2021). These estimates

separately quantified the precision of CDA for each set size (2, 4, and 6 cued items) using single-trial estimates of CDA (contralateral-ipsilateral activity differences). Psychometric internal consistency estimates used generalizability theory equations to compute coefficients of dependability for difference scores (Baldwin et al., 2015; Brennan, 1992; Clayson, Baldwin, et al., 2021; Clayson, Brush, et al., 2021; Sundre, 1993). Time-window mean amplitude estimates of single-trial scores of ipsilateral and contralateral activities were used to estimate the observed group-level internal consistency of the difference scores. Dependability of contralateral-ipsilateral activity difference scores was estimated separately for each set size and data collection site using the ERP Reliability Analysis Toolbox (Clayson, Carbine, et al., 2021; Clayson & Miller, 2017). Because CDA scores were calculated as the difference between activity from different electrode sites on the same trial, residual covariances were estimated because the constituent events of the difference scores are co-occurring (Clayson, Baldwin, et al., 2021).

Outcome-neutral test

To ensure that the data can test the stated hypotheses, we included quality checks and outcome-neutral tests. As an outcome-neutral test, we tested the presence of an asymmetry between contra- or ipsilateral electrode clusters time-locked to the memory array. For this, we averaged the ERPs across all set sizes and all subjects (i.e., grand averaged ERP) elicited by memory arrays that were either contra- or ipsilateral to electrode positions. A paired sample t-test for CDA between ipsilateral and contralateral sites was performed separately at each study site to verify the expected within-lab CDA experimental effect. If the t-test was significant ($p < 0.05$) with more negative CDA for contralateral activity than for ipsilateral activity, then this pattern of effect justified moving forward with testing the proposed hypotheses.

Statistical analysis

For all the statistical analyses, frequentist and Bayesian approaches were used. To estimate effect sizes, the statistical analyses were initially conducted for each participating lab separately. Because of the small sample size in each lab, we refrained from interpretation of the lab-specific statistics. However, the overall replication success for the project was determined based on meta-analytically pooled effect sizes, as per the defined criteria.

Statistical analysis for Hypothesis #1

A repeated-measures ANOVA of the CDA amplitude in the original study revealed a significant main effect for set size. Post-hoc t-tests showed significant increases in CDA amplitude for set sizes 4 and 6 compared to set size 2, with no significant differences between set sizes 4 and 6. In accordance with the original study, we also conducted repeated-measures ANOVA. The significance level was set to $p < 0.02$ uncorrected for multiple comparisons. If the ANOVA revealed a significant main effect, we further conducted post-hoc t-tests (with a significance level of $p < 0.02$, one-sided). We specifically tested one-sided, because we hypothesized a significant increase in CDA amplitude from arrays of 2 items per side to arrays of 4 items per side [H1.1] and 6 items per side [H1.2.]. As in the original study, we adjusted the p-values with the Greenhouse-Geisser correction for nonsphericity (Jennings & Wood, 1976). If the ANOVA revealed a significant main effect, and the post-hoc t-tests show a significant increase in the CDA amplitude between arrays of 2 items per side and arrays of 4 items per side or 6 items per side, it supported hypotheses [H1.1] and [H1.2], respectively.

We ran the corresponding analyses in a Bayesian analytical framework using a Bayesian generalized linear mixed models implemented in the brms R package (Bürkner, 2017). In the following formulas fixed effects are denoted by a "+" symbol and interaction effects by an "$*$" symbol, in line with the Wilkinson notation (Wilkinson & Rogers, 1973). The predictor variable was set size (factor with 3 levels: set sizes 2, 4 and 6; reference level = set size 2). The covariates included gender (factor with 2 levels: male, female; reference level = female), handedness (factor with 2 levels: right, left; reference level = right) from Edinburgh Handedness Inventory (EHI). Due to the very small number of participants in the other and ambidextrous categories, these subjects were excluded from the Bayesian analyses. Set size and all covariates were modeled as fixed effects, while site and subject were included as random effects. Please note that the Bayesian models included only random intercepts for subject and laboratory. More complex random effects structures, such as random slopes or fully specified hierarchical models, led to frequent non convergence and unstable parameter estimates. To ensure reliable and interpretable results, we therefore adopted a parsimonious random effects specification that captured between subjects and between laboratory variability while maintaining stable model estimation. Subsequently, the credible intervals (CIs) of the posterior distributions were calculated from the newly estimated levels of significance. We opted not to calculate Bayes factors for point estimates to determine whether the effect was zero or unequal to zero. This decision was made because these Bayes factors, which rely on the Savage-Dickey ratio, heavily depend on the selection of the prior distribution for each effect. Instead, we employed a different approach: we considered a model parameter to be significant if its 98%-CI did not include zero. As suggested by Gelman (2007), the predictors and outcome variables were scaled to achieve a mean of 0 and a standard deviation of 0.5

(Gelman, 2007). For initial prior distributions, uninformative Cauchy priors were set to a mean of 0 and a standard deviation of 2.5.

Importantly, the original Vogel and Machizawa (2004) study conducted ANOVA without covariates. To ensure direct compatibility with the original findings, our primary replicated analysis therefore also relied on an ANOVA without covariates. Additional analyses extending the original design conducted within the Bayesian framework included covariates to account for between subject and site-relevant variability.

Statistical analysis for Hypothesis #2

In the original paper, VWM capacity was positively correlated with the CDA amplitude increase from set size 2 to 4 (i.e., when the smaller set size is below typical adult working memory capacity estimates), but not from set size 4 to 6 (i.e., when both set sizes are at or exceed capacity estimates for typical adults). To replicate this, we calculated the mean CDA amplitude increase from set size 2 to 4, and from set size 4 to 6, for each subject individually. The VWM capacity was calculated using the same formula as in the original study. This formula was introduced by Pashler (1988) and refined by Cowan (2001). It is based on the assumption that if a person can retain K items from an S-item array, then the changed item should be among the K items being held in memory on (K/S) trials, leading to correct answers on (K/S) trials where an item changed (Cowan, 2001; Pashler, 1988). The formula accounts for the false alarm rate to adjust for guessing and is expressed as $K = S \times (H - F)$, where K is the memory capacity, S is the set size, H is the observed hit rate in the given set size, and F is the false alarm rate in the given set size. The resulting K scores from all set sizes (i.e., 2, 4, 6) were used to compute an average K score, which we used as the behavioral measure of VWM capacity. The relationship between VWM capacity and an increase in CDA amplitude from 2 to 4 items was statistically tested using Pearson's correlation. The significance level was set to $p < 0.02$ (one-sided). If the Pearson's correlation revealed a significant positive relationship between VWM capacity and the CDA amplitude increase from 2 to 4 items, it supported the hypothesis [H2.1]. Furthermore, we conducted a Bayesian linear mixed model with a prior assuming the reported correlation coefficient from the original study ($r = 0.78$). Again, significance is considered if the 98%-CI of the model parameter does not include zero.

Replication Success

Replication success was assessed for each hypothesis separately and was defined operationally as a statistically significant random-effects meta-analytic estimate (at $p < 0.02$) combining the results from the different laboratories, in the same direction as in the original study.

Hypothesis [H1.3] and [H2.2] were analyzed using an equivalence test for meta-analyses (Lakens, 2017). The equivalence test assesses whether the difference in CDA amplitude between arrays of 4 items and 6 items is as extreme as the smallest effect size of interest (SESOI) using the two one-sided tests (TOST) procedure implemented in the R package TOSTER (Caldwell, 2022; Lakens, 2017). To perform TOST, the SESOI and its lower and upper equivalence bounds must be established. Simonsohn recommended specifying the equivalence bounds for replication studies using the “small telescopes approach” (Simonsohn et al., 2015). The idea is to consider the effect size that would give the original study 33% power. If the original study had 33% power, the probability of observing a significant effect, if there was a true effect, is too low to reliably distinguish signal from noise. Using the small telescopes approach for hypothesis [H1.3], the SESOI is $d = 0.36$. An alternative approach would be to calculate the smallest effect size that can be detected at a predefined power level (e.g., 90%), given the sample size and alpha level. With this approach the smallest effect size would be very similar to the small telescope approach (i.e., $d = 0.44$). Therefore, we decided to define the SESOI based on the “small telescopes approach” (i.e., $d = 0.36$) as this approach was specifically recommended for replication studies. The TOST procedure was then conducted against these bounds based on the SESOI. If the 96% confidence interval of the meta-analytic effect size falls within the equivalence bounds, the observed meta-analytic effect is statistically equivalent (Lakens, 2017). In order to test hypothesis [2.2], which postulated that there is no correlation between the subject's VWM capacity and the CDA amplitude increase from 4 to 6 items, we conducted another equivalence test. Similar to hypothesis [1.3], we used the small telescope approach to specify the SESOI (i.e., $r = \pm 0.29$).

Finally, sequential Bayesian updating was employed by fitting a Bayesian model for each hypothesis separately to each dataset. The posterior distributions obtained from each analysis were used as priors for the next analysis, allowing evidence to be accumulated across the datasets from different labs. This approach was expected to produce greater statistical power than independent analyses and yield more robust outcome parameters.

Sensitivity analyses

Recently, there have been concerns regarding the validity of the K score as a measure of VWM capacity. Specifically, some researchers in the field have noted that K operates under the assumption of all-or-none memories and does not account for individual decision biases, which can lead to an overestimation of capacity depending on the observer's strategy (Brady et al., 2023; Williams et al., 2022).

In light of these concerns, we conducted an additional analysis using the d' (d prime) metric. d' is a commonly used measure in signal detection theory and provides a unitless, normalized measure of sensitivity that is independent of response bias. D prime is defined as d' = Z(hit rate) - Z(false alarm rate). A hit was defined as reporting a color change when there was one, and a false alarm as reporting such a change when no change occurred. The resulting d' scores from all set sizes (i.e., 2, 4, 6) were used to compute an average d' score. Our reasoning is that replicating the results using both K score (as the primary analysis) and d' (as an additional analysis) would provide stronger evidence for the observed effect, as it would demonstrate that the results are not solely dependent on the characteristics of the K measure.

Deviations from Preregistration

A small number of deviations from the preregistered Stage-1 protocol were necessary to ensure acceptable data quality across laboratories.

(1) We preregistered that data would be collected across 10 laboratories. However, after Stage 1 acceptance, the laboratory in Finland withdrew from the project due to regulatory reasons. After the laboratory in Finland withdrew from the project, several participating labs agreed to recruit additional participants to maintain the planned statistical power. Subsequently, we were also able to include an additional site (Prof. Sauseng's lab from Zurich; UZH-NCN), which further increased the total sample size. As a result, the total number of participants exceeded the preregistered target and was unevenly distributed across laboratories (Table 3).

(2) The preregistered blink threshold of 50 μV in the VEOG channel resulted in an excessive number of rejected trials in several laboratories. In the seminal CDA study by Vogel and Machizawa (2004), the authors state that "ERPs were recorded using our standard recording and analysis procedures, including rejection of trials contaminated by blinks or large eye movements," citing earlier work from 1998 (Vogel et al., 1998). However, neither the 2004 paper nor the referenced 1998 work specifies an explicit VEOG amplitude threshold for blink detection. Similarly, other early CDA studies do not provide a clearly defined numeric criterion for blink rejection based on VEOG amplitude (Drew & Vogel, 2008; Fukuda & Vogel, 2011;

McCollough et al., 2007; Vogel & Machizawa, 2004). In later work from the Vogel/Awh lab, blink detection was implemented using algorithmic, sliding-window step-function approaches applied to the VEOG signal. For example, Adam (2018) detected blinks using a sliding window on VEOG (window size = 200 ms, step size = 10 ms, threshold = 50 µV). Subsequent studies (Hakim et al., 2021) further refined this approach by using smaller thresholds (e.g., 30 µV; window size = 80 ms, step size = 10 ms). Importantly, these algorithmic thresholds are not directly comparable to our single-value VEOG amplitude cutoff, as sliding-window methods detect rapid step-like deflections over short temporal windows rather than absolute VEOG magnitude at a single time point. In the present study, the preregistered blink criterion of 50 µV was applied as a single-value VEOG threshold, which resulted in an excessive number of rejected epochs in several laboratories. Because single-value approaches require higher cutoffs to achieve a comparable level of stringency, we increased the blink threshold to 90 µV to retain an adequate number of usable epochs across laboratories.

(3) We used ZaplinePlus instead of the preregistered ZapLine algorithm. Several laboratories exhibited substantial line-noise contamination, and ZaplinePlus provided more robust identification and suppression of line-noise harmonics while preserving neural signals.

(4) The preregistered criterion defined blocking as amplitude ranges <1 µV for 30 ms. Applying this threshold would have resulted in the rejection of an unacceptably large proportion of trials (>30% on average). The threshold was therefore adjusted to <0.1 µV for 60 ms, which preserved data quality while avoiding disproportionate data loss and is in line with the logic of flatline detection implemented in automated preprocessing pipelines such as clean_rawdata (Kothe & Makeig, 2013).

(5) In a subset of laboratories (USF, OSU, UZH-NCN), the eye calibration task did not yield accurate HEOG amplitudes, resulting in estimates that were either unrealistically high or too low and therefore would have led to retaining too many trials or removing an excessive number of trials. For these laboratories, we applied a fixed HEOG threshold of 30 µV, which closely matched the heuristic value used in the seminal Vogel and Machizawa (2004) and the average 1° amplitude obtained from the laboratories with valid eye calibration data.

(6) The preregistered plan stated that equivalence testing would be evaluated using 90% confidence intervals, consistent with the standard TOST framework ($\alpha = 0.05$). However, all EEGManyLabs replication studies adopt a project-wide significance threshold of $\alpha = 0.02$ to ensure a uniform inferential standard across analyses and tasks. The threshold of $\alpha = 0.02$ was applied consistently throughout the present project for all frequentist statistical tests. To

maintain internal consistency with the EEGManyLabs inferential framework, the equivalence tests were therefore conducted using 96% confidence intervals (i.e., 1 - 2α), rather than the preregistered 90%.

(7) An exploratory preprocessing pipeline incorporating ICA followed by ICLabel-based component classification was implemented to assess robustness of the results. The ICA pipeline was not preregistered and is therefore reported only in the Supplement.

3. Results

To provide a concise overview of the results across hypotheses, preprocessing pipelines, and statistical frameworks, we summarize the replication outcomes in Table 3. Across the direct, advanced, and ICA pipelines, outcome neutral effects and set-size-related hypotheses (H1.1-H1.3) were consistently replicated using both frequentist and Bayesian approaches. In contrast, the hypothesis targeting the association between the CDA increase from set size 2 to 4 and individual VWM capacity (H2.1) showed less consistent support across analyses, with replication depending on the specific pipeline and modelling approach and effects consistently remaining close to the threshold of statistical significance. Finally, equivalence between the CDA increase from set size 4 to 6 and individual VWM capacity (H2.2) was again consistently supported across all analyses.

Detailed results for the advanced and ICA pipelines are reported in the Supplementary Materials, as their outcomes closely mirrored those of the direct pipeline and are included there to reduce redundancy and streamline the presentation of the results.

Table 3. Summary of the results across analyses and preprocessing pipelines

	Direct			Advanced			ICA		
	Frequen-tist	Bayesian LMM	Bayesian Seq.	Frequen-tist	Bayesian LMM	Bayesian Seq.	Frequen-tist	Bayesian LMM	Bayesian Seq.
Outcome N.	✓	✓	✓	✓	✓	✓	✓	✓	✓
H1.1.	✓	✓	✓	✓	✓	✓	✓	✓	✓
H1.2.	✓	✓	✓	✓	✓	✓	✓	✓	✓
H1.3.	✓	✓	✓	✓	✓	✓	✓	✓	✓
H2.1.	X	X	X	X	X	X	X	✓	✓
H2.2.	✓	✓	✓	✓	✓	✓	✓	✓	✓

Note. ✓ = indicates successful replication. X = indicates failure to replicate. LMM = Linear Mixed Model. Seq = Sequential Updating.

3. 1. Sample Characteristics and Artifact-Related Trial Rejection

Table 4 provides an overview of the demographic characteristics of all participants across the 10 laboratories. In total, 304 participants were recruited. After trial rejection and subject

exclusion, the final sample comprised 217 participants in the direct pipeline, 231 participants in the advanced pipeline, and 266 participants in the ICA pipeline. For comparison, the original study (Vogel & Machizawa, 2004) reported only limited demographic information, including the total sample size (N = 36) and an age range of 21-33 years. No further details (e.g., sex, handedness) were provided.

Table 4. Demographic characteristics of the full sample (N = 304).

	N	Age		Gender			Handedness		
		M	SD	Female	Male	Other	R	L	A
DART	28	20.29	3.41	9	18	1	23	3	0
USF	50	19.64	1.90	41	9	0	34	11	3
JGU	25	23.04	3.02	17	8	0	22	2	1
WWU	30	22.80	2.27	25	5	0	28	2	0
NDSU	28	24.93	5.80	15	13	0	23	5	0
OSU	24	24.13	3.39	18	6	0	NA	NA	NA
UI	30	24.50	4.90	21	7	0	25	4	0
TUOS	25	20.04	2.95	20	5	0	23	2	0
UZH-MPR	40	23.35	2.82	25	15	0	40	0	0
UZH-NCN	24	22.92	3.97	16	8	0	24	0	0

Note. M = Mean. SD = Standard deviation. R = Right. L = Left. A = Ambidextrous.

Table 5 summarizes the number of rejected trials across laboratories for each exclusion criterion and preprocessing pipeline. As preregistered, trials were rejected due to amplitude thresholds, blocking artefacts, VEOG artefacts (i.e., blinks), and HEOG artefacts (i.e., saccades). In addition, 4 laboratories collected eye-tracking data. For these sites, blink and saccade rejections in the advanced and ICA pipelines were based on ET-detected events rather than VEOG/HEOG thresholds. In contrast, the original study (Vogel & Machizawa, 2004) only reports that trials containing artefacts were rejected, without providing quantitative information on the number or proportion of rejected trials.

Table 5. Number of rejected trials by rejection criterion, laboratory, and preprocessing pipeline.

Pipeline	Amplitude		Blocking		VEOG		HEOG		ET Blink		ET Saccade		Final N
	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	
Dartmouth College													
Direct	48.18	109.06	0	0	76.79	122.39	81.89	126.94	–	–	–	–	23
Advanced	10.14	27.11	0	0	32.21	35.01	38.64	83.92	–	–	–	–	27
ICA	10.18	27.17	0	0	31.86	35.47	36.75	81.7	–	–	–	–	27
University of South Florida													
Direct	99.14	130.42	0	0	165.24	151.02	97.44	129.18	–	–	–	–	26
Advanced	67.1	99.34	0	0	134.58	143.48	71.4	91.71	–	–	–	–	33
ICA	60.22	96.66	0	0	83.64	150.88	40.56	75.12	–	–	–	–	40
Johannes Gutenberg University Mainz													
Direct	3.56	6.93	0	0	129.12	152.28	174.96	123.82	–	–	–	–	11
Advanced	3.52	6.95	0	0	128.68	152.34	174.84	123.76	–	–	–	–	11
ICA	1.36	2.97	0	0	3.72	7.23	4.52	11.72	–	–	–	–	25
University of Münster													
Direct	40.27	98.91	0.1	0.4	34.27	36.03	36.27	46.43	–	–	–	–	26
Advanced	38.47	134.37	0.43	2.03	36.67	43.52	35.27	93.58	21.97	22.33	19.53	20.08	28
ICA	32.83	132.98	0.47	2.37	28.2	130.24	24.2	131.42	21.97	22.33	19.53	20.08	28
North Dakota State University													
Direct	5.29	11.44	0	0	78.93	97.46	63.68	78.58	–	–	–	–	22
Advanced	2.57	6.81	0	0	65.89	98.55	63.82	78.82	49.18	74.41	52.43	85.41	25
ICA	2.25	6.54	0	0	6.86	12.75	6.32	23.76	49.18	74.41	52.43	85.41	25
The Ohio State University													
Direct	15.71	32.92	0	0	76.04	89.99	26.21	59.38	–	–	–	–	20
Advanced	8.88	14.74	0	0	35.21	61.92	13.96	17.78	82.42	94.46	100.25	109.43	19
ICA	8.21	14.69	0	0	6.58	15.08	11.33	18.89	82.42	94.46	100.25	109.43	19
University of Iceland													
Direct	11.8	32.44	0	0	79.83	121	79.6	82.64	–	–	–	–	24
Advanced	11.37	31.91	0	0	6.57	15.96	144.67	195.85	–	–	–	–	22
ICA	9.77	31.05	0	0	4.47	14.74	86.87	165.68	–	–	–	–	24
University of Sheffield													

Direct	11.92	24.8	0	0	135.6	127.05	106.96	93.48	–	–	–	–	15
Advanced	11.52	23.68	0	0	135.68	127.77	106.76	93.47	–	–	–	–	16
ICA	10.68	23.74	0	0	32.68	61.54	8.96	16.66	–	–	–	–	24
University of Zurich (UZH-MPR)													
Direct	41.1	88.56	0	0	103.98	131.67	86.98	114.74	–	–	–	–	30
Advanced	2.75	7.53	0	0	66.65	105.35	86.83	114.73	46.88	106.69	148.45	163.04	30
ICA	2.3	6.35	0	0	9.1	19.83	44.68	152.02	46.88	106.69	148.45	163.04	30
University of Zurich (UZH-NCN)													
Direct	11.33	25.82	0	0	93.17	137.66	119.71	115.23	–	–	–	–	20
Advanced	11.21	25.74	0	0	87.42	137.2	111.71	119.02	–	–	–	–	20
ICA	1.63	3.31	0	0	3.58	8.64	22.58	25.69	–	–	–	–	24

Note. M = Mean. SD = Standard deviation. Eye-tracking (ET) data and ET-based rejection criteria were available only for the WWU, NDSU, OSU, and UZH-MPR laboratories. An em dash (—) indicates that ET data were either not recorded at that site or were not included in the analysis pipeline.

3. 2. Behavioral Results

Performance in the color change detection task, expressed as accuracy, was 87.7% ± 14.5% for set size 2, 75.8% ± 13.0% for set size 4, and 65.2% ± 10.5% for set size 6 (Table 6). The average accuracy across all set sizes was 76.2% ± 12.0%. Performance in the color change detection task, expressed as K scores, was 1.61 ± 0.52 for set size 2, 2.26 ± 0.93 for set size 4, and 2.11 ± 1.10 for set size 6. The average K score across all set sizes was 1.99 ± 0.80. Performance expressed in d′ (d-prime) was 3.09 ± 1.26 for set size 2, 1.82 ± 0.87 for set size 4, and 1.14 ± 0.62 for set size 6, with an overall mean d′ of 2.01 ± 0.86. The average K score and average d′ were strongly correlated (r = 0.93, p = 2.93e-95), indicating high convergence between the two behavioral measures. Additionally, all further analyses based on K-score and d′ yielded highly similar results. The original study (Vogel & Machizawa, 2004) did not report detailed behavioral summary statistics (e.g., accuracy, K-scores, or d′), although an approximate average K-score of 2.8-2.9, averaged across set sizes 2, 4, and 6, can be visually inferred from the published figure.

Table 6. Performance in the color change detection task.

Accuracy (%)				K-score				D-prime			
2	4	6	Average	2	4	6	Average	2	4	6	Average

	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD	M	SD
DART	91	6	80	8	69	8	80	11	1.71	0.18	2.53	0.56	2.48	0.87	2.24	0.71	3.24	0.88	2.00	0.58	1.29	0.46	2.17	1.04
USF	75	23	65	18	56	12	65	20	1.27	0.89	1.67	1.25	1.41	1.13	1.45	1.11	2.19	1.69	1.26	1.02	0.73	0.65	1.39	1.33
JGU	90	10	77	9	67	8	78	13	1.67	0.37	2.36	0.78	2.23	1.03	2.09	0.82	3.25	1.01	1.89	0.75	1.2	0.55	2.11	1.16
WWU	94	10	85	9	73	8	84	13	1.81	0.39	2.93	0.72	2.97	1.02	2.57	0.92	3.99	1.07	2.48	0.73	1.6	0.6	2.69	§.28
NDSU	84	12	68	10	59	8	70	14	1.5	0.46	1.72	0.79	1.41	0.84	1.54	0.72	2.76	1.23	1.32	0.62	0.74	0.41	1.61	1.19
OSU	93	5	82	6	70	7	82	11	1.76	0.15	2.68	0.46	2.46	0.83	2.30	0.68	3.41	0.78	2.19	0.66	1.31	0.45	2.31	1.07
UI	86	16	72	11	61	7	73	16	1.49	0.65	1.83	0.84	1.5	0.77	1.61	0.77	2.75	1.36	1.43	0.69	0.83	0.43	1.67	1.21
TUOS	88	8	74	11	64	10	75	14	1.59	0.25	2.06	0.84	2.03	1.12	1.89	0.84	2.78	0.7	1.68	0.93	1.18	0.76	1.88	1.04
UZH-MPR	92	8	80	9	69	8	80	13	1.75	0.31	2.55	0.67	2.43	0.9	2.24	0.76	3.5	0.91	2.14	0.68	1.35	0.54	2.33	1.15
UZH-NCN	92	7	81	9	70	9	81	12	1.77	0.2	2.65	0.66	2.66	0.93	2.36	0.78	3.58	0.96	2.2	0.72	1.44	0.54	2.41	1.16

Note. M = Mean. SD = Standard deviation.

3. 3. Data Quality and Psychometric Internal Consistency

Estimates of data quality (i.e., standardized measurement error; SME) and psychometric internal consistency (i.e., dependability coefficients) were computed for each laboratory and each set size for all 3 preprocessing pipelines (Clayson, Brush, et al., 2021). Dependability coefficients for the CDA, derived from generalizability theory and conceptually analogous to coefficient alpha in classical test theory (Clayson, Carbine, et al., 2021; Shavelson & Webb, 2012) were generally high across sites, with most laboratories achieving values above 0.75 (Table 7). For set size 2, dependability ranged from 0.39 to 0.90. A similar pattern was observed for set size 4, with coefficients spanning 0.36 to 0.93. For set size 6, dependability values ranged from 0.35 to 0.92. Overall, CDA dependability coefficients in the present study were generally within or above the range considered acceptable for preliminary research (≥ 0.70) and, for many laboratories, approached or exceeded the more stringent threshold recommended for studies of group differences (≥ 0.80; Clayson & Miller, 2017; Nunnally & Bernstein, 1994).

Data quality, indexed by the SME of single-trial CDA amplitudes, showed comparable patterns across sites and set sizes (Table 7). SME values for set size 2 ranged from 0.297 to 1.439. For set size 4, SME values ranged from 0.299 to 1.382, and for set size 6, values ranged from 0.297 to 1.426. Across the majority of laboratories, SME values clustered between approximately 0.35-0.50, indicating precise trial-level CDA estimates. USF again showed substantially larger SME values across all set sizes, consistent with its lower dependability coefficients. Importantly, as emphasized by Luck and colleagues (2021), SME is intended as

a continuous index of data quality rather than a metric with predefined acceptability thresholds, and it is therefore difficult to classify SME values as intrinsically “good” or “bad”. Instead, lower SME values indicate higher precision and reliability. Together, these results demonstrate that CDA amplitudes were estimated with good precision and strong internal consistency across most participating laboratories, providing confidence in the reliability of the measurements underlying the primary analyses.

Table 7. Data Quality Metrics Across Laboratories.

	Dependability (Estimate and 98%-CI)			SME		
Pipeline	**Set size 2**	**Set size 4**	**Set size 6**	**Set size 2**	**Set size 4**	**Set size 6**
			Dartmouth College			
Direct	0.84 (0.71, 0.92)	0.88 (0.77, 0.94)	0.87 (0.76, 0.94)	0.36	0.35	0.38
Advanced	0.83 (0.69, 0.92)	0.88 (0.78, 0.94)	0.88 (0.78, 0.94)	0.36	0.34	0.36
ICA	0.83 (0.69, 0.92)	0.88 (0.78, 0.94)	0.87 (0.77, 0.94)	0.36	0.34	0.36
			University of South Florida			
Direct	0.39 (0.18, 0.60)	0.36 (0.15, 0.58)	0.35 (0.15, 0.57)	1.44	1.38	1.43
Advanced	0.51 (0.31, 0.70)	0.50 (0.28, 0.69)	0.51 (0.29, 0.69)	1.22	1.16	1.24
ICA	0.42 (0.21, 0.62)	0.51 (0.29, 0.70)	0.43 (0.22, 0.64)	1.30	1.20	1.16
			Johannes Gutenberg University Mainz			
Direct	0.85 (0.70, 0.93)	0.87 (0.75, 0.94)	0.89 (0.78, 0.95)	0.38	0.38	0.35
Advanced	0.84 (0.69, 0.93)	0.87 (0.75, 0.94)	0.88 (0.77, 0.95)	0.38	0.38	0.35
ICA	0.85 (0.71, 0.93)	0.88 (0.76, 0.94)	0.90 (0.80, 0.95)	0.39	0.27	0.28
			University of Münster			
Direct	0.78 (0.63, 0.89)	0.71 (0.50, 0.85)	0.80 (0.65, 0.90)	0.45	0.47	0.46
Advanced	0.79 (0.63, 0.89)	0.74 (0.56, 0.87)	0.81 (0.67, 0.91)	0.40	0.40	0.41
ICA	0.77 (0.61, 0.89)	0.78 (0.62, 0.89)	0.86 (0.74, 0.93)	0.41	0.40	0.41
			North Dakota State University			
Direct	0.77 (0.60, 0.89)	0.77 (0.59, 0.89)	0.79 (0.63, 0.90)	0.40	0.40	0.39
Advanced	0.80 (0.64, 0.90)	0.75 (0.57, 0.88)	0.82 (0.67, 0.91)	0.38	0.39	0.38
ICA	0.77 (0.59, 0.88)	0.74 (0.53, 0.87)	0.80 (0.63, 0.90)	0.36	0.38	0.37
			The Ohio State University			

Direct	0.76 (0.56, 0.89)	0.84 (0.69, 0.92)	0.86 (0.75, 0.94)	0.39	0.39	0.41
Advanced	0.75 (0.54, 0.88)	0.84 (0.71, 0.92)	0.86 (0.75, 0.94)	0.42	0.43	0.44
ICA	0.71 (0.48, 0.87)	0.84 (0.70, 0.92)	0.83 (0.69, 0.92)	0.44	0.44	0.46
			University of Iceland			
Direct	0.81 (0.67, 0.90)	0.84 (0.71, 0.92)	0.77 (0.62, 0.88)	0.42	0.42	0.43
Advanced	0.82 (0.68, 0.91)	0.84 (0.71, 0.92)	0.78 (0.60, 0.89)	0.47	0.49	0.57
ICA	0.81 (0.67, 0.90)	0.83 (0.70, 0.92)	0.79 (0.64, 0.89)	0.44	0.44	0.49
			University of Sheffield			
Direct	0.77 (0.57, 0.89)	0.80 (0.64, 0.91)	0.82 (0.66, 0.91)	0.50	0.49	0.48
Advanced	0.77 (0.57, 0.89)	0.80 (0.64, 0.90)	0.81 (0.64, 0.91)	0.49	0.49	0.48
ICA	0.85 (0.72, 0.93)	0.82 (0.66, 0.92)	0.78 (0.60, 0.90)	0.39	0.38	0.40
			University of Zurich (UZH-MPR)			
Direct	0.81 (0.68, 0.89)	0.91 (0.85, 0.95)	0.87 (0.78, 0.92)	0.30	0.30	0.30
Advanced	0.85 (0.74, 0.92)	0.91 (0.85, 0.95)	0.86 (0.76, 0.92)	0.33	0.33	0.34
ICA	0.86 (0.76, 0.92)	0.91 (0.85, 0.95)	0.85 (0.76, 0.92)	0.33	0.34	0.35
			University of Zurich (UZH-NCN)			
Direct	0.90 (0.81, 0.95)	0.93 (0.86, 0.96)	0.92 (0.84, 0.96)	0.50	0.36	0.70
Advanced	0.90 (0.80, 0.95)	0.93 (0.86, 0.97)	0.92 (0.84, 0.96)	0.50	0.36	0.70
ICA	0.88 (0.79, 0.95)	0.92 (0.85, 0.96)	0.91 (0.84, 0.96)	0.31	0.31	0.31

Note. CI = Confidence Interval. SME = Standardized Measurement Error.

3. 4. Results of the Direct Replication

Here, we report the results of the direct replication pipeline, which follows the procedures of the original study as closely as possible. Figure 3 displays the grand-average contralateral and ipsilateral ERPs, the grand-average CDA with set size effects, an error-bar plot with 98%-CI illustrating the set size effects, as well as the relationship between memory capacity defined as K-score and the amplitude increase from 2 to 4 items, with all subjects from all laboratories collapsed into a single combined dataset.

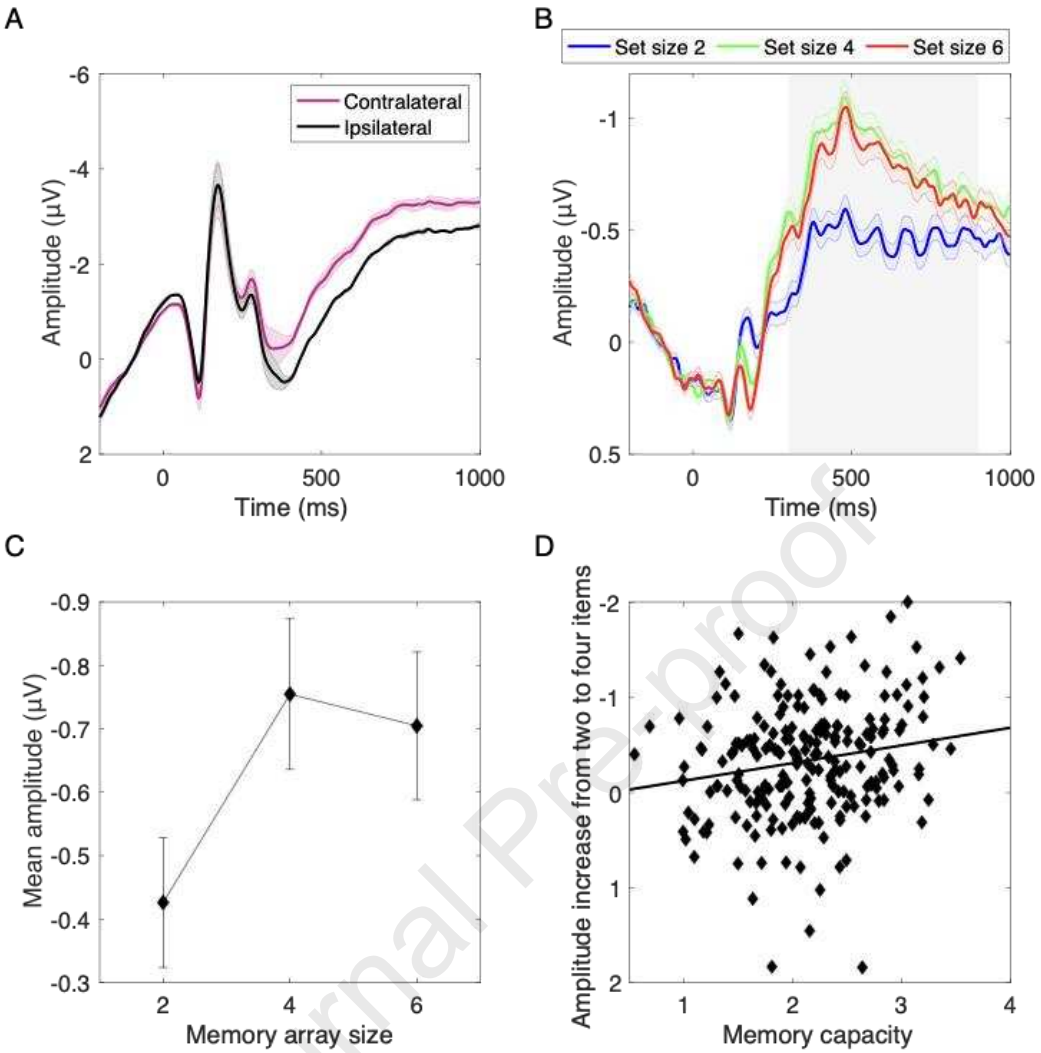

Figure 3. Direct Pipeline. Grand-average CDA effects across all participants and labs, averaged across posterior electrode clusters (left: P3, P7, O1; right: P4, P8, O2). (A) Outcome Neutral Test: Contralateral vs. ipsilateral ERP. Subplots (B) and (C) illustrate the set size effect specified in hypothesis #1, with (B) showing ERP responses by set size and (C) displaying CDA amplitudes for set sizes 2, 4, and 6 with 98% confidence intervals. The grey shaded region in (B) indicates the time window used for statistical analysis. (D) Association between the CDA increase from set size 2 to 4 and VWM capacity (K-score). The line represents the best linear fit. For visualization purposes only, participants were pooled across laboratories into a single combined sample.

3. 4. 1. Outcome Neutral Test

We first conducted the outcome-neutral test to assess whether the expected contralateral-ipsilateral asymmetry was present across labs, using a frequentist approach, a Bayesian generalized linear mixed-model and a Bayesian sequential updating procedure.

3. 4. 1. 1. Frequentist Approach

The frequentist approach revealed a robust contralateral-ipsilateral asymmetry across laboratories. Paired-sample t-tests at each lab consistently yielded significant contralateral-ipsilateral differences (all $p < 0.05$), with contralateral activity exhibiting more negative amplitudes than ipsilateral activity (Table 8).

Table 8. Direct Pipeline. Paired-sample t-tests showing contralateral-ipsilateral CDA asymmetry at each laboratory.

Lab	M_{Contra}	SD_{Contra}	M_{Ipsi}	SD_{Ipsi}	t-value	df	p-value	Hedges' gz
DART	-2.22	1.56	-1.48	1.69	-5.13	22	3.89e-05***	-1.03
USF	-0.33	1.02	-0.07	0.98	-2.76	25	0.011*	-0.53
JGU	-2.91	1.35	-2.42	1.22	-2.93	10	0.015*	-0.82
WWU	-2.66	2.00	-1.76	1.82	-8.61	25	6.02e-09***	-1.64
NDSU	-1.68	1.58	-1.16	1.44	-4.61	21	1.50e-04***	-0.95
OSU	-1.88	1.56	-1.00	1.16	-6.69	19	2.14e-06***	-1.44
UI	-2.94	1.54	-2.50	1.29	-3.45	23	0.002**	-0.68
TUOS	-2.35	1.41	-1.91	1.19	-2.53	14	0.024*	-0.62
UZH-MPR	-1.75	1.59	-1.09	1.57	-6.18	29	9.61e-07***	-1.10
UZH-NCN	-1.75	1.31	-0.84	1.09	-5.62	19	2.03e-05***	-1.21

Note. M = Mean. SD = Standard deviation. Df = Degrees of freedom.

**p < 0.05. **p < 0.01. ***p < 0.001*

A random-effects meta-analysis confirmed a significant overall asymmetry (average $g_z = -0.96$, 98%-CI = [-1.28, -0.65], t = -8.62, p = 1.21e-5), replicating the results from the original study (Figure 4). Taken together, these results verify that the dataset provides a reliable contralateral-ipsilateral asymmetry effect, thereby fulfilling the preregistered criterion for proceeding with the main analyses.

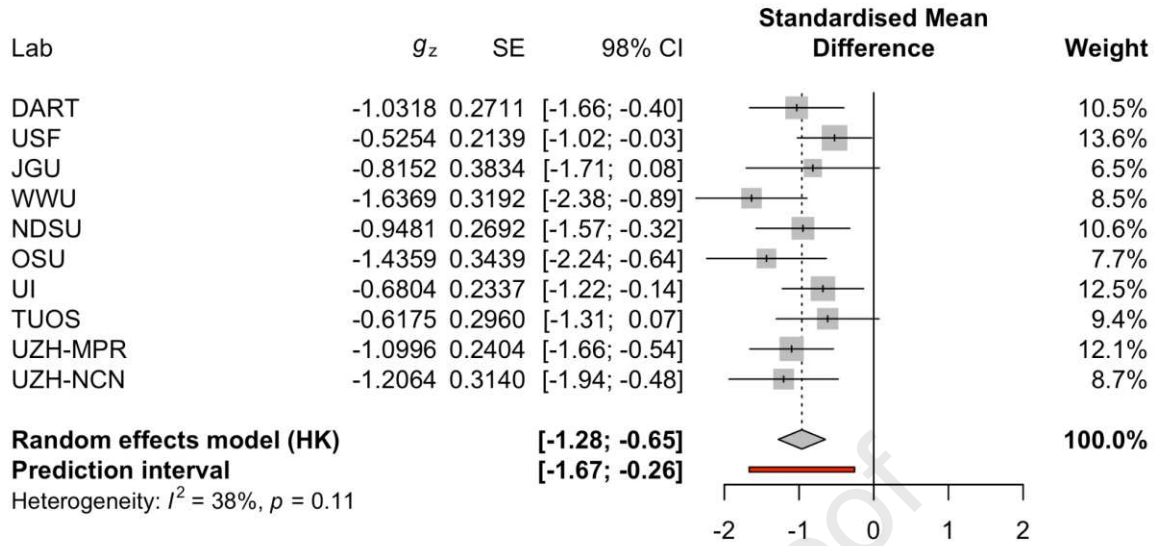

Figure 4. Direct Pipeline. Forest plot of the meta-analysis for outcome-neutral test: contralateral-ipsilateral asymmetry across labs. For each laboratory, the plot shows the effect size estimate Hedges' g with its standard error and corresponding 98% confidence interval. Squares represent lab-specific effect size estimates, with square size proportional to the inverse-variance weight, indicating the relative contribution of each laboratory to the pooled estimate (larger squares reflect greater weight due to higher precision). Horizontal lines denote 98% confidence intervals. The random-effects meta-analytic estimate (Hartung-Knapp adjustment) is shown as a diamond with its 98% confidence interval, together with the 98% prediction interval. Between-laboratory heterogeneity is quantified using I^2, with the associated p-value reported.

3. 4. 1. 1. Bayesian Approach

To complement the frequentist analyses, we estimated a Bayesian generalized linear mixed model predicting lateralized ERP amplitude from laterality (factor with 2 levels: contralateral, ipsilateral; reference level = contralateral) while including laboratory and subject as random intercepts. As preregistered (see Stage 1 Registered Report), all Bayesian models additionally controlled for gender and handedness.

ERP Amplitude ~ Laterality * Gender * Handedness + (1|Subject) + (1|Lab)

CDA amplitudes were credibly more positive at ipsilateral than contralateral electrodes (Estimate = 0.20, 98%-CI = [0.15, 0.24]), with the 98% credible interval for the laterality effect excluding zero, indicating robust evidence for the contralateral-ipsilateral asymmetry effect across laboratories. None of the covariates showed credible associations with CDA amplitude, as the 98% credible intervals for gender and handedness all included zero.

Finally, we implemented the Bayesian sequential updating procedure to evaluate the stability of the contralateral-ipsilateral asymmetry as evidence accumulated across laboratories. For the first site, we fit a Bayesian regression model predicting CDA amplitude from laterality using weakly informative (uniform) priors. The posterior from this model was then used as the prior for the next lab, and this process was repeated iteratively across all ten sites, allowing the posterior distribution to be continuously updated as new data were incorporated. The final posterior distribution after integrating evidence across all labs provided clear support for the predicted contralateral-ipsilateral asymmetry (Estimate = 0.21, 98%-CI = [0.18, 0.24]), with CDA amplitudes more positive at ipsilateral than contralateral electrodes.

3. 4. 2. Hypothesis #1: Set Size Effects

Having established that all laboratories reproduced the expected contralateral-ipsilateral asymmetry, we next turned to the analyses examining set-size-dependent changes in CDA amplitude. In line with the preregistration, we assessed these effects using both a frequentist repeated-measures ANOVA with post-hoc t-tests, a Bayesian generalized linear mixed-model approach and a Bayesian sequential updating procedure.

3. 4. 2. 1. Frequentist Approach

We first examined the set-size effect between arrays of 2, 4, and 6 items using the frequentist approach. A repeated-measures ANOVA was conducted separately for each laboratory (Table 9). Out of the 10 participating sites, 6 labs (DART, JGU, WWU, NDSU, and both UZH labs) showed a statistically significant main effect of set size (all $p < 0.02$), whereas 4 labs (USF, OSU, UI and TUOS) did not reach statistical significance ($p > 0.02$).

Table 9. Direct Pipeline. Results of the repeated-measures ANOVA assessing the set size effect across labs.

Lab	df	F-value	p-value
DART	(2, 44)	5.69	0.006**
USF	(2, 50)	0.80	0.455
JGU	(2, 20)	6.72	0.006**
WWU	(2, 50)	14.37	1.51e-5***
NDSU	(2, 42)	4.34	0.019*
OSU	(2, 38)	4.08	0.025
UI	(2, 46)	1.10	0.332
TUOS	(2, 28)	2.13	0.138
UZH-MPR	(2, 58)	31.54	5.28e-10***

UZH-NCN	(2, 38)	13.09	4.74e-5***

Note. Df = Degrees of freedom.

**p < 0.02. **p < 0.01. ***p < 0.001*

Hypothesis #1.1. CDA amplitude increase from set size 2 to 4

Next, we tested Hypothesis H1.1, which predicted an increase in CDA amplitude from set size 2 to 4. To this end, we computed post-hoc paired *t*-tests for all sites (Table 10). 6 labs (DART, JGU, WWU, OSU, and both UZH labs) showed statistically significant differences between set sizes 2 and 4, whereas in 4 labs the difference did not reach statistical significance (USF, NDSU, UI and TUOS).

Table 10. Direct Pipeline. Post-hoc t-tests for the set size 2 vs. 4 comparisons across laboratories.

Lab	M_{SZ2}	SD_{SZ2}	M_{SZ4}	SD_{SZ4}	t-value	df	p-value	Hedges' gz
DART	-0.61	0.62	-0.87	0.77	3.41	22	0.003**	0.69
USF	-0.31	0.75	-0.14	0.68	-0.98	25	0.334	-0.19
JGU	-0.26	0.53	-0.67	0.55	4.33	10	0.001**	1.20
WWU	-0.53	0.67	-1.11	0.56	4.82	25	5.90e-05***	0.92
NDSU	-0.33	0.64	-0.64	0.61	2.53	21	0.020	0.52
OSU	-0.64	0.46	-1.02	0.77	2.75	19	0.013*	0.59
UI	-0.35	0.73	-0.52	0.72	1.52	23	0.141	0.30
TUOS	-0.30	0.68	-0.58	0.74	2.12	14	0.052	0.52
UZH-MPR	-0.30	0.56	-0.85	0.70	6.85	29	1.58e-07***	1.22
UZH-NCN	-0.61	0.68	-1.14	0.85	4.54	19	2.24e-04***	0.97

Note. M = Mean. SD = Standard deviation. Df = Degrees of freedom.

**p < 0.02. **p < 0.01. ***p < 0.001*

To assess the overall effect across labs, we performed a random-effects meta-analysis on the site-wise effect sizes (Figure 5). The meta-analysis revealed a robust set size effect (average g_z = 0.63, 98%-CI = [0.25, 1.02], t = 4.63, p = 0.001), replicating the original effects. Thus, despite minor variability across individual labs, the meta-analytic evidence indicates a reliable increase in CDA negativity for set size 4 compared with set size 2.

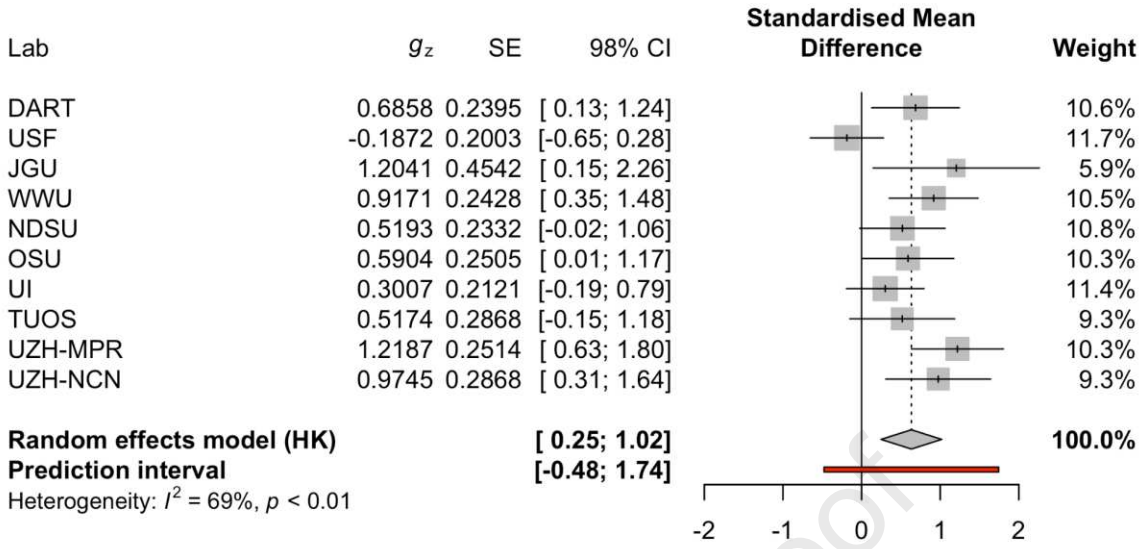

Figure 5. Direct Pipeline. Forest plot of the meta-analysis for set size 2 vs. 4 across laboratories. For each laboratory, the plot shows the effect size estimate Hedges' g with its standard error and corresponding 98% confidence interval. Squares represent lab-specific effect size estimates, with square size proportional to the inverse-variance weight, indicating the relative contribution of each laboratory to the pooled estimate (larger squares reflect greater weight due to higher precision). Horizontal lines denote 98% confidence intervals. The random-effects meta-analytic estimate (Hartung-Knapp adjustment) is shown as a diamond with its 98% confidence interval, together with the 98% prediction interval. Between-laboratory heterogeneity is quantified using I^2, with the associated p-value reported.

Hypothesis #1.2. CDA amplitude increase from set size 2 to 6

We next compared CDA amplitudes between set sizes 2 and 6 (Table 11). Post-hoc t-tests revealed statistically significant differences between set sizes 2 and 6 only in 3 out of 10 labs (WWU, and both UZH labs).

Table 11. Direct Pipeline. Post-hoc t-tests for the set size 2 vs. 6 comparison across laboratories.

Lab	M_{SZ2}	SD_{SZ2}	M_{SZ6}	SD_{SZ6}	t-value	df	p-value	Hedges' gz
DART	-0.61	0.62	-0.75	0.75	1.69	22	0.105	0.34
USF	-0.31	0.75	-0.35	0.71	0.21	25	0.837	0.04
JGU	-0.26	0.53	-0.53	0.72	2.22	10	0.051	0.62
WWU	-0.53	0.67	-1.05	0.69	4.10	25	3.78e-04***	0.78
NDSU	-0.33	0.64	-0.58	0.56	2.14	21	0.044	0.44
OSU	-0.64	0.46	-0.97	0.79	2.21	19	0.040	0.47
UI	-0.35	0.73	-0.44	0.68	0.62	23	0.541	0.12
TUOS	-0.30	0.68	-0.43	0.81	0.88	14	0.394	0.21
UZH-MPR	-0.30	0.56	-0.82	0.63	6.72	29	2.25e-07***	1.20
UZH-NCN	-0.61	0.68	-0.99	0.82	3.86	19	0.001**	0.83

Note. M = Mean. SD = Standard deviation. Df = Degrees of freedom.

*p < 0.02. **p < 0.01. ***p < 0.001

To quantify the overall effect across laboratories, we again performed a random-effects meta-analysis on the site-level effect sizes (Figure 6). This analysis yielded a significant overall difference between set sizes 2 and 6 (average g_z = 0.53, 98%-CI = [0.16, 0.81], t = 4.24, p = 0.002), replicating the original effect. These results indicate a reliable increase in CDA negativity for set size 6 compared with set size 2 across labs.

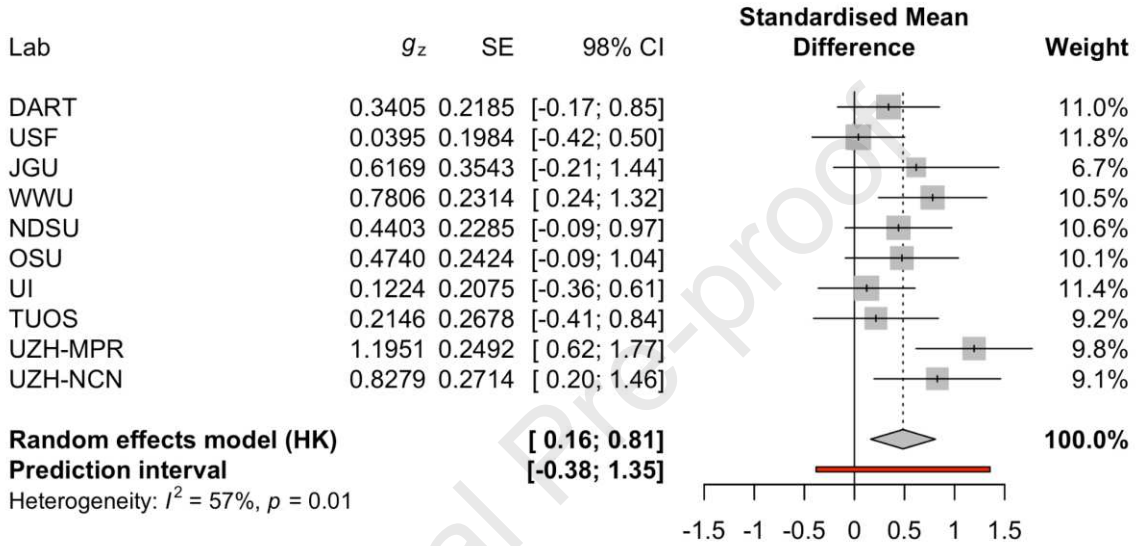

Figure 6. Direct Pipeline. Forest plot of the meta-analysis for set size 2 vs. 6 across laboratories. For each laboratory, the plot shows the effect size estimate Hedges' g with its standard error and corresponding 98% confidence interval. Squares represent lab-specific effect size estimates, with square size proportional to the inverse-variance weight, indicating the relative contribution of each laboratory to the pooled estimate (larger squares reflect greater weight due to higher precision). Horizontal lines denote 98% confidence intervals. The random-effects meta-analytic estimate (Hartung-Knapp adjustment) is shown as a diamond with its 98% confidence interval, together with the 98% prediction interval. Between-laboratory heterogeneity is quantified using I^2, with the associated p-value reported.

Hypothesis #1.3. Equivalence between CDA amplitude in set sizes 4 and 6

To assess whether CDA amplitudes for set sizes 4 and 6 were statistically equivalent, we conducted an equivalence test on the random-effects meta-analytic effect size, using equivalence bounds defined by the preregistered SESOI (Cohen's d = ±0.36; Figure 7). The TOST was performed on the pooled Hedges' g estimate across all labs and its standard error (average g_z = -0.13, 98%-CI = [-0.29, 0.03], t = -2.31, p = 0.046), with significance evaluated using 96% confidence intervals (α = 0.02), in line with the statistical framework adopted across analyses. The TOST indicated statistical equivalence between set sizes 4 and 6 (Z = 4.06,

96%-CI = [-0.25, -0.01], p = 2.50e-5). The traditional null-hypothesis significance test did not reach significance (Z = -2.31, 98%-CI = [-0.26, 0.003], p = 0.021), indicating no detectable difference between set sizes 4 and 6. Taken together, our analyses indicate that CDA amplitudes for set sizes 4 and 6 do not differ significantly and fall within the preregistered equivalence bounds, replicating the pattern reported in the original study.

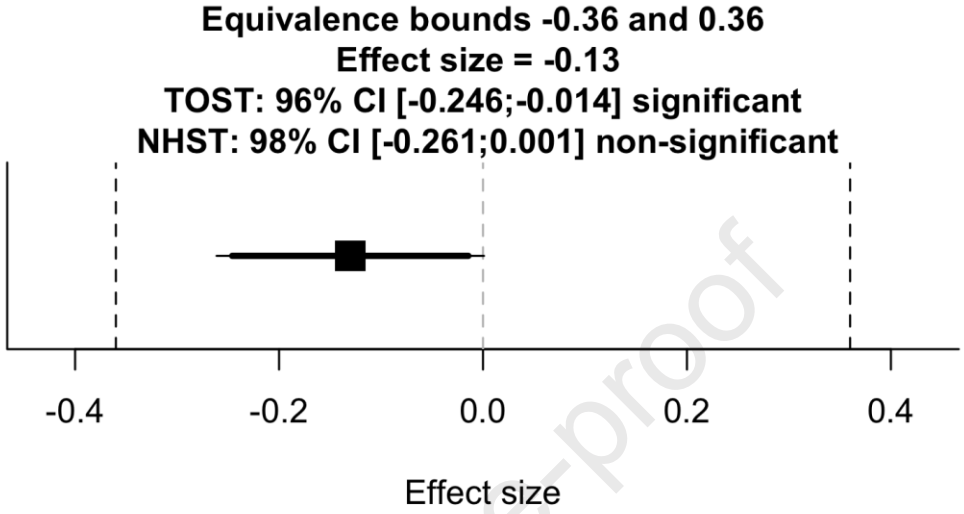

Figure 7. Direct Pipeline. Equivalence test results for the meta-analytic comparison of set sizes 4 and 6 (H1.3). Mean difference (black square) is shown together with 96% TOST confidence intervals (thick horizontal lines) and 98% NHST confidence intervals (thin horizontal lines). Equivalence bounds were set at $d = -0.36$ and $d = 0.36$ based on the small-telescope approach (dashed vertical lines).

3. 4. 2. 2. Bayesian Approach

To complement the frequentist analyses, we estimated a Bayesian generalized linear mixed model predicting CDA amplitude from set size (factor with 3 levels: set sizes 2, 4, and 6; reference level: set size 2) while including subject and laboratory as random intercepts. Again, the Bayesian linear mixed model additionally controlled for gender and handedness, in accordance with Stage 1 Registered Report.

$$\text{CDA} \sim \text{SetSize} * \text{Gender} * \text{Handedness} + (1|\text{Subject}) + (1|\text{Lab})$$

The posterior distributions provided clear evidence for set-size-dependent changes in CDA amplitude. Relative to set size 2 (reference level), CDA amplitudes were more negative for both set size 4 (Estimate = -0.22, 98%-CI = [-0.31, -0.13]) and set size 6 (Estimate = -0.19, 98%-CI = [-0.28, -0.10]). When re-leveling the model to use set size 4 as the reference, the posterior estimate for the contrast between set sizes 6 and 4 was small (Estimate = 0.03, 98%-

CI = [-0.04, 0.11]), with the credible interval including zero, indicating no evidence for a difference in CDA amplitude between the two larger set sizes. None of the covariates showed credible associations with CDA amplitude: the 98%-CIs for gender and handedness predictors all included zero, indicating no evidence that these factors meaningfully contributed to variability in CDA amplitude. Taken together, the Bayesian results closely mirror the frequentist findings, providing convergent evidence that CDA amplitudes reliably increase from set size 2 to 4 and from set size 2 to 6 across laboratories.

Finally, we implemented the Bayesian sequential updating procedure to evaluate the stability of the set-size effects as evidence accumulated across laboratories. For the first site, we fit a Bayesian regression model predicting CDA amplitude from set size using weakly informative (uniform) priors. The posterior from this model was then used as the prior for the next lab, and this process was repeated iteratively across all ten sites, allowing the posterior distribution to be continuously updated as new data were incorporated. The final posterior distribution after integrating evidence across all labs provided clear support for the predicted set-size effects. Relative to set size 2 (reference level), CDA amplitudes were credibly more negative for both set size 4 (Estimate = -0.24, 98%-CI = [-0.30, -0.18]) and set size 6 (Estimate = -0.18, 98%-CI = [-0.24, -0.13]). In both cases, the 98% credible intervals excluded zero, indicating strong cumulative evidence for load-dependent increases in CDA negativity as memory load increases. When re-leveling the model to use set size 4 as the reference, the posterior estimate for the contrast between set sizes 6 and 4 was small (Estimate = 0.06, 98%-CI = [-0.001, 0.12]), with the credible interval including zero, indicating no evidence for a difference in CDA amplitude between the set sizes 4 and 6.

3. 4. 3. Hypothesis #2: Correlation of CDA Increase with VWM Capacity

In the following analyses, we investigated whether individual differences in behavioral performance in the change detection task are related to CDA amplitude increases from 2 to 4 items and from 4 to 6 items, using both frequentist and Bayesian statistical models to test our preregistered hypotheses.

3. 4. 3. 1. Frequentist Approach

Hypothesis #2.1. Correlation between CDA increase from set size 2 to 4 and VWM capacity

First, we examined whether the increase in CDA amplitude from set size 2 to 4 correlates with VWM capacity, computed as average K-score over set sizes 2, 4, and 6. The correlation

coefficients were computed separately for each lab and then synthesized using a random-effects meta-analysis. The correlations did not reach statistical significance in a single lab (all $p > 0.02$). As shown in Table 12 and Figure 8, individual lab correlations varied in magnitude and direction, with confidence intervals often spanning zero.

Table 12. Direct Pipeline. Correlations between CDA amplitude increase from set sizes 2 to 4 and VWM capacity across laboratories

Lab	M_{SZ2}	SD_{SZ2}	M_{SZ4}	SD_{SZ4}	Pearson's r	p-value
DART	-0.25	0.36	2.22	0.52	0.38	0.074
USF	0.17	0.87	1.75	0.60	-0.33	0.094
JGU	-0.41	0.31	1.85	0.66	0.42	0.200
WWU	-0.59	0.62	2.69	0.45	0.14	0.485
NDSU	-0.31	0.58	1.67	0.61	0.16	0.488
OSU	-0.38	0.61	2.29	0.46	0.44	0.050
UI	-0.18	0.57	1.76	0.48	0.15	0.488
TUOS	-0.29	0.52	1.85	0.52	0.26	0.342
UZH-MPR	-0.55	0.44	2.26	0.62	-0.05	0.792
UZH-NCN	-0.54	0.53	2.36	0.61	0.04	0.862

Note. M = Mean. SD = Standard deviation.

*$p < 0.02$. **$p < 0.01$. ***$p < 0.001$

Across the ten labs, the random-effects meta-analysis yielded a small positive correlation that did not reach statistical significance ($r = 0.13$, 98%-CI = [-0.09, 0.36], $t = 1.68$, $p = 0.128$), indicating no reliable evidence for a positive association between VWM capacity defined as a K-score and the CDA increase from set size 2 to 4. Taken together, these results do not provide evidence for the positive relationship predicted by Hypothesis 2.1, nor do they replicate the strong association reported in the original study.

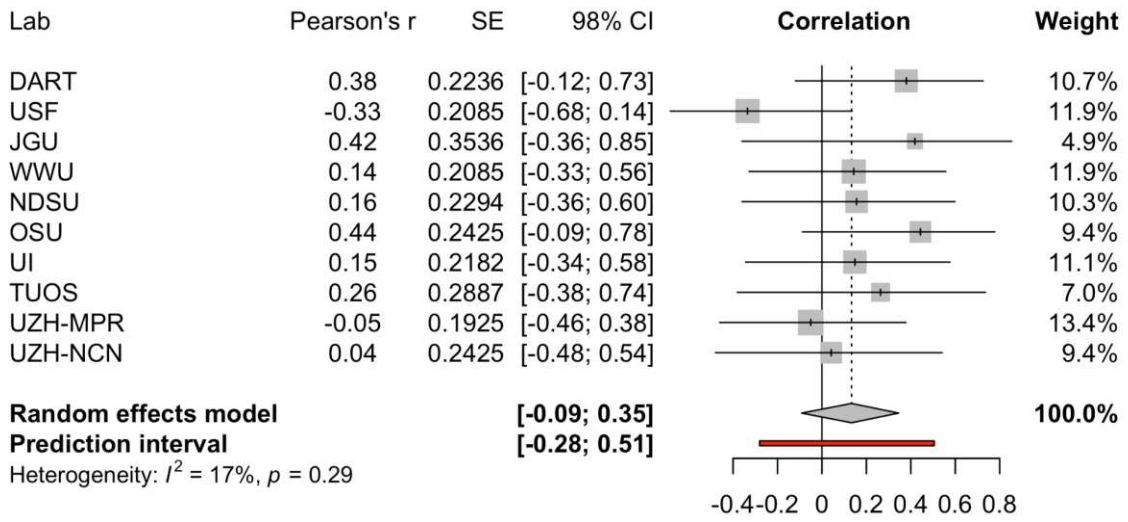

Figure 8. Direct Pipeline. Forest plot of the meta-analysis for the correlation between amplitude increase from set size 2 to 4 and VWM capacity defined as K-score. For each laboratory, the plot shows Pearson's r correlation coefficient with its standard error and corresponding 98% confidence interval. Squares represent lab-specific effect size estimates, with square size proportional to the inverse-variance weight, indicating the relative contribution of each laboratory to the pooled estimate (larger squares reflect greater weight due to higher precision). Horizontal lines denote 98% confidence intervals. The random-effects meta-analytic estimate (Hartung-Knapp adjustment) is shown as a diamond with its 98% confidence interval, together with the 98% prediction interval. Between-laboratory heterogeneity is quantified using I^2, with the associated p-value reported.

As a sensitivity analysis, we conducted an analogous meta-analysis using d′ instead of K-score to quantify performance. Across the ten laboratories, the random-effects meta-analysis revealed a small positive but non-significant correlation (r = 0.09, 98% CI = [-0.14, 0.32], t = 1.10, p = 0.301), supporting the results of the analyses using K-score.

Given the comparatively poor data quality observed for the USF laboratory (Table 7), we conducted an exploratory sensitivity analysis to assess whether this site disproportionately influenced the correlation between the CDA increase from set size 2 to 4 and individual VWM capacity. Specifically, we repeated the full set of analyses for H2.1 after excluding the USF dataset. Across the 9 remaining laboratories, the random-effects meta-analysis revealed a slightly stronger positive correlation between CDA increase and K score that reached statistical significance (r = 0.19, 98%-CI [0.02, 0.35], t = 3.16, p = 0.013). In contrast, the corresponding analysis using d′ as the measure of VWM capacity showed a smaller, non-significant association (r = 0.13, 98%-CI [-0.10, 0.35], t = 1.63, p = 0.142).

<u>Hypothesis #2.2. Correlation between CDA increase from set size 4 to 6 and VWM capacity</u>

Subsequently, to evaluate whether the association between VWM capacity and the CDA amplitude increase from set size 4 to 6 was small enough to be considered statistically equivalent, we conducted an equivalence test on the random-effects meta-analytic correlation, using the preregistered SESOI (r = ±0.29) derived from the small telescopes approach (Figure 9). The TOST indicated statistical equivalence (Z = -3.03, 96%-CI = [-0.20, 0.20], p = 0.002). The traditional null-hypothesis significance test did not reach significance (Z = 0.01, 98%-CI = [-0.23, 0.23], p = 0.994), indicating no detectable correlation between VWM capacity and the CDA increase from set sizes 4 to 6. Taken together, these results show that the CDA increase from 4 to 6 items is both statistically indistinguishable from zero and statistically equivalent within the preregistered bounds, thereby replicating the pattern reported in the original study.

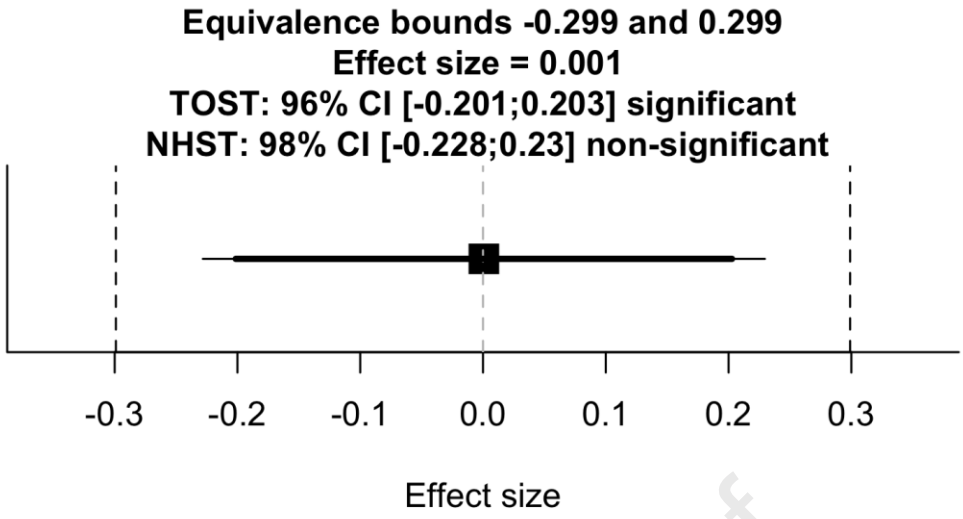

Figure 9. Direct Pipeline. Equivalence test for the correlation between VWM capacity and CDA increase from set size 4 to 6 (H2.2). Mean difference (black square) is shown together with 96% TOST confidence intervals (thick horizontal lines) and 98% NHST confidence intervals (thin horizontal lines). Equivalence bounds were set at $d = -0.299$ and $d = 0.299$ based on the small-telescope approach (dashed vertical lines).

3. 4. 3. 2. Bayesian Approach

Hypothesis #2.1. Correlation between CDA increase from set size 2 to 4 and VWM capacity

To complement the frequentist analyses, we examined the relationship between VWM capacity and the CDA increase from set size 2 to 4 within a Bayesian framework. We fitted a Bayesian linear mixed model with VWM capacity as the predictor and the CDA difference (set size 4 minus set size 2) as the outcome, including laboratory as a random intercept, and gender and handedness as covariates. Consistent with our preregistration, the model used a prior centered on the correlation reported in the original study (r = 0.78).

$$CDA_{4\text{-}2} \sim K * Gender * Handedness + (1|Lab)$$

The posterior distribution provided no evidence for a relationship between the CDA increase and VWM capacity (Estimate = 0.16, 98%-CI = [-0.06, 0.38]). The posterior mean was close to zero, and the credible interval spanned both positive and negative values, indicating a high probability that any true association is negligible. The posterior distributions for all covariates (gender and handedness) likewise overlapped zero, suggesting that none of the covariates contributed meaningfully to the model.

Finally, we applied the Bayesian sequential updating procedure to evaluate whether evidence for a relationship between VWM capacity and the CDA increase from set size 2 to 4 would accumulate across laboratories. Starting with weakly informative priors for the first dataset, the posterior from each lab was used as the prior for the subsequent lab, allowing the evidence to be integrated iteratively across all ten sites. The final posterior distribution converged tightly around zero, with the posterior mean close to zero and the 98% credible interval including both positive and negative values (Estimate = 0.14, 98%-CI [-0.03, 0.31]). Taken together, the sequential updating analysis confirms the results from the frequentist, and standard Bayesian models: across accumulating evidence from all laboratories, there is no indication that CDA increase from set size 2 to 4 is related to VWM capacity.

Hypothesis #2.2. Correlation between CDA increase from set size 4 to 6 and VWM capacity

Finally, we examined the relationship between VWM capacity and the CDA increase from set size 4 to 6 within a Bayesian framework. We fitted a Bayesian linear mixed model with VWM capacity as the predictor and the CDA difference (set size 6 minus set size 4) as the outcome, including laboratory as a random intercept. The original study did not report the correlation between amplitude increase from set size 4 to 6, therefore, we used weakly informative (uniform) priors.

$$CDA_{6\text{-}4} \sim K * Gender * Handedness + (1|Lab)$$

The posterior distribution provided no evidence for a relationship between the CDA increase from set size 4 to 6 and VWM capacity (Estimate = 0.04, 98%-CI = [-0.17, 0.26]). The posterior mean was close to zero, and the credible interval spanned both positive and negative values, indicating a high probability that any true association is negligible. The posterior distributions for all covariates (gender and handedness) likewise overlapped zero, suggesting that none of the covariates contributed meaningfully to the model.

Consistent with this result, the Bayesian sequential updating analysis yielded an estimate close to zero (Estimate = 0.002, 98%-CI = [-0.17, 0.18]), with the credible interval again spanning both positive and negative values, providing no evidence for a reliable association between the CDA increase from set size 4 to 6 and VWM capacity.

Discussion

In this multi-site replication study, we aimed to evaluate two central claims from Vogel and Machizawa (2004): that the CDA reliably tracks increases in VWM load, and that the magnitude of the CDA increase from set size 2 to 4 correlates with individual VWM capacity. Consistent with the original findings, we reproduced a clear CDA lateralization effect and robust set size increases in CDA amplitude from set size 2 to 4 and from set size 2 to 6, demonstrating that the component is a robust neural marker of VWM maintenance across diverse samples, EEG systems, and testing environments. In contrast, we found no compelling evidence for the predicted positive correlation between CDA amplitude increase from set size 2 to 4 and VWM capacity. With the exception of a small effect that reached statistical credibility only in the Bayesian sequential-updating analysis of the advanced pipeline, all other estimates were smaller than those reported in the original study and previous meta-analyses. Taken together, these findings indicate that while the CDA is a robust indicator of memory load, its value as a reliable marker of individual differences in VWM capacity is likely substantially more limited than previously assumed.

CDA as a Marker of Memory Load (Hypothesis #1)

Across laboratories, our results provide strong support for the robustness of the CDA as a neural marker of visual working memory load, in line with a large body of prior work (Adam et al., 2018; Diamantopoulou et al., 2011; Drew & Vogel, 2008; Feldmann-Wüstefeld, 2021; Feldmann-Wüstefeld et al., 2018; Hakim et al., 2019, 2020; Kang & Woodman, 2014; Kundu et al., 2013; Kuo et al., 2012; Lefebvre et al., 2013; Leonard et al., 2013; Luria et al., 2016; Ngiam et al., 2021; Roy & Faubert, 2023; Störmer et al., 2013; Tröndle & Langer, 2024; Tsubomi et al., 2013; Unsworth et al., 2015; Villena-González et al., 2020; Vogel & Machizawa, 2004). In both the direct and advanced pipelines, all (or all but one) laboratories showed a robust contralateral-ipsilateral asymmetry, and the meta-analytic effect sizes were large ($g_z \approx -0.95$), confirming that a sustained contralateral negativity emerges consistently during the retention interval. Importantly, CDA amplitudes increased systematically from set size 2 to 4 and from 2 to 6, with random-effects meta-analyses yielding medium-to-large effects ($g_z \approx 0.45$-0.65) despite some variability in statistical significance at the level of individual sites, which is expected when the same underlying effect is tested repeatedly across multiple independent samples. Equivalence tests further indicated that CDA amplitudes for set sizes 4 and 6 were statistically equivalent and fell within the preregistered equivalence bounds, replicating the plateau pattern reported by Vogel and Machizawa (2004), which parallels the average human working memory capacity limit of around four items. The Bayesian mixed

models and sequential updating analyses converged with the frequentist results, providing clear evidence for load-dependent increases from set size 2 to 4 and 2 to 6, but no credible difference between 4 and 6 items. Together with the high dependability coefficients and low SME values observed at most sites, these findings demonstrate that the CDA can be measured with good precision and internal consistency across different samples, EEG systems, laboratories, and recordings environments. Thus, at the process level, our multi-site study confirms that the CDA is a robust and psychometrically reliable electrophysiological index of the number of items held in visual working memory.

Correlation of CDA Increase with VWM Capacity (Hypothesis #2)

In contrast to the robust set-size effects, our multi-lab replication provided only limited evidence for the predicted positive association between VWM capacity and the CDA increase from set size 2 to 4. In both, direct and advanced pipelines, correlations were non-significant in every laboratory, and the random-effects meta-analysis yielded a small, non-significant estimate. The Bayesian mixed model likewise produced a posterior centered near zero, indicating that the current data provided little evidence for a positive relationship between CDA amplitude increase and VWM capacity. In the ICA pipeline, frequentist correlations were again non-significant across all sites. However, the Bayesian mixed model and the sequential Bayesian updating both indicated a small positive association, with an estimated effect of approximately $r \approx 0.20$. Notably, an effect of $r \approx 0.20$ is small, accounting for roughly 4% of the variance, and thus represents only a weak association between neural and behavioral measures. Taken together, results from all pipelines converge on the conclusion that the CDA increase does not reliably track individual differences in VWM capacity and that any true relationship, if present, is substantially smaller and less robust than previously suggested.

The lack of a strong correlation between CDA and behavior deviates from a major finding of the original (2004) study and from later meta-analyses (Luria et al., 2016; Roy & Faubert, 2023). For instance, the meta-analysis of 11 studies by Luria et al. (2016) yielded a combined correlation of $r = 0.596$ (98%-CI = [0.51, 0.67]), suggesting a robust relationship between CDA amplitude increase and VWM capacity in the existing literature. Similarly, Roy and Faubert (2023) reported three medium-to-strong correlations ranging from $r = 0.26$ to 0.45.

One likely source of discrepancy between our findings and those in the existing literature is the limited sample size of most published studies examining the correlations between CDA and VWM capacity, together with substantial between-study variability in how the relationship

between CDA and VWM capacity is operationalized. Previous studies have correlated VWM capacity with differences between two different set sizes (e.g., 4-2, 3-1) or with CDA amplitude at a single set size (e.g., set size 3 or 6). Psychometrically, the reliability of a difference score is inherently limited when the two contributing variables are highly correlated, as is likely for capacity estimates at set sizes 4 and 6 provided that variability exists at both levels. Furthermore, difference scores may capture a distinct, and often smaller, portion of variance in individual differences in VWM than performance assessed at a single set size or across conditions, because shared variance between conditions is effectively removed. Given that mean capacity typically asymptotes around set size 4, the difference between set sizes 4 and 6 may, for many individuals, reflect compensatory strategies or the ability to maintain information beyond nominal capacity limits rather than core storage capacity per se. These analytic choices complicate direct comparisons across studies and may contribute to variability in reported effect sizes.

Additionally, the majority of studies included in Luria et al.'s (2016) meta-analysis were small (e.g., Diamantopoulou et al., N = 14; Drew et al., N = 33/18; Jost et al., N = 25; Kang & Woodman, N = 24; Kundu et al., N = 30; Kuo et al., N = 18; Lefebvre et al., N = 39; Leonard et al., N = 23; Störmer et al., N = 35; Tsubomi et al., N = 25), with sample sizes between 14 and 39 participants (Table 13). Such sample sizes fall far below the 250 participants needed for correlation estimates to stabilize (Schönbrodt & Perugini, 2013), making inflated and unstable correlations statistically likely. Consistent with this interpretation, the few larger studies in the literature report substantially smaller effects. For instance, Unsworth et al. (2015; N = 170) observed r = 0.33 between the CDA amplitude at set size 6 (i.e., instead of a difference between 2 set sizes) and VWM capacity, Adam et al. (2018; N = 72) reported r = 0.27 (again, CDA amplitude at set size 6) and Tröndle and Langer (2024; N = 55) found only a modest association (r = 0.22; CDA amplitude at set size 3). These correlations are closer to our own meta-analytic estimate and suggest that the true relationship, if present, is likely smaller, around r = 0.20-0.30. For context, when directly using CDA amplitude at individual set sizes instead of the CDA increase from set size 2 to 4, the meta-analytic correlations in our study were small and showed an inconsistent pattern across preprocessing pipelines. At set size 2, the meta-analytic correlations were near zero and non-significant (direct: r = 0.09, p = 0.234; advanced: r = 0.06, p = 0.382; ICA: r = 0.02, p = 0.786). At set size 4, correlations were slightly larger but remained non-significant (direct: r = 0.17, p = 0.079; advanced: r = 0.16, p = 0.122; ICA: r = 0.17, p = 0.055). The strongest associations were observed at set size 6, where correlations were small but statistically significant across pipelines (direct: r = 0.17, p = 0.019; advanced: r = 0.15, p = 0.034; ICA: r = 0.15, p = 0.022). Overall, this pattern

further supports the conclusion that the association between CDA amplitude and VWM capacity is modest at best.

More broadly, the multitude of analytical choices researchers face when quantifying the relationships between CDA and behavior is referred to as researcher degrees of freedom (Gelman & Loken, 2023; Simmons et al., 2011; Trübutschek et al., 2023). When researchers' decisions are made post-hoc or without explicit justification, they can inflate variability, obscure true effect magnitudes, and contribute to inconsistent findings across studies (Götz et al., 2024; Simmons et al., 2011). Standardizing CDA operationalizations, or systematically evaluating alternative definitions within, for example, a multiverse framework, would therefore be an important step toward determining whether, and under which conditions, the CDA indexes individual differences in VWM capacity (Götz et al., 2024; Sarma et al., 2024).

Another potential source of the discrepancy between our findings and previously reported medium-to-large effects is publication bias. We therefore conducted funnel-plot diagnostics (Sterne & Egger, 2001; Viechtbauer, 2010) on the complete set of 27 studies known to us that investigated the association between CDA and VWM capacity (Adam et al., 2018; Diamantopoulou et al., 2011; Drew & Vogel, 2008; Feldmann-Wüstefeld, 2021; Hakim et al., 2019, 2020; Kang & Woodman, 2014; Kundu et al., 2013; Kuo et al., 2012; Lefebvre et al., 2013; Leonard et al., 2013; Luria et al., 2016; Ngiam et al., 2021; Roy & Faubert, 2023; Störmer et al., 2013; Tröndle & Langer, 2024; Tsubomi et al., 2013; Unsworth et al., 2015; Villena-González et al., 2020; Vogel & Machizawa, 2004). For an overview of the studies published prior to EEGManyLabs, please refer to Table 13.

First, when restricting the meta-analysis to all prior #EEGManyLabs studies (k = 17; blue studies on Figure 10), the random-effects model yielded a large, pooled effect (Fisher's z = 0.58, 95%-CI = [0.45, 0.72], p = 2.22e-18), accompanied by substantial heterogeneity (I^2 = 56.73%). Funnel-plot asymmetry was pronounced in this restricted dataset with the Egger regression test statistically significant (z = 2.83, p = 0.005), providing strong evidence for small-study effects consistent with selective reporting. In contrast, when the full dataset including the #EEGManyLabs replication was analyzed (k = 27), evidence for publication bias was strongly attenuated. The Egger regression test was non-significant (z = 0.72, p = 0.474), and the limit estimate as the standard error approached zero was b = 0.28 (98%-CI = [-0.19, 0.76]), indicating no reliable inflation of effects in smaller samples. The random-effects model yielded a pooled Fisher's z of 0.45 (95%-CI = [0.33, 0.57], p = 4.64e-13), corresponding to a significant medium effect. However, heterogeneity was substantial (I^2 = 63.43%), indicating considerable variability in effect sizes across studies.

Taken together, these results indicate that the large correlations between VWM capacity and CDA amplitude reported in the early literature are substantially inflated by small-study effects, whereas the inclusion of the present large-scale multi-laboratory data strongly attenuates both effect size estimates and evidence for publication bias. These results highlight the need for large-scale, well-powered, and preregistered studies to approximate the true magnitude and robustness of a given effect.

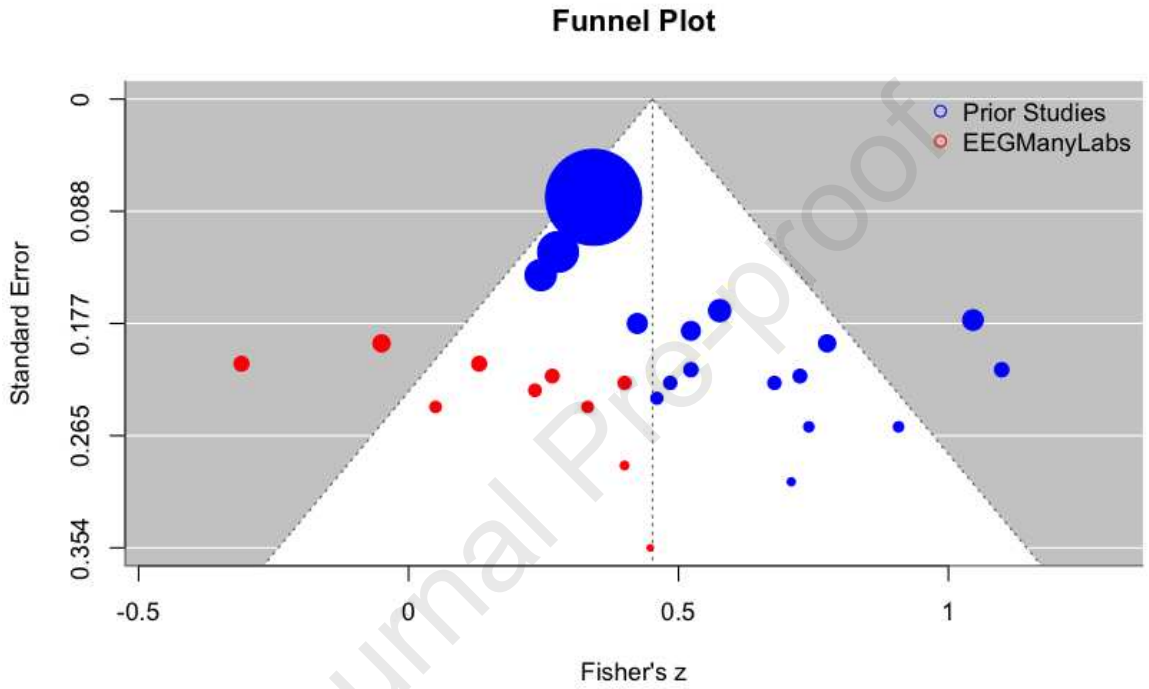

Figure 10. Funnel plot of the CDA-VWM capacity correlations across all 27 studies. Each point represents an individual study, plotted as Fisher's z against its standard error, and the size of each point is proportional to the study's sample size. The dashed vertical line indicates the pooled random-effects estimate (Fisher's z = 0.45), and diagonal dashed lines denote the 98% pseudo-confidence limits around the summary effect. Correlations from the EEGManyLabs datasets were derived from the direct replication.

Theoretical Implications for Interpreting the CDA

The dissociation between robust process-level effects and weak individual-differences correlations has important implications for the interpretation of CDA as a neural marker of VWM capacity. First, if the true brain-behavior correlation is modest ($r \approx 0.20$-0.30 at best), then CDA measurements may not reliably capture the same sources of individual variation that drive behavioral performance differences. Therefore, researchers should exercise caution when interpreting individual differences in CDA as reflecting differences in VWM capacity, and

that CDA-based inferences about capacity changes may need to be reconsidered or at minimum supported by parallel behavioral evidence.

Second, the present findings highlight a fundamental theoretical dissociation: although CDA is highly sensitive to within-individual variations in memory load, it shows little consistent association with between-person differences in capacity. Several non-mutually exclusive explanations warrant consideration. One possibility is that CDA primarily reflects the maintenance or active storage component of VWM, whereas behavioral capacity estimates such as Cowan's K conflate maintenance with a host of other processes including encoding efficiency, attentional filtering during stimulus presentation, resistance to interference, and decision or comparison processes during retrieval. If individuals vary substantially in these non-maintenance components, then CDA amplitude might correlate only weakly with overall task performance despite being a valid index of the number of items actively maintained. Another related possibility is that individuals with lower CDA amplitudes may engage compensatory strategies that preserve behavioral performance. For instance, some individuals might rely more heavily on verbal recoding, hierarchical chunking, or a quality-quantity tradeoff (maintaining fewer items at higher precision) to achieve similar behavioral capacity estimates despite differences in the neural signature of visual maintenance indexed by CDA. Such strategic flexibility could obscure the brain-behavior relationship at the individual-differences level while leaving within-person load effects intact.

Third, it is important to recognize that between-person variance in CDA amplitude likely reflects a mixture of cognitive and non-cognitive sources. Anatomical and biophysical factors such as skull thickness, cortical geometry, sulcal depth, and electrode-to-source distance can all influence EEG amplitude measures and vary considerably across individuals. Critically, these anatomical factors contribute to stable between-person differences in CDA amplitude but are entirely orthogonal to cognitive capacity. Within-person experimental manipulations (such as varying set size) effectively control for these individual anatomical differences, which may explain why load effects are robust and replicable while individual-differences correlations are weak. Future work employing source localization methods or individual structural MRI data might help to disentangle cognitive from anatomical sources of CDA variance.

Conclusion

In summary, our large-scale, preregistered multi-lab replication provides evidence that the CDA is a reliable neural marker of visual working memory load, consistently reproducing the

contralateral-ipsilateral asymmetry and the characteristic load-dependent increases observed in the seminal work. At the same time, our results do not support the widely cited claim that individual differences in CDA amplitude reflect individual differences in visual working memory capacity. Across pipelines, statistical frameworks, and laboratories, the association between CDA increase and behavioral capacity was small, inconsistent, and far weaker than previously reported in the literature. Together with evidence of substantial heterogeneity among past studies, the present findings suggest that earlier reports of medium-to-large correlations were likely inflated by small sample sizes, analytic flexibility, and variation in measurement reliability. These results highlight the critical role of large, collaborative, preregistered projects for resolving long-standing debates and establishing the true size and robustness of foundational cognitive neuroscience effects, as well as the necessity of adequately powered studies to obtain stable and interpretable estimates of individual differences.

Table 13. Overview of studies published prior to EEGManyLabs examining contralateral delay activity and working memory capacity.

Study	N	K-score (Mean, SD)						CDA definition	r
		SZ1	SZ2	SZ3	SZ4	SZ6	Average K		
Adam et al., 2018, Exp1	72	0.95 (0.04)	—	2.41 (0.33)	—	2.53 (0.53)	2.62 (1.00)	SZ6	0.26
Diamantopoulou et al., 2011	14	NR	NR	NR	NR	—	1.73	SZ3 - SZ2	0.61
Drew and Vogel, 2008, Exp 4	33	NR	—	NR	—	—	NR	SZ3 - SZ1	0.48
Drew and Vogel, 2008, Exp 5	18	NR	—	NR	—	—	NR	SZ3 - SZ1	0.72
Feldmann-Wüstefeld et al., 2021	21	—	NR	—	NR	NR	2.5	Mean of SZ2 and SZ4;SZ4 - SZ2	0.43; 0.39
Jost et al., 2011 (Young)	25	NR	—	2.14	—	—	NR	NR	0.48
Kang and Woodman, 2014	24	NR	NR	—	NR	NR	2.59	SZ4 - SZ1	0.62
Kundu et al., 2013	30	—	NR	—	NR	—	2.16	NR	0.62
Kuo et al., 2012, Exp1	18	—	1.53 (0.16)	—	2.66 (0.58)	—	NR	SZ4 - SZ2	0.63
Lefebvre et al., 2013	39	—	1.9	—	3.3	3.75	NR	SZ6 - SZ2	0.52
Leonard et al., 2013	23	NR	—	NR	—	—	NR	SZ3 - SZ1	0.59
Stormer et al., 2013 (Young)	35	NR	—	2.1	—	—	NR	SZ3 - SZ1	0.40
Tröndle & Langer, 2024 (Young)	55	0.96 (0.05)	—	2.26 (0.42)	—	—	—	SZ3	0.24
Tsubomi et al., 2013	25	—	NR	—	NR	NR	NR	SZ4 - SZ2	0.64

Unsworth et al., 2015	170	—	NR	—	NR	1.90 (0.77)	NR	SZ6	0.30
Villena-Gonzalez et al., 2020	23	NR	NR	—	NR	—	2.3 (0.8)	SZ4 - SZ2	0.45
Vogel and Machizawa, 2004	36	NR	NR	NR	NR	NR	2.8	SZ4 - SZ2	0.78

Note. SZ = set size. NR = not reported. R = Pearson's correlation coefficient. SD = standard deviation. SD values are shown only when they were reported in the publication. Often, publications do not report K scores separately for each investigated set size, which is indicated as NR in the table. An em dash (—) indicates that a specific set size was not investigated in the given study.

Acknowledgements

#EEGManyLabs is supported by a DFG (PA 4005/1-1) grant to YGP and a UKRI BBSRC grant (BB/X008428/1) awarded to FM. The Lead Author, NL, is funded by the Swiss National Science Foundation (SNSF) Grant 100014_175875. HMS, HD, and AL are supported by the Icelandic Research Fund (228916-051) and the University of Iceland Research Fund. CCvB and SJ are supported by the Economic and Social Research Council (UK; ES/V013610/1). MGM is supported by the Moonshot R&D Goal 9 (JPMJMS2296) and JST COI (JPMJCE1311, JPMJCA2208). PEC is funded by a grant from the National Institute of Mental Health (MH128208). These funding sources were not involved in data analysis, interpretation, or the preparation of this submission for publication.

Competing Interests

The authors disclose no conflicts of interest related to this manuscript.

Author Contributions

Conceptualization: D.S., F.M., Y.G.P., M.G.M., W.X.Q.N., E.K.V., and N.L.

Data curation: D.S. and N.L.

Formal analysis: D.S. and N.L.

Funding acquisition: P.E.C., H.M.S., F.M., Y.G.P., V.S.S., A.-L.S., J.S.J., J.D.G., C.C.v.B., N.A.B., and N.L.

Investigation: D.S., P.E.C., H.M.S., H.D., A.L., H.A.R., Y.H.C., K.M.O., V.S.S., J.C.G.A., C.L., A.-L.S., A.-L.B., S.A.B., E.M.J., J.S.J., Z.L., Y.M.C., E.L., J.D.G., S.J., M.J., E.M., C.C.v.B., N.A.B., C.P., L.B., Y.H., W.X.Q.N., and N.L.

Methodology: F.M., Y.G.P., K.M.O., M.G.M., W.X.Q.N., E.K.V., and N.L.

Project administration: F.M., Y.G.P., and N.L.

Resources: D.S., P.E.C., H.M.S., F.M., Y.G.P., V.S.S., A.-L.S., J.S.J., J.D.G., C.C.v.B., N.A.B., W.X.Q.N., and N.L.

Software: D.S., W.X.Q.N., and N.L.

Supervision: P.E.C., H.M.S., F.M., Y.G.P., V.S.S., A.-L.S., J.S.J., J.D.G., C.C.v.B., N.A.B., and N.L.

Validation: D.S., F.M., Y.G.P., and N.L.

Visualization: D.S., H.M.S., and N.L.

Writing - original draft: D.S. and N.L.

Writing - review & editing: D.S., P.E.C., H.M.S., F.M., Y.G.P., H.D., A.L., H.A.R., Y.H.C., K.M.O., V.S.S., J.C.G.A., C.L., A.-L.S., A.-L.B., S.A.B., E.M.J., J.S.J., Z.L., Y.M.C., E.L., J.D.G., S.J., M.J., E.M., C.C.v.B., N.A.B., C.P., L.B., Y.H., M.G.M., W.X.Q.N., E.K.V., and N.L.

References

Adam, K. C. S., Robison, M. K., & Vogel, E. K. (2018). Contralateral delay activity tracks fluctuations in working memory performance. *Journal of Cognitive Neuroscience*, *30*(9), 1229–1240. https://doi.org/10.1162/jocn_a_01233

Asp, I. E., Störmer, V. S., & Brady, T. F. (2021). Greater visual working memory capacity for visually matched stimuli when they are perceived as meaningful. *Journal of Cognitive Neuroscience*, 1–17. https://doi.org/10.1162/jocn_a_01693

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In Psychology of Learning and Motivation (pp. 89–195). Elsevier.

Baddeley, A. (2003). Working memory: looking back and looking forward. Nature Reviews. Neuroscience, 4(10), 829–839.

Baddeley, A. D. (1986). Working Memory. OUP Australia and.

Baddeley, A. D., Gathercole, S. E., & Papagno, C. (2017). The phonological loop as a language learning device. In Exploring Working Memory (pp. 164–198). Routledge.

Baddeley, A. D., & Hitch, G. (1974). Working Memory. In Psychology of Learning and Motivation (pp. 47–89). Elsevier.

Baddeley, A., Papagno, C., & Vallar, G. (1988). When long-term learning depends on short-term storage. Journal of Memory and Language, 27(5), 586–595.

Baldwin, S. A., Larson, M. J., & Clayson, P. E. (2015). The dependability of electrophysiological measurements of performance monitoring in a clinical sample: A generalizability and decision analysis of the ERN and Pe: EEG dependability. *Psychophysiology*, *52*(6), 790–800. https://doi.org/10.1111/psyp.12401

Brady, T. F., Robinson, M. M., Williams, J. R., & Wixted, J. T. (2023). Measuring memory is harder than you think: How to avoid problematic measurement practices in memory research. *Psychonomic Bulletin & Review*, *30*(2), 421–449. https://doi.org/10.3758/s13423-022-02179-w

Brady, T. F., Störmer, V. S., & Alvarez, G. A. (2016). Working memory is not fixed-capacity:

More active storage capacity for real-world objects than for simple stimuli. *Proceedings of the National Academy of Sciences of the United States of America*, *113*(27), 7459–7464. https://doi.org/10.1073/pnas.1520027113

Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. https://doi.org/10.1163/156856897x00357

Brennan, R. L. (1992). Generalizability theory. *Educational Measurement Issues and Practice*, *11*(4), 27–34. https://doi.org/10.1111/j.1745-3992.1992.tb00260.x

Brisson, B., & Jolicoeur, P. (2007). A psychological refractory period in access to visual short-term memory and the deployment of visual-spatial attention: multitasking processing deficits revealed by event-related potentials. *Psychophysiology*, *44*(2), 323–333. https://doi.org/10.1111/j.1469-8986.2007.00503.x

Bürkner, P.-C. (2017). Brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1). https://doi.org/10.18637/jss.v080.i01

Caldwell, A. R. (2022). Exploring equivalence testing with the updated TOSTER R package. In *PsyArXiv*. https://doi.org/10.31234/osf.io/ty8de

Caplan, D., & Waters, G. S. (1999). Verbal working memory and sentence comprehension. The Behavioral and Brain Sciences, 22(1), 77–94; discussion 95–126.

Champely, S. (2020). *Basic Functions for Power Analysis [R package pwr version 1.3-0]*. https://CRAN.R-project.org/package=pwr

Clayson, P. E., Baldwin, S. A., & Larson, M. J. (2021). Evaluating the internal consistency of subtraction-based and residualized difference scores: Considerations for psychometric reliability analyses of event-related potentials. *Psychophysiology*, *58*(4), e13762. https://doi.org/10.1111/psyp.13762

Clayson, P. E., Brush, C. J., & Hajcak, G. (2021). Data quality and reliability metrics for event-related potentials (ERPs): The utility of subject-level reliability. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *165*, 121–136. https://doi.org/10.1016/j.ijpsycho.2021.04.004

Clayson, P. E., Carbine, K. A., Baldwin, S. A., & Larson, M. J. (2019). Methodological

reporting behavior, sample sizes, and statistical power in studies of event-related potentials: Barriers to reproducibility and replicability. *Psychophysiology*, *56*(11), e13437. https://doi.org/10.1111/psyp.13437

Clayson, P. E., Carbine, K. A., Baldwin, S. A., Olsen, J. A., & Larson, M. J. (2021). Using generalizability theory and the ERP Reliability Analysis (ERA) Toolbox for assessing test-retest reliability of ERP scores part 1: Algorithms, framework, and implementation. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *166*, 174–187. https://doi.org/10.1016/j.ijpsycho.2021.01.006

Clayson, P. E., & Miller, G. A. (2017). Psychometric considerations in the measurement of event-related brain potentials: Guidelines for measurement and reporting. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *111*, 57–67. https://doi.org/10.1016/j.ijpsycho.2016.09.005

Cowan, N. (2001). The magical number 4 in short-term memory: a reconsideration of mental storage capacity. *The Behavioral and Brain Sciences*, *24*(1), 87–114; discussion 114–185. https://doi.org/10.1017/s0140525x01003922

Cowan, N., Elliott, E. M., Scott Saults, J., Morey, C. C., Mattox, S., Hismjatullina, A., & Conway, A. R. A. (2005). On the capacity of attention: its estimation and its role in working memory and cognitive aptitudes. *Cognitive Psychology*, *51*(1), 42–100. https://doi.org/10.1016/j.cogpsych.2004.12.001

Cowan, N., & Morey, C. C. (2006). Visual working memory depends on attentional filtering. Trends in Cognitive Sciences, 10(4), 139–141.

Cowan, N. (2014). Working memory underpins cognitive development, learning, and education. Educational Psychology Review, 26(2), 197–223.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. Journal of Verbal Learning and Verbal Behavior, 19(4), 450–466.

de Cheveigné, A. (2020). ZapLine: A simple and effective method to remove power line artifacts. *NeuroImage*, *207*, 116356. https://doi.org/10.1016/j.neuroimage.2019.116356

Delorme, A., & Makeig, S. (2004). EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, *134*(1), 9–21. https://doi.org/10.1016/j.jneumeth.2003.10.009

Diamantopoulou, S., Poom, L., Klaver, P., & Talsma, D. (2011). Visual working memory capacity and stimulus categories: a behavioral and electrophysiological investigation. *Experimental Brain Research*, *209*(4), 501–513. https://doi.org/10.1007/s00221-011-2536-z

Drew, T., & Vogel, E. K. (2008). Neural measures of individual differences in selecting and tracking multiple moving objects. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *28*(16), 4183–4191. https://doi.org/10.1523/JNEUROSCI.0556-08.2008

Duval, S., & Tweedie, R. (2000). Trim and fill: A simple funnel-plot-based method of testing and adjusting for publication bias in meta-analysis. *Biometrics*, *56*(2), 455–463. https://doi.org/10.1111/j.0006-341x.2000.00455.x

Emrich, S. M., Al-Aidroos, N., Pratt, J., & Ferber, S. (2009). Visual search elicits the electrophysiological marker of visual working memory. *PloS One*, *4*(11), e8042. https://doi.org/10.1371/journal.pone.0008042

Feldmann-Wüstefeld, T. (2021). Neural measures of working memory in a bilateral change detection task. *Psychophysiology*, *58*(1), e13683. https://doi.org/10.1111/psyp.13683

Feldmann-Wüstefeld, T., Vogel, E. K., & Awh, E. (2018). Contralateral delay activity indexes working memory storage, not the current focus of spatial attention. *Journal of Cognitive Neuroscience*, *30*(8), 1185–1196. https://doi.org/10.1162/jocn_a_01271

Feuerstahler, L. M., Luck, S. J., MacDonald, A., 3rd, & Waller, N. G. (2019). A note on the identification of change detection task models to measure storage capacity and attention in visual working memory. *Behavior Research Methods*, *51*(3), 1360–1370. https://doi.org/10.3758/s13428-018-1082-z

Forsberg, A., Adams, E. J., & Cowan, N. (2023). Why does visual working memory ability improve with age: More objects, more feature detail, or both? A registered report.

Developmental Science, *26*(2), e13283. https://doi.org/10.1111/desc.13283

Fukuda, K., & Vogel, E. K. (2011). Individual differences in recovery time from attentional capture. *Psychological Science*, *22*(3), 361–368. https://doi.org/10.1177/0956797611398493

Garrett-Ruffin, S., Hindash, A. C., Kaczkurkin, A. N., Mears, R. P., Morales, S., Paul, K., Pavlov, Y. G., & Keil, A. (2021). Open science in psychophysiology: An overview of challenges and emerging solutions. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *162*, 69–78. https://doi.org/10.1016/j.ijpsycho.2021.02.005

Gathercole, S. E., & Baddeley, A. D. (1989). Evaluation of the role of phonological STM in the development of vocabulary in children: A longitudinal study. Journal of Memory and Language, 28(2), 200–213.

Gelman, A. (2007). Struggles with survey weighting and regression modeling. *Statistical Science: A Review Journal of the Institute of Mathematical Statistics*, *22*(2), 153–164. https://doi.org/10.1214/088342306000000691

Gelman, A., & Loken, E. (2023). *The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time*. Retrieved August 21, 2023, from http://stat.columbia.edu/~gelman/research/unpublished/forking.pdf

Götz, M., Sarma, A., & O'Boyle, E. H. (2024). The multiverse of universes: A tutorial to plan, execute and interpret multiverses analyses using the R package multiverse. *International Journal of Psychology: Journal International de Psychologie*, *59*(6), 1003–1014. https://doi.org/10.1002/ijop.13229

Hakim, N., Adam, K. C. S., Gunseli, E., Awh, E., & Vogel, E. K. (2019). Dissecting the neural focus of attention reveals distinct processes for spatial attention and object-based storage in visual working memory. *Psychological Science*, *30*(4), 526–540. https://doi.org/10.1177/0956797619830384

Hakim, N., Feldmann-Wüstefeld, T., Awh, E., & Vogel, E. K. (2020). Perturbing neural

representations of working memory with task-irrelevant interruption. *Journal of Cognitive Neuroscience*, *32*(3), 558–569. https://doi.org/10.1162/jocn_a_01481

Hakim, N., Feldmann-Wüstefeld, T., Awh, E., & Vogel, E. K. (2021). Controlling the flow of distracting information in working memory. *Cerebral Cortex (New York, N.Y.: 1991)*, *31*(7), 3323–3337. https://doi.org/10.1093/cercor/bhab013

Heuer, A., & Schubö, A. (2016). The focus of attention in visual working memory: Protection of focused representations and its individual variation. *PloS One*, *11*(4), e0154228. https://doi.org/10.1371/journal.pone.0154228

Jennings, J. R., & Wood, C. C. (1976). Letter: The epsilon-adjustment procedure for repeated-measures analyses of variance. *Psychophysiology*, *13*(3), 277–278. https://doi.org/10.1111/j.1469-8986.1976.tb00116.x

Jongbloed-Pereboom, M., Nijhuis-van der Sanden, M. W. G., & Steenbergen, B. (2019). Explicit and implicit motor sequence learning in children and adults; the role of age and visual working memory. *Human Movement Science*, *64*, 1–11. https://doi.org/10.1016/j.humov.2018.12.007

Kang, M.-S., & Woodman, G. F. (2014). The neurophysiological index of visual working memory maintenance is not due to load dependent eye movements. *Neuropsychologia*, *56*, 63–72. https://doi.org/10.1016/j.neuropsychologia.2013.12.028

Klaver, P., Talsma, D., Wijers, A. A., Heinze, H. J., & Mulder, G. (1999). An event-related brain potential correlate of visual short-term memory. *Neuroreport*, *10*(10), 2001–2005. https://doi.org/10.1097/00001756-199907130-00002

Klug, M., & Kloosterman, N. A. (2022). Zapline-plus: A Zapline extension for automatic and adaptive removal of frequency-specific noise artifacts in M/EEG. *Human Brain Mapping*, *43*(9), 2743–2758. https://doi.org/10.1002/hbm.25832

Kothe, C. A., & Makeig, S. (2013). BCILAB: a platform for brain-computer interface development. *Journal of Neural Engineering*, *10*(5), 056014. https://doi.org/10.1088/1741-2560/10/5/056014

Kundu, B., Sutterer, D. W., Emrich, S. M., & Postle, B. R. (2013). Strengthened effective

connectivity underlies transfer of working memory training to tests of short-term memory and attention. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *33*(20), 8705–8715. https://doi.org/10.1523/JNEUROSCI.5565-12.2013

Kuo, B.-C., Stokes, M. G., & Nobre, A. C. (2012). Attention modulates maintenance of representations in visual short-term memory. *Journal of Cognitive Neuroscience*, *24*(1), 51–60. https://doi.org/10.1162/jocn_a_00087

Lakens, D. (2017). TOSTER: Two one-sided tests (TOST) equivalence testing. *R Package Version 0. 2*, *5*, 648.

Lefebvre, C., Vachon, F., Grimault, S., Thibault, J., Guimond, S., Peretz, I., Zatorre, R. J., & Jolicœur, P. (2013). Distinct electrophysiological indices of maintenance in auditory and visual short-term memory. *Neuropsychologia*, *51*(13), 2939–2952. https://doi.org/10.1016/j.neuropsychologia.2013.08.003

Leonard, C. J., Kaiser, S. T., Robinson, B. M., Kappenman, E. S., Hahn, B., Gold, J. M., & Luck, S. J. (2013). Toward the neural mechanisms of reduced working memory capacity in schizophrenia. *Cerebral Cortex (New York, N.Y.: 1991)*, *23*(7), 1582–1592. https://doi.org/10.1093/cercor/bhs148

Liesefeld, H. R., & Müller, H. J. (2019). Current directions in visual working memory research: An introduction and emerging insights. *British Journal of Psychology (London, England: 1953)*, *110*(2), 193–206. https://doi.org/10.1111/bjop.12377

Lotfi, S., Ward, R., Mathew, A., Shokoohi-Yekta, M., Rostami, R., Motamed-Yeganeh, N., Christine, C., & Lee, H.-J. (2022). Limited visual working memory capacity in children with dyslexia: An ERP study. *NeuroRegulation*, *9*(2), 98–109. https://doi.org/10.15540/nr.9.2.98

Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd ed.). Bradford Books.

Luck, S. J., & Gaspelin, N. (2017). How to get statistically significant effects in any ERP experiment (and why you shouldn't). *Psychophysiology*, *54*(1), 146–157. https://doi.org/10.1111/psyp.12639

Luck, S. J., Stewart, A. X., Simmons, A. M., & Rhemtulla, M. (2021). Standardized measurement error: A universal metric of data quality for averaged event-related potentials. *Psychophysiology*, *58*(6), e13793. https://doi.org/10.1111/psyp.13793

Luck, S. J., & Vogel, E. K. (2013). Visual working memory capacity: from psychophysics and neurobiology to individual differences. *Trends in Cognitive Sciences*, *17*(8), 391–400. https://doi.org/10.1016/j.tics.2013.06.006

Luria, R., Balaban, H., Awh, E., & Vogel, E. K. (2016). The contralateral delay activity as a neural measure of visual working memory. *Neuroscience and Biobehavioral Reviews*, *62*, 100–108. https://doi.org/10.1016/j.neubiorev.2016.01.003

McCollough, A. W., Machizawa, M. G., & Vogel, E. K. (2007). Electrophysiological measures of maintaining representations in visual working memory. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *43*(1), 77–94. https://doi.org/10.1016/s0010-9452(08)70447-7

Martin, R. C., & Romani, C. (1994). Verbal working memory and sentence comprehension: A multiple-components view. Neuropsychology, 8(4), 506–523.

Miyake, A., & Shah, P. (Eds.). (1999). *Models of working memory: Mechanisms of active maintenance and executive control* (Vol. 506). Cambridge University Press. https://doi.org/10.1017/cbo9781139174909

Naveh-Benjamin, M., & Cowan, N. (2023). The roles of attention, executive function and knowledge in cognitive ageing of working memory. *Nature Reviews Psychology*, *2*(3), 151–165. https://doi.org/10.1038/s44159-023-00149-0

Ngiam, W. X. Q., Adam, K. C. S., Quirk, C., Vogel, E. K., & Awh, E. (2021). Estimating the statistical power to detect set-size effects in contralateral delay activity. *Psychophysiology*, *58*(5), e13791. https://doi.org/10.1111/psyp.13791

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

Oldfield, R. C. (1971). The assessment and analysis of handedness: the Edinburgh inventory. *Neuropsychologia*, *9*(1), 97–113. https://doi.org/10.1016/0028-3932(71)90067-4

Olivers, C. N. L. (2008). Interactions between visual working memory and visual attention. Frontiers in Bioscience, 13(13), 1182–1191.

Pashler, H. (1988). Familiarity and visual change detection. *Perception & Psychophysics*, *44*(4), 369–378. https://doi.org/10.3758/bf03210419

Paul, M., Govaart, G. H., & Schettino, A. (2021). Making ERP research more transparent: Guidelines for preregistration. *International Journal of Psychophysiology: Official Journal of the International Organization of Psychophysiology*, *164*, 52–63. https://doi.org/10.1016/j.ijpsycho.2021.02.016

Pavlov, Y. G., Adamian, N., Appelhoff, S., Arvaneh, M., Benwell, C. S. Y., Beste, C., Bland, A. R., Bradford, D. E., Bublatzky, F., Busch, N. A., Clayson, P. E., Cruse, D., Czeszumski, A., Dreber, A., Dumas, G., Ehinger, B., Ganis, G., He, X., Hinojosa, J. A., … Mushtaq, F. (2021). #EEGManyLabs: Investigating the replicability of influential EEG experiments. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *144*, 213–229. https://doi.org/10.1016/j.cortex.2021.03.013

Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision*, *10*(4), 437–442. https://doi.org/10.1163/156856897x00366

Perron, R., Lefebvre, C., Robitaille, N., Brisson, B., Gosselin, F., Arguin, M., & Jolicoeur, P. (2009). Attentional and anatomical considerations for the representation of simple stimuli in visual short-term memory: evidence from human electrophysiology. *Psychological Research*, *73*(2), 222–232. https://doi.org/10.1007/s00426-008-0214-y

Quirk, C., Adam, K. C. S., & Vogel, E. K. (2020). No evidence for an object working memory capacity benefit with extended viewing time. *eNeuro*, *7*(5), ENEURO.0150–20.2020. https://doi.org/10.1523/ENEURO.0150-20.2020

Roy, Y., & Faubert, J. (2023). Is the Contralateral Delay Activity (CDA) a robust neural correlate for Visual Working Memory (VWM) tasks? A reproducibility study. *Psychophysiology*, *60*(2), e14180. https://doi.org/10.1111/psyp.14180

Sarma, A., Hwang, K., Hullman, J., & Kay, M. (2024). Milliways: Taming multiverses through

principled evaluation of data analysis paths. *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–15. https://doi.org/10.1145/3613904.3642375

Schneider, D., Barth, A., Getzmann, S., & Wascher, E. (2017). On the neural mechanisms underlying the protective function of retroactive cuing against perceptual interference: Evidence by event-related potentials of the EEG. *Biological Psychology*, *124*, 47–56. https://doi.org/10.1016/j.biopsycho.2017.01.006

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*(5), 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Shavelson, R. J., & Webb, N. M. (2012). *Generalizability Theory*. SAGE Publications.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2015). Specification curve: Descriptive and inferential statistics on all reasonable specifications. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2694998

Sterne, J. A. C., & Egger, M. (2001). Funnel plots for detecting bias in meta-analysis. *Journal of Clinical Epidemiology*, *54*(10), 1046–1055. https://doi.org/10.1016/S0895-4356(01)00377-8

Störmer, V. S., Li, S.-C., Heekeren, H. R., & Lindenberger, U. (2013). Normative shifts of cortical mechanisms of encoding contribute to adult age differences in visual-spatial working memory. *NeuroImage*, *73*, 167–175. https://doi.org/10.1016/j.neuroimage.2013.02.004

Strzelczyk, D., Clayson, P. E., Sigurdardottir, H. M., Mushtaq, F., Pavlov, Y. G., Devillez, H., Lukashevich, A., Rocha, H. A., Chung, Y. H., Ortego, K. M., St�rmer, V. S., Garc�a Alanis, J. C., L�ffler, C., Schubert, A.-L., Biel, A. L., Tretow, A., Xu, W., Hamalainen, J.,

Lu, Z., … et al. (2023). Contralateral delay activity as a marker of visual working memory capacity: a multi-site registered replication. In *PsyArXiv Preprints* (Issue shdea). University of Zurich. https://scholar.google.com/citations?view_op=view_citation&hl=en&citation_for_view=xHS491AAAAAJ:9yKSN-GCB0IC

Sundre, D. L. (1993). Book reviews : Generalizability theory: A primer, by Richard J. shavelson and Noreen M. webb. Newbury Park, CA: Sage publications, 1991,137 pp. *Evaluation Practice*, *14*(2), 207–209. https://doi.org/10.1177/109821409301400219

Tröndle, M., & Langer, N. (2024). Decomposing neurophysiological underpinnings of age-related decline in visual working memory. *Neurobiology of Aging*, *139*, 30–43. https://doi.org/10.1016/j.neurobiolaging.2024.03.004

Trübutschek, D., Yang, Y., Gianelli, C., Cesnaite, E., Fischer, N. L., Vinding, M. C., Marshall, T. R., Algermissen, J., Pascarella, A., Puoliväli, T., Vitale, A., Busch, N. A., & Nilsonne, G. (2023). EEGManyPipelines: a large-scale, grassroots multi-analyst study of electroencephalography analysis practices in the wild. Journal of Cognitive Neuroscience, 36(2), 217–224. https://doi.org/10.1162/jocn_a_02087

Tsubomi, H., Fukuda, K., Watanabe, K., & Vogel, E. K. (2013). Neural limits to representing objects still within view. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *33*(19), 8257–8263. https://doi.org/10.1523/JNEUROSCI.5348-12.2013

Unsworth, N., Fukuda, K., Awh, E., & Vogel, E. K. (2015). Working memory delay activity predicts individual differences in cognitive abilities. *Journal of Cognitive Neuroscience*, *27*(5), 853–865. https://doi.org/10.1162/jocn_a_00765

Viechtbauer, W. (2010). Conducting Meta-Analyses in R with the metafor Package. *Journal of Statistical Software*, *36*(3). https://doi.org/10.18637/jss.v036.i03

Villena-González, M., Rubio-Venegas, I., & López, V. (2020). Data from brain activity during visual working memory replicates the correlation between contralateral delay activity and memory capacity. *Data in Brief*, *28*(105042), 105042.

https://doi.org/10.1016/j.dib.2019.105042

Vogel, E. K., Luck, S. J., & Shapiro, K. L. (1998). Electrophysiological evidence for a postperceptual locus of suppression during the attentional blink. *Journal of Experimental Psychology. Human Perception and Performance*, *24*(6), 1656–1674. https://doi.org/10.1037//0096-1523.24.6.1656

Vogel, E. K., & Machizawa, M. G. (2004). Neural activity predicts individual differences in visual working memory capacity. *Nature*, *428*(6984), 748–751. https://doi.org/10.1038/nature02447

von Bastian, C. C., Belleville, S., Udale, R. C., Reinhartz, A., Essounni, M., & Strobach, T. (2022). Mechanisms underlying training-induced cognitive change. *Nature Reviews Psychology*, *1*(1), 30–41. https://doi.org/10.1038/s44159-021-00001-3

Wang, J., Huo, S., Wu, K. C., Mo, J., Wong, W. L., & Maurer, U. (2022). Behavioral and neurophysiological aspects of working memory impairment in children with dyslexia. *Scientific Reports*, *12*(1), 12571. https://doi.org/10.1038/s41598-022-16729-8

Westheimer, G. (1954a). Eye movement responses to a horizontally moving visual stimulus. *A.M.A. Archives of Ophthalmology*, *52*(6), 932–941. https://doi.org/10.1001/archopht.1954.00920050938013

Westheimer, G. (1954b). Mechanism of saccadic eye movements. *A.M.A. Archives of Ophthalmology*, *52*(5), 710–724. https://doi.org/10.1001/archopht.1954.00920050716006

Widmann, A., & Schröger, E. (2012). Filter effects and filter artifacts in the analysis of electrophysiological data. *Frontiers in Psychology*, *3*, 233. https://doi.org/10.3389/fpsyg.2012.00233

Wilkinson, G. N., & Rogers, C. E. (1973). Symbolic description of factorial models for analysis of variance. *Journal of the Royal Statistical Society. Series C, Applied Statistics*, *22*(3), 392.

Williams, J. R., Robinson, M. M., Schurgin, M. W., Wixted, J. T., & Brady, T. F. (2022). You cannot “count” how many items people remember in visual working memory: The

importance of signal detection-based measures for understanding change detection performance. *Journal of Experimental Psychology. Human Perception and Performance*, *48*(12), 1390–1409. https://doi.org/10.1037/xhp0001055

Scientific transparency statement

DATA: All raw and processed data supporting this research are publicly available: https://gin.g-node.org/EEGManyLabs/EEGManyLabs_Replication_VogelMachizawa2004_Raw, https://gin.g-node.org/EEGManyLabs/EEGManyLabs_Replication_VogelMachizawa2004_Processed

CODE: All analysis code supporting this research is publicly available: https://github.com/ksgfan/EEGManyLabs

MATERIALS: All study materials supporting this research are publicly available: https://github.com/ksgfan/EEGManyLabs

DESIGN: This article reports, for all studies, how the author(s) determined all sample sizes, all data exclusions, all data inclusion and exclusion criteria, and whether inclusion and exclusion criteria were established prior to data analysis.

PRE-REGISTRATION: At least part of the study procedures was pre-registered in a time-stamped, institutional registry prior to the research being conducted: https://doi.org/10.31234/osf.io/shdea At least part of the analysis plans was pre-registered in a time-stamped, institutional registry prior to the research being conducted: https://doi.org/10.31234/osf.io/shdea The analyses that were undertaken deviated from the preregistered analysis plans. All such deviations are fully disclosed in the manuscript.

For full details, see the *Scientific Transparency Report* in the supplementary data to the online version of this article.