UNIVERSITY OF CALIFORNIA
RIVERSIDE

Task and Stimulus Processes May Explain Diverse/Inconsistent Working Memory Training Outcomes

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Psychology

by

Mahsa Alizadeh Shalchy

September 2021

Dissertation Committee:
Dr. Aaron R. Seitz, Chairperson
Dr. Megan A. K. Peters
Dr. Steven E. Clark

The Dissertation of Mahsa Alizadeh Shalchy is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

Dedication and Acknowledgements

To begin with, I would like to thank my family, who bear the pain of me being far away from home for years so that I can pursue my passion of studying the human mind, brain, and behavior. To my mom, Laya Haddadi, who has taught me to be strong, couragous, and compassionate. To my dad, Mehdi Alizadeh Shalchy, who has taught me the beauty of mathematics, how to fix electrical devices and to never give up. To my aunt, Fateme Alizadeh Shalchy, who has taught me the joy of learning through playing games and solving puzzles. To my cousin, Dr. Nastaran Ghiasvand who has been supporting me like a sister for as long as I can remember; completing my degree wouldn't be possible without her mental and emotional support. I am forever indebted to my amazing family and I dedicate this work to the first teachers of my life, my family.

Foremost, I owe my deepest gratitude to my graduate teacher, mentor, supervisor and committee chair Dr. Aaron Seitz who taught me to become an independent researcher from my very first day in the program. Aaron taught me the beauty of electrophysiological and neurophysiological signals and he has introduced me to novel tools to investigate data. He has taught me to seek the true meaning of psychological concepts and mathematical/statistical tools, and to understand every step of experiment design, data collection and data processing rather than being a mere button-pressing operator. His invaluable and genius insight on making scientific inquiries has shaped my research and his emphasis on "every hypothesis is a model" has pushed me to be a better consumer and producer of scientific research. I could not have asked for a better supervisor with the

multidisciplinary research to quench my thirst for understanding human cognition and acquiring a versatile skill set. Finally, I am grateful that despite his busy schedule, he continuously supported my research by making time for regular meetings/chats even during the weekends and holidays.

I am grateful to my committee member, Dr. Megan Peters for the valuable inputs throughout my PhD. I especially learned a lot from Dr. Peters regarding computational modeling and perceptual decision-making. I will never forget her modeling class where, for the first time, I fit a diffusion model to an audiovisual experiment, and I was in awe of understanding the parameters. I am grateful to my committee member, Dr. Steven Clark who taught me to appreciate the intricacy of signal detection theory and further my understanding of memory models, eye-witness memory and impact of stress on memory. I will never forget the fun and wit in Dr. Clark's classes.

I am also grateful to Dr. Weiwei Zhang who has shaped my understanding of working memory, cognitive processes, and eye-tracking methods. He has been the most outstanding international advisor one could ask for, supporting me as an international student throughout graduate school. I am also grateful to Dr. Illana Bennet for her insights regarding memory processing, impact of aging on memory and neuroimaging techniques. I had the honor of having Drs. Peters, Zhang and Bennett as my qualifying exam and/or second-year project committee who provided me with great feedback.

I would like to express my gratitude to Dr. Edward Zagha for his amazing graduate and undergraduate classes that taught me a lot regarding neuroscience.

I would like to also extend my gratitude to Dr. Robert Rosenthal and Dan Ozer for giving me deep insight into statistical modeling. I would like to thank Dr. Kate Sweeny for being a great and hands-on graduate advisor. I would like to thank Ms. Vanessa Lee and Ms. Nina Mandracchia for their valuable help with my professional development. I would like to thank Ms. Kirsten Susan Alonso and Ms. Sarah Turnbull for their help with various logistic processes during my PhD. I would also like to thank the experts of the Center for Advanced Neuroimaging at UCR for our collaborations and all their efforts in collecting neuroimaging data possible: Dr. Xiaoping Hu, Ms. Chelsea Savina Evelyn, Dr. Xu Chen and Dr. Jason Langley.

A part of this work was supported by a grant from NASA MIRO, and I would like to thank Dr. Bahram Mobasher, Xinnan Du and the NASA team for this amazing opportunity to be a part of the BIG DATA team and meeting with amazing researchers.

A special thanks to my amazing lab-mates and RAs at the Perception and Learning Lab and UCR's Brain Game Center who have helped me with providing feedback and insight throughout various stages of my research.

Finally, I would like to acknowledge that I am using previously published material in my thesis document as follows: Chapter 1 was authored by Pergher, V., Shalchy, M. A., Pahor, A., Van Hulle, M. M., Jaeggi, S. M., & Seitz, A. R. (2020). Chapter 2 was authored by Shalchy, M. A., Pergher, V., Pahor, A., Van Hulle, M. M., & Seitz, A. R. (2020). These previously published material appeared in the journals as follow: a) Chapter 1 was published in 2019 in the Journal of Cognitive Enhancement, 4(1), 100-120.

https://doi.org/10.1007/s41465-019-00134-7

b) Chapter 2 was published in 2020 in Frontiers in human neuroscience, 14, 453

https://doi.org/10.3389/fnhum.2020.549966

These materials will appear in this dissertation as follows:

a) Refers to Chapter 1, b) Refers to Chapter 2

The entire articles were used with the necessary parts being re-written to adhere to the general format of the dissertation.

ABSTRACT OF THE DISSERTATION

Task and Stimulus Processes May Explain Diverse/Inconsistent Working Memory Training Outcomes

by

Mahsa Alizadeh Shalchy

Doctor of Philosophy, Graduate Program in Psychology
University of California, Riverside, September 2021
Dr. Aaron R. Seitz, Chairperson

Working Memory (WM) is a fundamental cognitive ability that encodes, manipulates, and maintains information for a brief amount of time. WM is involved in vital daily functions such as reasoning, problem-solving, and learning. Thus, there has been an increasing interest in enhancing WM by the use of "training" interventions. As such, many researchers have utilized various experimental paradigms in the form of varied tasks and stimuli to train WM. On one hand, the training results report benefits of WM training and transfer of these beneficial effects to domains similar to WM as well as different domains. On the other hand, some studies fail to observe beneficial effects. These inconsistent results have brought controversy, dividing researchers into believers and non-believers of WM training interventions. This is a general problem as many studies in the WM field assume a straightforward relationship between the construct of WM and its measurement through any of the variety of available WM measures, such as N-back.

In this dissertation, we aim to further our understanding of WM by investigating WM through distinct but complementary lenses. In chapter 1, we focus on WM training

studies that use N-back, a popular training task, to understand the efficacy of WM training interventions. In this chapter, we aim to point out that the features of training/assessment tasks and stimuli differ on many levels across different studies of WM, and that combining the results of these studies would be similar to mixing apples and oranges. In chapter 2, we bring mechanistic understanding by using brain signals as important mediators of behavior. Our goal is to not only understand the early and late brain mechanisms during various tasks and stimuli, but also to investigate other potential factors that can produce inconsistent results. After investigating the existing experimental paradigms and related brain signals, in chapter 3 we combine our knowledge of chapter 1 and chapter 2 by doing a multi-measure experiment. Our goal is to set the stage for studying the cortical arousal system, a key influencer of memory processes and learning through altering brain states. In conclusion, this dissertation furthers our understanding of WM, its training efficacy, and factors impacting it.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

General Introduction

Working Memory (WM) is defined as a limited capacity system responsible for temporary storage and manipulation of relevant information over a limited time (Baddley, 2012). WM has been studied extensively over the last few decades due to the fact that it correlates with a wide range of complex cognitive abilities such as problem-solving, reasoning, learning and planning of goal-directed behaviors (Miyake, & Shah, 1999; Swanson & Alloway, 2012). Due to its distributed site of functioning (Christophel et al., 2017) and involvement in multiple processes such as encoding of information, maintenance, and retrieval (D'Esposito, & Postle, 2015), many researchers suggest that WM is the "hub" of cognition. Given its crucial role in many cognitive abilities, neurological disorders affecting WM, such as depression (Rose & Ebmeier, 2006), schizophrenia (Frydecka et al., 2014), attention-deficit hyperactivity disorder (ADHD; Arjona et al., 2020), Alzheimer's disease (AD; Zokaei & Husain, 2019), and learning difficulties (Swanson & Seigel, 2011), can seriously impact patients' lives.

Thus, there has been a great interest in enhancing WM by use of "training" interventions (Anguera et al., 2012): numerous studies (Blacker et al., 2017; Minear et al., 2016) have trained participants using a variety of WM tasks such as N-back, span tasks, immediate recall, etc. (see Pergher, Shalchy, Pahor et al., 2019). While most studies find improvements in the training task, it is controversial whether this gain transfers to similar WM tasks (near transfer) and even more so to different tasks that may involve WM (far transfer), such as reasoning (Melby-Lervåg & Hulme, 2013; Au et al., 2015). For instance,

some studies offered evidence of transfer from WM training to fluid intelligence and complex reasoning (Jaeggi et al., 2008; 2010; Klingberg et al., 2005), reading comprehension (Loosli et al., 2012), and arithmetic (Bergman-Nutley, & Klingberg, 2014), while others reported no transfer to fluid intelligence or any other cognitive domains (Thompson et al., 2013; Estrada et al., 2015). Several meta-analyses interpreted these findings in support of the hypothesis that WM training is only beneficial to improve the trained task but has limited effect on other cognitive abilities (Sala, & Gobet, 2019), while others (Au et al., 2016) supported generalized efficacy of WM and attributed the discrepant results to incorrect analysis.

However, the studies covered in these meta-analyses share limited consistency in training tasks, methodology and outcome measures, which makes interpretation of their conclusions challenging. This is a general problem as many studies in WM field assume a straightforward relationship between the construct of WM and its measurement through any of the variety of available WM measures, such as N-back, simple span tasks, etc. (see Wilhelm et al., 2013). Regardless of the task at hand, the measurement of WM is assumed to represent the 'ground truth' of participants' WM; however, our recent review of the WM training literature (Pergher, Shalchy, Pahor et al., 2019) as well as our empirical comparative study (Shalchy, Pergher, Pahor et al, 2020) suggest that the WM performance, and thus the WM estimation, can be influenced by many factors. Moreover, with more researchers using new tasks or new variations of established tasks and producing highly variable in findings in WM (Blasiman & Was, 2018) and WM training in particular

(Melby-Lervåg & Hulme, 2013; Au et al., 2015), this is an important concern that needs to be addressed. This issue is further complicated by the burgeoning number of studies that use neurophysiological data, which is sensitive not only to task types and task variations, but also to many other situational factors such as type of the machines used to collect the data, pre-processing steps, etc. (Shalchy, Pergher, Pahor et al., 2020; Brouwer et al., 2015).

This work presents my efforts to understand the problem of inconsistent results in the WM training literature and by doing so to further understand WM. To do so we used various approaches that are divided into three chapters. First, in chapter 1 (Pergher, Shalchy, Pahor et al., 2019), we investigate the divergent results of WM training studies (Au et al., 2015; Melby-Lervåg, et al., 2016) by characterizing the broad diversity of features employed in N-back training tasks and behavioral outcome measures in published WM training literature. The N-back task is a well-established WM task (Kirchner, 1958; Jaeggi et al., 2008) that requires the participant to continuously store and update the last N items of a sequence in memory. N-back is a popular tool as it can be easily adapted to experimental needs and generate multiple versions that can serve either as training or assessment tasks. In this chapter, our aim is to point out that the features of training/assessment tasks and stimuli differ on many levels across different studies of WM, and combining the results of these studies would be similar to comparing apples and oranges. Therefore, the question of "Does WM training work?" does not have a clear dichotomous outcome and further investigation is required to provide us with mechanistic understanding of the differences in task and stimulus in brain and behavior.

In chapter 2 (Shalchy, Pergher, Pahor et al., 2020), we investigate EEG signatures during variations of N-back task structures and stimulus types. We test whether different combinations of experimental design parameters (i.e., stimulus type, stimulus duration, inter-stimulus interval, and response contingency) differentially affect electrophysiological signatures. This helps us understand how variations in task and stimulus may contribute to differences found across training approaches. In this chapter, our aim is to gain a mechanistic understanding of how different task structures and stimuli are processed by the brain during early (i.e., encoding) and late (i.e., retrieval) WM processes and to examine the so-called electrophysiological correlates of WM. Moreover, we evaluate the effect of other unconsidered factors (i.e., pre-processing pipelines, type of EEG machines, etc.) that are proper to electrophysiological testing.

In chapter 3, we examine stimulus processing that may underlie encoding of information, a crucial first step in creating new memories. A candidate neuromodulatory system that affects the moment-by-moment variation in stimulus processing is the Locus Coeruleus (LC) – norepinephrine (NE) system (Vazey et al., 2018; Cohen Hoffing & Seitz, 2015). LC is a nucleus in the pons of the brainstem that projects NE widely throughout the cortex and by doing so affects the brain states which can be manifested in stimulus processing, memory encoding, memorization and learning (Sara, 2009; Aston-Jones & Cohen, 2005; Clewett et al., 2018). Our aim in this chapter is to establish a viable experimental paradigm as a first necessary step to study the LC-NE system and its effect on WM.

In sum, the overarching goal of this work is to understand WM and WM training efficacy by testing the human behavior and related brain mechanisms. To do this, we first characterized the existing experimental paradigms and behavioral performance. Next, we brought mechanistic understanding by taking advantage of brain signals as crucial intermediaries of human behavior. Finally, we advanced to developing a new experimental paradigm combined with a tri-variate model: a model including behavioral, physiological and neural data. By doing this, we set the stage for studying the cortical arousal system, a key influencer of memory processes and learning through altering brain states.

References

Anguera, J. A., Bernard, J. A., Jaeggi, S. M., Buschkuehl, M., Benson, B. L., Jennett, S., ... & Seidler, R. D. (2012). The effects of working memory resource depletion and training on sensorimotor adaptation. Behavioural brain research, 228(1), 107-115.

Arjona Valladares, A., Gómez, C. M., Rodríguez-Martínez, E. I., Barriga-Paulino, C. I., Gómez-González, J., & Diaz-Sánchez, J. A. (2020). Attention-deficit/hyperactivity disorder in children and adolescents: An event-related potential study of working memory. *European Journal of Neuroscience*, *52*(10), 4356-4369.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.*, *28*, 403-450.

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: a meta-analysis. Psychonomic bulletin & review, 22(2), 366-377.

Au, J., Katz, B., Buschkuehl, M., Bunarjo, K., Senger, T., Zabel, C., ... & Jonides, J. (2016). Enhancing working memory training with transcranial direct current stimulation. *Journal of cognitive neuroscience*, *28*(9), 1419-1432.

Baddeley, A. (2012). Working memory: theories, models, and controversies. Annual review of psychology, 63, 1-29.

Bergman-Nutley, S., & Klingberg, T. (2014). Effect of working memory training on working memory, arithmetic and following instructions. *Psychological research*, *78*(6), 869-877.

Blacker, K. J., Negoita, S., Ewen, J. B., & Courtney, S. M. (2017). N-back versus complex span working memory training. *Journal of cognitive enhancement*, *1*(4), 434-454.

Blasiman, R. N., & Was, C. A. (2018). Why is working memory performance unstable? A review of 21 factors. *Europe's journal of psychology*, *14*(1), 188.

Brouwer, A. M., Zander, T. O., Van Erp, J. B., Korteling, J. E., & Bronkhorst, A. W. (2015). Using neurophysiological signals that reflect cognitive or affective state: six recommendations to avoid common pitfalls. *Frontiers in neuroscience*, *9*, 136.

Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., & Haynes, J. D. (2017). The distributed nature of working memory. *Trends in cognitive sciences*, *21*(2), 111-124.

Clewett, D. V., Huang, R., Velasco, R., Lee, T. H., & Mather, M. (2018). Locus coeruleus activity strengthens prioritized memories under arousal. *Journal of Neuroscience*, *38*(6), 1558-1574.

Cohen Hoffing, R., & Seitz, A. R. (2015). Pupillometry as a glimpse into the neurochemical basis of human memory encoding. *Journal of cognitive neuroscience*, *27*(4), 765-774.

D'Esposito, M., & Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annual review of psychology*, *66*, 115-142.

Estrada, E., Ferrer, E., Abad, F. J., Román, F. J., & Colom, R. (2015). A general factor of intelligence fails to account for changes in tests' scores after cognitive practice: A longitudinal multi-group latent-variable study. *Intelligence*, *50*, 93-99.

Frydecka, D., Eissa, A. M., Hewedi, D. H., Ali, M., Drapała, J., Misiak, B., ... & Moustafa, A. A. (2014). Impairments of working memory in schizophrenia and bipolar disorder: the effect of history of psychotic symptoms and different aspects of cognitive task demands. *Frontiers in behavioral neuroscience*, *8*, 416.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. Proceedings of the National Academy of Sciences, 105(19), 6829-6833.

Jaeggi, S. M., Studer-Luethi, B., Buschkuehl, M., Su, Y. F., Jonides, J., & Perrig, W. J. (2010). The relationship between n-back performance and matrix reasoning—implications for training and transfer. Intelligence, 38(6), 625-635.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology*, *55*(4), 352.

Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., Dahlström, K., ... & Westerberg, H. (2005). Computerized training of working memory in children with ADHD-a randomized, controlled trial. *Journal of the American Academy of child & adolescent psychiatry*, *44*(2), 177-186.

Loosli, S. V., Buschkuehl, M., Perrig, W. J., & Jaeggi, S. M. (2012). Working memory training improves reading processes in typically developing children. *Child Neuropsychology*, *18*(1), 62-78.

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental psychology*, *49*(2), 270.

Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of "far transfer"

evidence from a meta-analytic review. *Perspectives on Psychological Science*, *11*(4), 512-534.

Minear, M., Brasher, F., Guerrero, C. B., Brasher, M., Moore, A., & Sukeena, J. (2016). A simultaneous examination of two forms of working memory training: Evidence for near transfer only. *Memory & Cognition*, *44*(7), 1014-1037.

Miyake, A., & Shah, P. (Eds.). (1999). Models of working memory: Mechanisms of active maintenance and executive control. Cambridge University Press.

Pergher, V., Shalchy, M. A., Pahor, A., Van Hulle, M. M., Jaeggi, S. M., & Seitz, A. R. (2019). Divergent research methods limit understanding of working memory training. Journal of Cognitive Enhancement, 1-21.

Rose, E. J., & Ebmeier, K. P. (2006). Pattern of impaired working memory during major depression. *Journal of affective disorders*, *90*(2-3), 149-161.

Sala, G., & Gobet, F. (2019). Cognitive training does not enhance general cognition. *Trends in cognitive sciences*, *23*(1), 9-20.

Sara, S. J., & Bouret, S. (2012). Orienting and reorienting: the locus coeruleus mediates cognition through arousal. *Neuron*, *76*(1), 130-141.

Shalchy, M. A., Pergher, V., Pahor, A., Van Hulle, M. M., & Seitz, A. R. (2020). N-Back Related ERPs Depend on Stimulus Type, Task Structure, Pre-processing, and Lab Factors. Frontiers in human neuroscience, 14.

Swanson, H. L., & Alloway, T. P. (2012). Working memory, learning, and academic achievement. In *APA educational psychology handbook, Vol 1: Theories, constructs, and critical issues.* (pp. 327-366). American Psychological Association.

Swanson, H. L., & Siegel, L. (2011). Learning disabilities as a working memory deficit. *Experimental Psychology*, *49*(1), 5-28.

Thompson, T. W., Waskom, M. L., Garel, K. L. A., Cardenas-Iniguez, C., Reynolds, G. O., Winter, R., ... & Gabrieli, J. D. (2013). Failure of working memory training to enhance cognition or intelligence. *PloS one*, *8*(5), e63614.

Vazey, E. M., Moorman, D. E., & Aston-Jones, G. (2018). Phasic locus coeruleus activity regulates cortical encoding of salience information. *Proceedings of the National Academy of Sciences*, *115*(40), E9439-E9448.

Wilhelm, O., Hildebrandt, A. H., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it?. Frontiers in psychology, 4, 433.

Zokaei, N., & Husain, M. (2019). Working memory in Alzheimer's disease and Parkinson's disease. In *Processes of Visuospatial Attention and Working Memory* (pp. 325-344). Springer, Cham.

Chapter 1

This chapter has been previously published as

Divergent Research Methods Limit Understanding of Working Memory Training

Valentina Pergher[1*], Mahsa Alizadeh Shalchy[2*], Anja Pahor[2*], Marc M. Van Hulle[1a], Susanne M. Jaeggi[3a], and Aaron R. Seitz[2a]

1 KU Leuven - University of Leuven, Department of Neurosciences, Laboratory for Neuro- & Psychophysiology, Leuven, Belgium
2 University of California, Riverside, Department of Psychology
Riverside, California, USA
3 University of California, Irvine, School of Education, School of Social Sciences (Department of Cognitive Sciences)
Irvine, California, USA

[*]Contributed equally to the work.
[a]Share senior authorship.

The Journal of Cognitive Enhancement (2019)
https://doi.org/10.1007/s41465-019-00134-7

Abstract

Working memory training has been a hot topic over the last decade. Although studies show benefits in trained and untrained tasks as a function of training, there is an ongoing debate on the efficacy of working memory training. There have been numerous meta-analyses put forth to the field, some finding overall broad transfer effects while others do not. However, discussion of this research typically overlooks specific qualities of the training and transfer tasks. As such, there has been next to no discussion in the literature on what training and transfer tasks features are likely to mediate training outcomes. To address this gap, here, we characterized the broad diversity of features employed in N-Back training tasks and outcome measures in published working memory training studies. Extant meta-analyses have not taken into account the diversity of methodology at this level, primarily because there are too few studies using common methods to allow for a robust meta-analysis. We suggest that these limitations preclude strong conclusions from published data. In order to advance research on working memory training, and in particular, N-Back training, more studies are needed that systematically compare training features and use common outcome measures to assess transfer effects.

Introduction

A longstanding debate has regarded the extent to which training can improve our basic cognitive functions (Katz et al., 2018). Here we address this issue in reference to working memory (WM), defined as a limited-capacity system responsible for temporary storage and manipulation of relevant information. WM is important for a wide range of complex cognitive activities, such as reading or problem-solving (Shah & Miyake, 1999). In the last decade, there has been a considerable amount of literature focused on WM training (Jaeggi, Buschkuehl, Jonides, & Perrig, 2008; Von Bastian, & Oberauer, 2014; Morrison, & Chein, 2011; Klingberg, 2012). For example, WM training on a given task can transfer to improvements in untrained working memory tasks (Blacker et al., 2017; Lilienthal, Tamez, Shelton, Myerson, & Hale, 2013; Chein, & Morrison, 2010; Borella, Carretti, Riboldi, & De Beni, 2010), as well as tasks pertaining to other cognitive domains such as fluid intelligence (Jaeggi et al., 2008; Heinzel et al., 2017; Chein, & Morrison, 2010; Borella et al., 2010). While there are numerous reports of transfer in the literature, there is also a substantial literature showing failures of transfer (Thompson et al., 2013; Jackson, Hill, Payne, Roberts, & Stine-Morrow, 2012). The field has reached a point in which there is a battle of meta-analyses lingering with roughly half of them finding evidence of transfer while the others do not (see Table 1.1 for variety of individual studies upon which these meta-analyses are based).

Table 1. 1. Summary of training features from the 56 studies selected. Transfer effects tests, N-Back type, N-Back modality, feedback, Inter-stimulus interval (ISI), Intervention length/training sessions (short < 10 sessions ≤ long), adaptivity, control group, education, single/double blind, strategies, motivation/expectation, payment, achieved N-Back levels, blocks, main results, transfer tasks.

STUDIES	**TRANSFER EFFECTS TESTS**	**SINGLE/ DUAL**	**N-BACK MODALITY**	**FEED BACK**	**ISI**	**TRAINING SESSIONS**	**ADAPTIVITY**
1. Anguera et al. (2012)	within and beyond WM	dual	Audio/Spatial	Yes	Long	Long	nonforgiving
2. Beavon et al. (2012)	within and beyond WM	single	Spatial	unknown	Long	Long	nonforgiving
3. Blacker et al. (2017)	within and beyond WM	dual	Audio/Spatial	Yes	Long	Long	forgiving
4. Burki et al. (2014)	within and beyond WM	single	Visual	unknown	Long	Long	unknown
5. Buschkuehl et al. (2014)	within WM	single	Spatial	No	Long	Short	nonforgiving
6. Chooi et al. (2012)	beyond WM	dual	Audio/Spatial	unknown	Long	Long	nonforgiving
7. Clark et al. (2017)	beyond WM	dual	Audio/Spatial	Yes	Long	Long	unknown
8. Clouter et al. (2013)	within and beyond WM	dual	Audio/Spatial	Yes	Long	Long	forgiving
9. Colom et al. (2013)	beyond WM	dual	Audio/Spatial	Yes	Long	Long	unknown
10. Feiyue et al. (2009)	beyond WM	dual	Audio/Visual	unknown	Long	Long	nonforgiving
11. Heinzel et al. (2014)	within and beyond WM	single	Visual	No	Short	Long	nonforgiving
12. Heinzel et al. (2016)	within and beyond WM	single	Visual	No	Short	Long	forgiving

STUDIES	TRANSFER EFFECTS TESTS	SINGLE/ DUAL	N-BACK MODALITY	FEED BACK	ISI	TRAINING SESSIONS	ADAPTIVITY
13. Heinzel et al. (2017)	within WM	single	Visual	No	Short	Long	forgiving
14. Hogrefe et al. (2017)	within WM	single	Spatial	Yes	Short	Long	nonforgiving
15. Hussey et al. (2017)	within WM	single	Visual	Yes	Long	Long	nonforgiving
16. Jaeggi et al. (2008)	beyond WM	dual	Audio/Spatial	No	Long	Long	nonforgiving
17. Jaeggi et al. (2010)	within and beyond WM	dual	Audio/Spatial	Yes	Long	Long	nonforgiving
18. Jaeggi et al. (2014)	within and beyond WM	single	Audio	No	Long	Long	nonforgiving
19. Jaeggi et al. (2014)	within and beyond WM	dual	Audio/Spatial	No	Long	Long	nonforgiving
20. Jonasson et al. (2011)	within WM	dual	Audio/Spatial	Yes	Long	Short	nonforgiving
21. Katz et al. (2018)	within and beyond WM	dual	Audio/Spatial	No	Long	Long	nonforgiving
22. Kühn et al. (2013)	within WM	single	Visual/Spatial (numerical updating and spatial N-Back)	No	Long	Long	unknown

STUDIES	TRANSFER EFFECTS TESTS	SINGLE/ DUAL	N-BACK MODALITY	FEED BACK	ISI	TRAINING SESSIONS	ADAPTIVITY
23. Kundu et al. (2013)	within WM	dual	Visual/Spatial	Yes	Long	Long	forgiving
24. Kuper et al. (2016)	within and beyond WM	single	Visual	No	Short	short	forgiving
25. Lawlor-Savage et al. (2016)	within and beyond WM	dual	Audio/Spatial	Yes	Long	Long	forgiving
26. Li et al. (2008)	within WM	single	Spatial (mental spatial shifting and updating)	Yes	Long	Long	unknown
27. Lilienthal et al. (2013)	within WM	dual	Audio/Spatial	No	Long	short	nonforgiving
28. Loosli et al. (2016)	within and beyond WM	single	Visual	No	Long	Long	unknown
29. Maraver et al. (2016)	within and beyond WM	single	Audio/Spatial (N-Back, WM search, WM updating)	Yes	Short	Short	nonforgiving
30. Marcek et al. (2015)	beyond WM	single	Spatial	No	Long	Long	unknown
31. Minear et al. (2016)	within and beyond WM	single	Spatial	No	Long	Long	nonforgiving
32. Mohammed et al. (2017); game N-Back	within and beyond WM	game	Visual/Spatial gaming task	Yes	unknown	Long	forgiving

STUDIES	TRANSFER EFFECTS TESTS	SINGLE/ DUAL	N-BACK MODALITY	FEED BACK	ISI	TRAINING SESSIONS	ADAPTIVITY
33. Mohammed et al. (2017); standard N-back;	within and beyond WM	single	Visual	Yes	unknown	Long	forgiving
34. Nagle et al. (2015)	within and beyond WM	game	Visual/Spatial gaming task	Yes	Long	Long	forgiving
35. Preece et al. (2012)	beyond WM	single	Spatial	Yes	Long	Long	nonforgiving
36. Redick et al. (2013)	within and beyond WM	dual	Audio/Spatial	No	Long	Long	nonforgiving
37. Rudebeck et al. (2012)	within and beyond WM	dual	Visual/Spatial	No	Long	Long	nonforgiving
38. Salminen et al. (2012)	within and beyond WM	dual	Audio/Spatial	Yes	Long	Long	nonforgiving
39. Salminen et al. (2015)	within WM	dual	Audio/Spatial	Yes	Long	Long	nonforgiving
40. Schwarb et al. (2015)	within and beyond WM	single	Visual/Spatial	unknown	Long	Short	nonforgiving
41. Schweizer et al. (2011)	beyond WM	dual	Audio/Visual	unknown	Long	Long	nonforgiving
42. Shahar et al. (2015)	beyond WM	single	Visual/Spatial	Yes	Short	Long	forgiving
43. Smith et al. (2013)	beyond WM	dual	Audio/Spatial	unknown	Long	Long	unknown

STUDIES	TRANSFER EFFECTS TESTS	SINGLE/ DUAL	N-BACK MODALITY	FEED BACK	ISI	TRAINING SESSIONS	ADAPTIVITY
44. Soveri et al. (2017)	within WM	dual	Audio/Spatial	No	Long	Long	nonforgiving
45. Stepankova et al. (2013); long intervention	within and beyond WM	single	Visual	Yes	Short	Long	nonforgiving
46. Stepankova et al. (2013); short intervention	within and beyond WM	single	Visual	Yes	Short	Short	nonforgiving
47. Stephenson et al. (2013); single visual N-Back	beyond WM	single	Visual	No	Long	Long	nonforgiving
48. Stephenson et al. (2013); single audio N-Back	beyond WM	single	Audio	No	Long	Long	nonforgiving
49. Stephenson et al. (2013); dual N-Back	beyond WM	dual	Audio/Spatial	No	Long	Long	nonforgiving
50. Thompson et al. (2013)	within and beyond WM	dual	Audio/Spatial	No	Long	Long	nonforgiving
51. Urbanek et al. (2015)	beyond WM	single	Spatial	No	Long	short	forgiving
52. Vartanian et al. (2013)	beyond WM	single	Visual	No	Long	short	unknown

STUDIES	TRANSFER EFFECTS TESTS	SINGLE/ DUAL	N-BACK MODALITY	FEED BACK	ISI	TRAINING SESSIONS	ADAPTIVITY
53. Waris et al. (2015)	within and beyond WM	dual	Audio/Spatial (selective updating task, moving figures, dual N-Back)	No	Long	Long	nonforgiving
54. Zajac-Lamparska et al. (2016)	beyond WM	single	Visual	No	Long	short	unknown
55. Zhao et al. (2017); HAM group	within and beyond WM	single	Spatial	Yes	Long	Long	nonforgiving
56. Zhao et al. (2017); LAM group	within and beyond WM	single	Spatial	Yes	Long	Long	nonforgiving

STUDIES	Control group	Education info	Single/ double blind info	Strategies info	Motivation/ expectation info	Payment	Starting/ ending (achieved) n-back levels	Blocks
1. Anguera et al. (2012)	Active (knowledge training)	No	No	No	Yes	Low	3-Back to 5-Back	9 blocks (20 trials/block)
2. Beavon et al. (2012)	Active (knowledge training)	No	No	No	No	unknown	2-Back to 5-Back	15 blocks (20 trials/block)
3. Blacker et al. (2017)	Active (adaptive non-WM task called Permuted Rule Operations)	Yes	No	No	Yes	High	1-Back to 2-Back	20 blocks (20 trials/block)
4. Burki et al. (2014)	Active (implicit sequence learning training), Passive	Yes	No	Yes	No	unknown	1-Back to 8-Back	15 block (30 trials/block)
5. Buschkuehl et al. (2014)	Active (knowledge training)	No	No	No	No	Low	1-Back to 4-Back	15 blocks (20+n trials/block)
6. Chooi et al. (2012)	Active (dual 1-Back training), Passive	Yes	No	No	No	Low	0-Back to 4-Back	
7. Clark et al. (2017)	Active (processing speed training), Passive	Yes	Yes (single blind)	No	Yes	High		4 blocks (20 trials/block)
8. Clouter et al. (2013)	Active (dual 1-Back training)	No	Yes (single blind)	No	Yes	unknown	1-Back to 4-Back	20 blocks (20 trials/block)

STUDIES	Control group	Education info	Single/ double blind info	Strategies info	Motivation/ expectation info	Payment	Starting/ ending (achieved) n-back levels	Blocks
9. Colom et al. (2013)	Passive	No	No	No	Yes	unknown	2-Back to 5-Back	
10. Feiyue et al. (2009)	Passive	No	No	No	No	unknown	2-Back to 3-Back	20 blocks (20+n trails/block)
11. Heinzel et al. (2014)	Passive	Yes	No	No	No	unknown	0-Back to 12-Back	27 blocks (20-28 trials/block)
12. Heinzel et al. (2016)	Passive	Yes	No	No	No	unknown		36 blocks
13. Heinzel et al. (2017)	Passive	Yes	No	No	No	unknown		36 blocks
14. Hogrefe et al. (2017)	Active (N-Back training with no immediate feedback), Passive	No	No	No	No	unknown	2-Back to 9-Back	10 blocks (20 trials/each)
15. Hussey et al. (2017)	Active (adaptive N-Back training without lures; non-adaptive 3-Back training without lures)	Yes	Yes (double blind)	No	No	High	unknown	2 blocks (15 trials/block)
16. Jaeggi et al. (2008)	Passive	No	No	No	No	unknown	2-Back to 5-Back	20 blocks (20 trials/block)

STUDIES	Control group	Education info	Single/ double blind info	Strategies info	Motivation/ expectation info	Payment	Starting/ ending (achieved) n-back levels	Blocks
17. Jaeggi et al. (2010)	Active (adaptive single N-Back training)	No	No	No	No	High	2-Back to 4-Back	9 blocks (20 trials/block)
18. Jaeggi et al. (2014)	Active (knowledge training)	No	No	No	Yes	High	2-Back to 6-Back	15 blocks (20 trials/block)
19. Jaeggi et al. (2014)	Active (knowledge training)	No	No	No	Yes	unknown	2-Back to 4-Back	15 blocks (20 trials/block)
20. Jonasson et al. (2011)	Active (face-name recall training)	No	No	No	No	unknown	1-Back to 4-Back	15 rounds
21. Katz et al. (2018)	Active (knowledge training)	No	No	No	Yes	High	2-Back to 4-Back	15 blocks (20 trials/block)
22. Kühn et al. (2013)	Active (numerical updating and N-Back training with fixed difficulty level)	No	No	No	No	High	2-Back to 4-Back	8 blocks (39 trials/block)
23. Kundu et al. (2013)	Active (adaptive Tetris)	No	No	No	No	unknown	2-Back to 4-Back	25 blocks (25 trials/block)
24. Kuper et al. (2016)	Passive	Yes	No	No	No	Low	2-Back to 5-Back	15 blocks (20 trials/block)
25. Lawlor-Savage et al. (2016)	Active (1-Back training)	No	Yes (double blind)	No	No	unknown	1-Back to 3-Back	15 blocks (20 trials/block)

STUDIES	Control group	Education info	Single/ double blind info	Strategies info	Motivation/ expectation info	Payment	Starting/ ending (achieved) n-back levels	Blocks
26. Li et al. (2008)	Passive	No	No	No	No	High		4 blocks (22 trials/block)
27. Lilienthal et al. (2013)	Active (non-adaptive dual N-Back), Passive	No	No	No	No	unknown	2-Back to 3-Back	20 blocks (20 trials/block)
28. Loosli et al. (2016)	Active (recent-probes and N-Back training with low proactive interference)	Yes	Yes (double blind)	No	No	Low	2-Back	(100 trials/block)
29. Maraver et al. (2016)	Active (processing speed training), Passive	No	No	No	Yes	Low	1-Back to 3-Back	Not fixed (18 trials/block)
30. Marcek et al. (2015)	Active (non-adaptive Sudoku)	No	No	Yes	Yes	Low		20 blocks (20 trials/block)
31. Minear et al. (2016)	Active (non-adaptive N-Back training, real time strategy video game)	No	No	Yes	Yes	High	2-Back to 4-Back	15 blocks (20 trials/block)
32. Mohammed et al. (2017); game N-Back	None	No	No	No	Yes	Low	3-Back to 5-Back	8-15 blocks (20-40 trials/block)
33. Mohammed et al. (2017); standard N-back;	None	No	No	No	Yes	Low	3-Back to 5-Back	8-15 blocks (20-40 trials/block)

STUDIES	Control group	Education info	Single/ double blind info	Strategies info	Motivation/ expectation info	Payment	Starting/ ending (achieved) n-back levels	Blocks
34. Nagle et al. (2015)	None	No	No	No	Yes	unknown	2-Back to 3-Back	4 blocks (15 trials/block)
35. Preece et al. (2012)	Active (vocabulary and knowledge training)	No	No	No	No	unknown	3-Back to 5-Back	15 rounds
36. Redick et al. (2013)	Active (visual search training), Passive	No	No	Yes	Yes	High	2-Back to 4-Back	20 blocks (20 trials for each)
37. Rudebeck et al. (2012)	Passive	Yes	No	No	No	unknown	2-Back to 3-Back	12 blocks (30 trials for each)
38. Salminen et al. (2012)	Passive	No	No	No	No	unknown	2-Back to 5-Back	20 blocks (22 trials for each)
39. Salminen et al. (2015)	Passive	Yes	No	No	No	Low	1-Back to 2-Back	20 blocks (22 trials for each)
40. Schwarb et al. (2015)	Passive	No	No	No	No	Low	4-Back to 6-Back	18 blocks (20 trials/block)
41. Schweizer et al. (2011)	Active (feature matching training)	Yes	No	/	No	unknown	4-Back to 7-Back	20 blocks (20 trials/block)
42. Shahar et al. (2015)	Passive	No	Yes (double blind)	No	No	Low	1-Back-16-Back	10,7 blocks (64 trials/block)

STUDIES	Control group	Education info	Single/ double blind info	Strategies info	Motivation/ expectation info	Payment	Starting/ ending (achieved) n-back levels	Blocks
43. Smith et al. (2013)	Active (strategy video game training), Passive	No	No	No	No	unknown	2-Back-5-Back	(20 trials/block)
44. Soveri et al. (2017)	Active (non adaptive game Bejeweled 2)	Yes	No	No	Yes	Low	2-Back to 4-Back	20 blocks (20 trials/block)
45. Stepankova et al. (2013); long intervention	Passive	Yes	No	No	No	unknown	2-Back to 4-Back	20 blocks (20 trials/block)
46. Stepankova et al. (2013); short intervention	Passive	Yes	No	No	No	unknown	2-Back to 4-Back	20 blocks (20 trials/block)
47. Stephenson et al. (2013); single visual N-Back	Active (spatial matrix span training), Passive	Yes	No	No	No	unknown		(20 trials/block)
48. Stephenson et al. (2013); single audio N-Back	Active (spatial matrix span training), Passive	Yes	No	No	No	unknown		(20 trials/block)
49. Stephenson et al. (2013); dual N-Back	Active (spatial matrix span training), Passive	Yes	No	No	No	unknown		(20 trials/block)
50. Thompson et al. (2013)	Active (multiple object tracking training), Passive	No	No	Yes	Yes	High	3-Back to 5-Back	30 blocks (20 trials/block)

STUDIES	Control group	Education info	Single/ double blind info	Strategies info	Motivation/ expectation info	Payment	Starting/ ending (achieved) n-back levels	Blocks
51. Urbanek et al. (2015)	Active (non-adaptive Sudoku)	No	No	No	No	Low	2-Back to 4-Back	15 blocks (20 trials/block)
52. Vartanian et al. (2013)	Active (reaction time training)	No	No	No	No	unknown		4
53. Waris et al. (2015)	Active (Angry Birds, Bejeweled 2, Peggle)	No	No	No	Yes	Low	2-Back to 4-Back	(20 trials/block)
54. Zajac-Lamparska et al. (2016)	Active (attentional control training), Passive	Yes	No	No	No	unknown		(20 trials/block)
55. Zhao et al. (2017); HAM group	None	No	No	No	Yes	unknown	3-Back to 5-Back	15 blocks (20 trials/block)
56. Zhao et al. (2017); LAM group	None	No	No	No	Yes	unknown	3-Back to 5-Back	15 blocks (20 trials/block)

STUDIES	MAIN RESULTS (TRANSFER EFFECTS)	TRANSFER TASKS
1. Anguera et al. (2012)	Transfer to the 3-Back task and visuospatial tasks. No transfer to visuo-motor task.	N-Back, Operation span (OSPAN), Card rotation, Digit symbol substitution (WAIS-R), Visuomotor adaptation
2. Beavon et al. (2012)	No transfer effects to STM and WM.	Numbers Reversed (WJ-III), Auditory WM (WJ-III)
3. Blacker et al. (2017)	Transfer to near WM task (objects N-Back). No far-transfer to fluid intelligence.	N-Back, Symmetry Span, Spatial locations and relations, BOMAT
4. Burki et al. (2014)	Near-transfer effects to a similar spatial N-Back task.	N-Back, Numerical updating, Reading span, RSPM, RAPM, Stroop, Letter and number comparison (pattern comparison), SRT
5. Buschkuehl et al. (2014)	Near transfer to N-Back task with different stimuli.	N-Back
6. Chooi et al. (2012)	No transfer effects.	OSPAN, Vocabulary (Mill-Hill, PMA), Word beginning & ending, Colorado perceptual speed, Identical pictures, Finding A's, Card rotation (ETS), Paper folding (ETS), Mental rotation (Shepard–Metzler), RAPM
7. Clark et al. (2017)	No transfer effects.	RAPM, WAIS-IV, CCFT, Lexical decision, Ospan, Spatial delayed response task
8. Clouter et al. (2013)	Transfer effects to fluid intelligence, WM and conflict resolution.	CFIT, Stroop, Monty Hall Problem, OSPAN, Symmetry span
9. Colom et al. (2013)	No transfer to fluid intelligence, crystallized intelligence, WM and attention control.	RAPM, Abstract reasoning (DAT-AR), Inductive reasoning (PMA-R), WM: reading span, computation span, dot matrix
10. Feiyue et al. (2009)	Transfer effects to fluid intelligence, larger for the training group.	RSPM

STUDIES	MAIN RESULTS (TRANSFER EFFECTS)	TRANSFER TASKS
11. Heinzel et al. (2014)	Transfer to Verbal Fluency and Digit Symbol, Digit Span Fwd, Digit Symbol and CERAD Del Recall	Digit span Fwd (WAIS), Digit span Bwd (WAIS), Recall (CERAD) (immediate and delayed), Digit Symbol (WAIS), Verbal fluency (COWAT), RSPM, Figural relations (LPS)
12. Heinzel et al. (2016)	Transfer to Sternberg task, processing speed, executive functions and fluid intelligence.	Digit Span Fwd (Weschler), Digit Span Bwd (Weschler), Digit symbol substitution (Weschler), D2, Stroop, Verbal fluency (COWAT), RSPM, Figural relations (LPS)
13. Heinzel et al. (2017)	Near-transfer to N-Back.	Visual and auditory single tasks, dual task
14. Hogrefe et al. (2017)	Transfer effects to the N-Back task and to numerical memory updating.	N-Back, Task switching, Flanker, Stroop, Numerical updating, RAPM
15. Hussey et al. (2017)	Transfer to untrained memory and language conditions.	N-Back, recognition memory, Verb generation, Stroop, Garden-path recovery, Relative clause processing
16. Jaeggi et al. (2008)	Transfer to fluid intelligence based on training amount.	RAPM, BOMAT
17. Jaeggi et al. (2010)	Transfer effects to fluid intelligence for both groups (single and dual N-Back tasks)	N-Back, OSPAN, RAPM, BOMAT
18. Jaeggi et al. (2014)	Transfer to fluid intelligence	RAPM, CFIT, BOMAT, Surface Development Test (ETS), Space Relations (DAT), Form Board Test (ETS), Interference Test (ETS), Reading Comprehension (AFOQT), Verbal Analogies, Digit Symbol (WAIS)
19. Jaeggi et al. (2014)	Transfer to fluid intelligence	RAPM, CFIT, BOMAT, Surface Development Test (ETS), Space Relations (DAT), Form Board Test (ETS), Interference Test (ETS), Reading Comprehension (AFOQT), Verbal Analogies, Digit Symbol (WAIS)

STUDIES	MAIN RESULTS (TRANSFER EFFECTS)	TRANSFER TASKS
20. Jonasson et al. (2011)	No transfer effects.	N-Back, OSPAN, Addition, Trail making (TMT), dual task
21. Katz et al. (2018)	Transfer to visuo-spatial composite.	RAPM, CFIT, BOMAT, Surface Development Test (ETS), Space Relations (DAT), Form Board Test (ETS), Interference Test (ETS), Reading Comprehension (AFOQT), Verbal Analogies, Digit Symbol (WAIS)
22. Kuhn et al. (2013)	Improvements on untrained working-memory tasks.	N-Back, Spatial updating, Figural and numerical reasoning (Berlin Intelligence Structure Test)
23. Kundu et al. (2013)	Transfer to stimulus processing, STM and visual search performance	Visual-array comparison task, Visual search, OSPAN, Stroop, RAPM
24. Kuper et al. (2016)	Near-transfer to a WM updating task. No far-transfer to switch costs, Stroop and matrix reasoning tasks.	N-Back, Task switching, Stroop, RAPM, Digit symbol substitution, Spot-a-word
25. Lawlor-Savage et al. (2016)	No transfer to fluid intelligence.	Digit span (WAIS), Symbol search (WAIS), Coding (WAIS), OSPAN, RAPM, CFIT
26. Li et al. (2008)	Transfer to a more complex spatial N-Back task and numerical N-Back task. No far-transfer to complex span.	N-Back, OSPAN, Rotation span, Decision speed
27. Lilienthal et al. (2013)	Transfer effects from adaptive training to a running span task (focus of attention).	Cued recall, Focus-switching, Grid span, Operation span, Running span
28. Loosli et al. (2016)	No transfer effects to untrained tasks.	Verb-generation, Paired associates, Stroop, Digit symbol substitution (WAIS-R), TONI

STUDIES	MAIN RESULTS (TRANSFER EFFECTS)	TRANSFER TASKS
29. Maraver et al. (2016)	Transfer effects to reasoning for the inhibitory control training group.	N-Back, Stroop, OSPAN, Stop-signal, AX-CPT, RAPM
30. Marcek et al. (2015)	No transfer for the single N-Back group, transfer to RAPM for the control group (triple N-Back task).	RAPM, BOMAT
31. Minear et al. (2016)	Near-transfer effects to a different N-Back task for both adaptive and non-adaptive N-Back training group.	N-Back, Speed of processing, Dot judgement, Array matching, (Letter-Digit-Arrow-Circle-Reading-Operation-Letter-number-Rotation-Alignment) span, Attention network, Simon, Nonsense Syllogisms (ETS), Inference Tests (ETS), RPM, CFIT, Mathematical aptitude (ETS)
32. Mohammed et al. (2017); game N-Back	Transfer to untrained N-Back, Far transfer to DRM free recall (falsely remembered), DRM (recognition), Space relations, Surface development, Form board, Delay discounting	N-Back, AX-CPT, DRM, Space relations (DAT), Surface development (ETS), Form board test (ETS), BOMAT, Learning from lectures, Math, Delay discounting
33. Mohammed et al. (2017); standard N-back;	Transfer to untrained N-Back, Far transfer to DRM free recall (falsely remembered), Space relations, Surface development, Form board, Math, Delay discounting	N-Back, AX-CPT, DRM, Space relations (DAT), Surface development (ETS), Form board test (ETS), BOMAT, Learning from lectures, Math, Delay discounting
34. Nagle et al. (2015)	No transfer effects.	RAPM, Digit span Fwd, Digit span Bwd, N-Back
35. Preece et al. (2012)	No transfer effects to fluid intelligence compared to the control group.	Figure Weights (WAIS), RAPM

STUDIES	MAIN RESULTS (TRANSFER EFFECTS)	TRANSFER TASKS
36. Redick et al. (2013)	No transfer effects in fluid intelligence, WM, crystallized intelligence and perceptual speed tasks.	RAPM, RSPM, CFIT, Paper folding, Letter sets, Number series, Inference, Verbal analogies, SynWin, Control tower, ATClab, Symmetry span, Running letter span, Vocabulary, General knowledge, Letter and number comparison
37. Rudebeck et al. (2012)	Transfer effects to episodic memory and fluid intelligence.	BOMAT, Recognition memory
38. Salminen et al. (2012)	Transfer to WM updating task, switching situation task and attentional processing. No transfer to reasoning or dual N-Back task.	Updating (AV numbers, VS black bars that was shown in 4 different locations), dual task, Task switching, Attentional blink, RAPM
39. Salminen et al. (2015)	Transfer effects to a WM updating task for both young and older adults.	Updating (AV numbers, VS black bars, one block with VS another block with AV), Task switching, Attentional blink
40. Schwarb et al. (2015)	Transfer effects to visual short-term memory capacity.	OSPAN, Symmetry span, RAPM, Motion interference, Rapid decision-making, Change detection, Short term recall
41. Schweizer et al. (2011)	Transfer to fluid intelligence. Transfer to emotional Stroop task only for affective training group.	RPM, Stroop, Digit span
42. Shahar et al. (2015)	Transfer to processing speed compared to the control group.	Digits updating task, shape/digit classification task, Stroop, stop-signal, RAPM
43. Smith et al. (2013)	No transfer effect to fluid intelligence.	RAPM
44. Soveri et al. (2017)	Transfer effects to different single N-Back task and to a WM updating task. No transfer effects for active or passive groups.	N-Back, Verbal running span, Visuo-spatial running span, Number substitution, Digit span Fwd, Digit span Bwd, Corsi Block

STUDIES	MAIN RESULTS (TRANSFER EFFECTS)	TRANSFER TASKS
45. Stepankova et al. (2013); long intervention	Transfer effects to WM and visuospatial skills.	Digit span (WMS), Letter number sequencing (WAIS), Block design (WAIS), Matrix reasoning (WAIS)
46. Stepankova et al. (2013); short intervention	Transfer effects to WM and visuospatial skills.	Digit span (WMS), Letter number sequencing (WAIS), Block design (WAIS), Matrix reasoning (WAIS)
47. Stephenson et al. (2013); single visual N-Back	Transfer to 4 fluid intelligence tests (APM, Cattell, WASI, BETA-III).	RAPM, CFIT, WASI, BETA, Mental rotation, Paper folding, Vocabulary, Lexical decision
48. Stephenson et al. (2013); single audio N-Back	Transfer to 3 fluid intelligence tests (APM, Cattell, WASI).	RAPM, CFIT, WASI, BETA, Mental rotation, Paper folding, Vocabulary, Lexical decision
49. Stephenson et al. (2013); dual N-Back	Transfer to 4 fluid intelligence tests (APM, Cattell, WASI, BETA-III).	RAPM, CFIT, WASI, BETA, Mental rotation, Paper folding, Vocabulary, Lexical decision
50. Thompson et al. (2013)	No transfer effects to fluid intelligence and other cognitive tasks.	Operation span, Reading span, RAPM, WASI, WAIS-III, Reading comprehension (Nelson Denny), Digit-symbol coding (WAIS-III), Visual matching (Woodcock-Johnson III), Pair cancellation (Woodcock-Johnson III)
51. Urbanek et al. (2015)	No transfer effects to fluid intelligence.	RAPM, BOMAT
52. Vartanian et al. (2013)	Transfer effects to fluid intelligence.	RAPM, Alternate Uses Task (AUT) - test of divergent thinking
53. Waris et al. (2015)	Near transfer effects in a different N-Back task, WM updating and in a WM task.	Verbal running span, Digit span, Corsi block, Set shifting, Visuo-spatial running span, CFIT, Simon, Number substitution, Numerical updating, N-Back

STUDIES	MAIN RESULTS (TRANSFER EFFECTS)	TRANSFER TASKS
54. Zajac-Lamparska et al. (2016)	Weak transfer effects to fluid intelligence.	Attentional control, RSPM, N-Back
55. Zhao et al. (2017); HAM group	Near-transfer effects to WM. No far-transfer effects to executive functions and fluid intelligence.	N-Back task, Running digit span, Go/no-go, Stroop, Task switching, RAPM.
56. Zhao et al. (2017); LAM group	Near-transfer effects to WM. No far-transfer effects to executive functions and fluid intelligence.	N-Back task, Running digit span, Go/no-go, Stroop, Task switching, RAPM.

Studies mentioned in this table are placed in the References section with a *.
Legend: HAM = high achievement motivated group; LAM = low achievement motivated group

The lack of explanation regarding this variability not only casts a shadow on WM training research but also poses a significant hurdle when evaluating the effectiveness of WM training.

One of the most common measures of WM is the N-Back task, an updating task that requires multiple processes (storage, maintenance, and manipulation of information) and is predictive of inter-individual differences in higher cognitive functions (Jaeggi, Buschkuehl, Perrig, & Meier, 2010). Since the N-Back task is also one of the most prominent tasks used in WM training studies, here we limit our discussion on WM training to interventions using N-Back tasks. However, with as many studies using the N-Back task, there are as many variants in methodology. These range from the adopted training approaches (e.g. varying in terms of task timing, types of stimuli, number of stimulus streams, adaptive algorithms, feedback provided, number of training sessions, blind/not-blind; see Fig. 1 for illustration; Table 1.1) to the transfer tasks that are rarely consistent from one study to the next with over 120 different transfer tasks used across the 56 experiments reviewed in 51 studies (see Figure 1.1 for illustration and Table 1.1 for details). For example, across these experiments, 30 different tasks assess aspects of WM and short-term memory (STM), including N-Back and other updating tasks, simple span tasks and various complex WM tasks. Another 28 tasks assess aspects of fluid intelligence, the content of which is predominantly visuospatial (Matrix Reasoning, Block Design, Figure Weights, Paper Folding, Form Board, Surface Development, Space Relations, Abstract Reasoning, Mental Rotation, Card Rotation, TONI etc.) followed by verbal (Letter

Sets, Inference test, Nonsense syllogisms, Inductive Reasoning PMA-R, Verbal Analogies, Reading Comprehension), and quantitative (Number Series) (cf. Table 1.2). With many unique combinations of training methodologies and transfer tasks, and no model to interpret these differences (Katz et al., 2018), we are left with the difficulty of understanding what approaches might give rise to which cognitive outcomes, and what features might determine the boundary conditions of N-Back training.

To date, discrepant findings regarding transfer effects reported by meta-analytic studies, focusing primarily on healthy adults, have been discussed in regard to important moderators such as population demographics, training dose, training type (e.g. single task, multiple tasks), training task (e.g., single N-Back, dual N-Back), training modality (visual, auditory, both), stimulus content (verbal, nonverbal), type of transfer tasks, design type (randomized/not randomized), type of control group (active/passive), attrition rate, training location, supervision, instructional support, feedback, and publication bias (Au et al., 2015; Soveri, Karlsson, Waris, Grönholm-Nyman, & Laine, 2017; Melby-Lervåg & Hulme, 2013; Melby-Lervåg, Redick, & Hulme, 2016; Schwaighofer, Fischer, & Büchner, 2015). While these moderators are certainly relevant, the details of procedures employed in each training study, such as trained and transfer tasks features, which may mediate learning, have been largely ignored.

In this qualitative review, we examine a variety of design factors previously overlooked in N-Back training that bear potential to affect learning and transfer, such as task timing and adaptive procedures, types of stimuli and sensory modality.

Table 1. 2. Transfer tasks categorized by cognitive domain

		N			***N***
Working Memory	N-Back	19		Identical pictures	1
	Operation span (OSPAN)	13	Vocabulary	Verbal fluency (COWAT)	2
	Symmetry span	4		Lexical decision	2
	Numerical updating	6		Word beginning and ending	1
	Reading span	4		Verb generation	2
	Spatial updating	3		Vocabulary (Mill-Hill, PMA)	4
	Rotation span	2	*Fluid intelligence**	Letter sets	1
	Grid span	1		Inference	1
	Running digit span	1		Space relations (DAT)	5
	Running letter span	2		Abstract reasoning (DAT-AR)	1
	Verbal running span	2		Matrix resoning (BETA-III)	1
	Visuo-spatial running span	2		Matrix reasoning (WAIS)	1
	Arrow/Circle span	2		Block design (WAIS)	1
	Visuospatial and auditory-verbal updating	1		Figure weights (WAIS)	1
	Computation span	1		Paper folding (ETS)	3
	Dot matrix	1		Surface development test (ETS)	5
	Digit span (forward, backward)	14		Form board test (ETS)	5
	Digit symbol substitution (WAIS-R)	8		Interference test (ETS)	4
	Corsi block	2		RSPM	8
	Letter number sequencing (WAIS)	1		RAPM	27
	Number substitution	2		BOMAT	11
	Spatial locations and relations	1		Space relations	1
	Visual array comparison task	1		Inductive reasoning (PMA-R)	1

	Array matching	*1*		*CFIT*	*10*
	Spatial delayed response task	1		TONI	1
	Delayed match to sample (single and dual)	2		Number series	1
	Auditory WM (WJ-III)	1		Mental rotation (Shepard-Metzler)	2
LTM	Recall (CERAD) (delayed, immediate)	2		Figural and numerical reasoning (BIST)	1
	Cued recall	1		Verbal analogies	4
	Recognition memory	2		Reading comprehension (AFOQT)	4
	Paired associates	1		Card rotation	2
	Learning from lectures	2	*Crystallized/general intelligence*	General knowledge	1
False memory	Deese–Roediger–McDermott (DRM)	2		WAIS-IV	1
Visual search	Visual search	1		Spot a word	1
	Symbol search	1		Similarities (WASI)	1
	Finding A's	1		Vocabulary (WASI)	1
	Extended Range Vocabulary Test (ETS)	1	Attention /cognitive control	Garden path recovery	1
Reading	Nelson-Denny Comprehension	1		Set shifting	2
	Lexical Decision Test	1		Trail making (TMT)	1
	Nelson-Denny Reading Rate	1		Stroop	12
Math	Mathematical aptitude (ETS)	1		Task switching	5
	Arithmetic aptitude test (ETS)	1		Focus switching	1
	Addition	1		Attentional blink	2
	Math	2		Pair cancellation (WJ-III)	1
Processing speed	Letter and number comparison (pattern comparison)	4		Stop-signal	2
	Simon	2		Go no go	1

	Coding (WAIS)	2		Flanker	1
	Visual matching (WJ-III)	1		Attention network	1
	Colorado Perceptual Speed Test	1		Motion interference	1
	Shape/Digit Classification	1		AX-CPT	3
	SRT	2		D2	1
	Decision speed	1		Attentional control	1
	Dot judgement	1		Visuomotor adaptation	1
Decision making, problem-solving	Monty Hall problem	1	*Motor learning*	Control tower	1
	Rapid decision making	1	*Multitasking*	Synwin	1
	Delay discounting	2		Atclab	1
	Relative clause processing	1	*Divergent thinking*	Alternate Uses Task (AUT)	1

*Fluid intelligence classification was based on Au et al. (2015), Table S3.
Legend: WM = working memory; STM = short-term memory; LTM = long-term memory; COWAT = Controlled Oral Word Association Test; DAT = Differential Aptitude Test; WAIS = Wechsler Adult Intelligence Scale; WJ-III = Woodcock-Johnson III); ETS = Educational Testing Service Kit; RSPM = Raven's Standard Progressive Matrices; RAPM = Raven's Advanced Progressive Matrices; CFIT = Culture Fair Intelligence Test; LPS = Leistungsprüfsystem; PMA-R = Primary Mental Abilities Battery; TONI = Test of Nonverbal Intelligence; WASI = Wechsler Abbreviated Scales of Intelligence; BOMAT = Bochumer Matrizen test; AFOQT = Air Force Officer Qualifying Test; BIST = Berlin Intelligence Structure Test; CERAD = Consortium to Establish a Registry for Alzheimer's Disease; SRT = Simple Reaction Time; Ax-CPT = Ax- continuous performance task; EF = executive functions.

A summary of all training features can be found in Table 1.1. Interestingly, only 8 experiments relied on the same training method, whereas 48 experiments had unique training conditions (Figure 1.1). In addition, we discuss issues pertaining to the size of the transfer battery and the inconsistency in transfer tasks across studies, and how these factors can affect the findings and their interpretation. The novelty of this review is to highlight

the fact that different training protocols and transfer tasks might differentially affect training efficacy and transfer results.

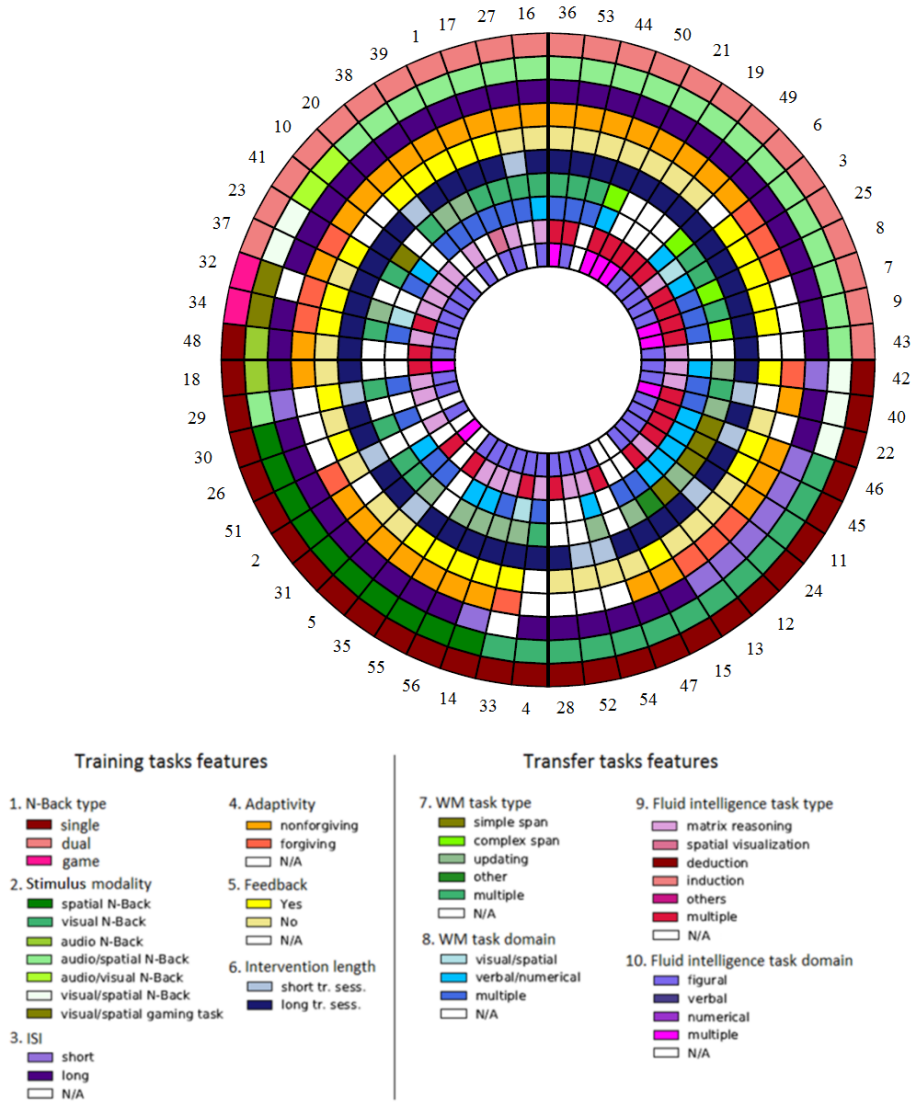

Figure 1. 1. Diversity of training and transfer procedures. Each circle contains 56 sectors, each one corresponding to an N-Back training group included in this review (see Table 1.1). The 6 outer circles reflect training task features whereas the 4 inner circles reflect transfer task features. Starting from the outer circle, each sector is colored in terms of N-Back type (1), Stimulus modality (2), Inter-stimulus interval (ISI) (3), Adaptivity (forgiving vs. non-forgiving) (4), Feedback (5), Intervention length (short < 10 sessions ≤ long) (6), WM (transfer) task type (7), WM (transfer) task domain (8) Fluid intelligence (transfer) task type (9), and Fluid intelligence (transfer) task domain (10).

Training Task Features

We highlight six training task attributes (types of N-Back task and stimulus modality, task timing, adaptive threshold, feedback, and intervention length) that commonly vary across implementations of N-Back training studies. In addition to these, numerous other factors varied across studies within training tasks, such as the number of blocks for each training session, response types (e.g., requiring participants to respond to targets only or also to non-targets), and how feedback was provided (visual/auditory). Within participants, there are additional factors that might determine training outcome, such as N-Back levels achieved, used strategies, or motivation. Note that in many cases, details of the procedures that might be important are simply not reported (see Table S1.1, Supplemental Material). Another source of variation is the inclusion of training procedures that go beyond the N-Back task, thereby targeting additional cognitive processes. For example, Li et al. (2008) incorporated mental spatial shifting in the N-Back training procedure and Mohammed et al. (2017) used a 2D game version of the N-Back task that required navigational skills. In 4 studies, participants trained on other types of updating WM tasks in addition to the N-Back, which precludes understanding of the individual contributions of these training tasks to transfer (Maraver et al; 2016, Waris et al. 2015; Kühn et al., 2013; Loosli et al., 2015).

N-Back task type – single vs. dual – A main area of variation is the use of single or dual N-Back training. Conducting multiple N-Back tasks simultaneously places different demands on attentional and WM resources as compared with a single N-Back.

For example, Jaeggi et al. (2003) showed that single and dual N-Back tasks differ at the behavioral level with longer reaction times and more errors on dual N-Back tasks compared to single N-Back. On the other hand, no differentiation between single and dual N-Back tasks was observed at the neural level: prefrontal activation increased with higher load irrespective of task type. This may explain why single N-Back training seems to be as effective as dual N-Back training (Jaeggi et al., 2008; 2010 b). In the current sample, 29 out of the 56 experiments adopted single N-Back training (13 reporting transfer within WM, 10 reporting transfer beyond WM, and 6 reporting no transfer[1]) and 27 experiments employed dual N-Back training (8 reporting transfer within WM, 9 reporting transfer beyond WM, and 10 reporting no transfer[2]). While this may suggest that dual N-Back training is more likely to yield transfer within and beyond WM, as compared to single N-Back which seems more likely to show transfer within WM, it should be noted that not all studies assessed both types of transfer. Within the single N-Back studies, 2 experiments tested untrained WM tasks, 10 experiments tested for far transfer (6 experiments focusing on fluid intelligence), and 17 experiments tested both. Within the dual N-Back studies, 1 experiment tested untrained WM tasks, 9 tested for far transfer (4 experiments using fluid

[1,2] Note that only studies that assessed transfer are reported here.

intelligence), and 17 experiments tested both. Even though the single vs dual N-Back dichotomy is the most powered of available comparisons, the differences between study methodologies, as described below, largely preclude strong meta-analytic conclusions.

Stimulus Modalities – While WM is often discussed as a domain-general process (Kane et al., 2004), there is substantial evidence that stimuli presented in different modalities (i.e. visual, spatial or auditory stimuli) are processed differently in WM. Owen, McMillan, Laird, and Bullmore (2005) showed changes in brain activation between different N-Back modalities, specifically for location and for non-verbal stimuli. Similarly, Crottaz-Herbette, Anagnoson, and Menon (2004) found differences in neural activation for auditory and non-spatial WM tasks. The authors used, in a randomized order, a visual and an auditory N-Back task. The stimuli were either single-digit numbers (0-9) presented visually at the center of the screen, or binaurally in case of the auditory version. The results showed bilateral suppression of the superior and middle temporal (auditory) cortex during visual (non-spatial) WM, and changes in the occipital (visual) cortex during auditory WM, suggesting that although similar prefrontal and parietal regions are involved in both auditory and visual WM, there are important modality differences in the way neural signals are generated and processed.

For the current review, we define modalities used to categorize the N-Back stimuli as follows: 1) 'spatial N-Back' is a single N-Back task that requires the processing of spatial locations of visual stimuli; 2) 'visual N-Back' describes a single N-Back task that requires the processing of visual stimuli (objects, colors or letters) irrespective of their

spatial location; and 3) ‘audio N-Back’ describes a single N-Back in which stimuli are presented in the auditory domain (e.g. letters, numbers or other sounds). Dual N-Back stimulus modalities are categorized as combinations of the three type of modalities described above: 1) ‘audio-spatial N-Back’ involves concurrent processing of auditory stimuli and spatial locations of visual stimuli; 2) ‘audio-visual N-back’ requires simultaneous processing of auditory stimuli and visual stimuli irrespective of their spatial location; and 3) ‘visual-spatial N-Back’ requires the processing of both visual stimuli and the spatial locations of these stimuli. In addition, ‘visual/spatial gaming N-back’ refers to a gamified (dual) N-Back task that involves processing of different types of visual stimuli presented at different locations.

In our sample, we find that training task modalities vary widely, with 26 using auditory stimuli (7 reporting transfer within WM, 11 reporting transfer beyond WM, and 8 reporting no transfer), 13 using visual stimuli (non-spatial) (5 reporting transfer within WM, 6 reporting transfer beyond WM, and 2 reporting no transfer), and 17 using spatial stimuli (9 reporting transfer within WM, 2 reporting transfer beyond WM, and 6 reporting no transfer). Within those using auditory stimuli, 2 experiments employed a single audio N-Back, 22 used dual audio/spatial N-Back, and 2 used audio/visual N-Back for training. The variety of the auditory stimuli is further highlighted by some studies using letters or syllables for the audio/spatial sub-group, others using words or other type of sounds for the audio/visual sub-group. Overall, N-Back training tasks implement a variety of stimuli (shapes, objects, letters, numbers, etc.) in different modalities (visual, auditory, with or

without a spatial component) (see Figure 1.1), which can be problematic for cross-study comparisons of transfer effects.

Task Timing – Another training feature rarely considered as a relevant factor impacting WM training is the timing between stimuli in the N-Back tasks. Inter-stimulus intervals (ISI) can have an important impact on the time available to process each stimulus and to engage in strategies such as rehearsal or grouping and comparison. The use of these strategies can modify performance levels, give rise to very different experiences during training, and thus likely impact learning outcomes (Laine, Fellman, Waris, & Nyman, 2018). Strüber and Polich (2002) showed that during an oddball task, in which participants needed to press a button every time the visual target stimulus appeared, shorter ISIs were associated with smaller P300 amplitudes. They suggested that long ISIs enables a 'recovery cycle' that can reduce task difficulty. To date, ISI has not been considered a factor relevant to WM training.

In the papers that we reviewed, we screened 56 experiments across single and dual N-Back training and found 45 experiments that reported long ISIs (between 1800 ms and 2500 ms; 20 reporting transfer within WM, 13 reporting transfer beyond WM and 10 reporting no transfer), 9 that used short ISIs (between 500 ms and 1800 ms; 8 reporting transfer within WM, and 1 reporting transfer beyond WM), while 2 experiments did not report ISI information (and did not report any transfer either).

Adaptive threshold – The extent to which training adapts to participants' abilities is another factor that can have a substantial impact on learning and transfer. For example,

in the case of perceptual learning, transfer is greatly impacted by task difficulty with more difficult/precise tasks giving rise to more specificity of learning than found through training involving easier/less-precise stimulus judgements (Hung and Seitz, 2014; Ahissar and Hochstein, 1997). Most N-Back training studies utilize adaptive training by adjusting the level of task difficulty based on individual performance, and it has been shown that adapting the difficulty level of the task is engaging for the participant (Jaeggi, Buschkuehl, Shah, & Jonides, 2014). Moreover, Holmes, Gathercole, & Dunning (2009) showed that WM training gains were significantly greater for an adaptive training group compared to a non-adaptive training group, although others have failed to observe any effects of adaptivity on learning outcome (von Bastian, & Eschen, 2016).

In the papers that we reviewed, we distinguished experiments based on the adaptive threshold used to pass to the next difficulty level: most experiments used a threshold of 90% correct responses (non-forgiving), whereas others used a threshold of 65% or 80% (forgiving). Of 41 experiments, 12 adopted a threshold lower than 90% to achieve the next level (7 reporting transfer within WM, 1 reporting transfer beyond WM, and 4 reporting no transfer), while 33 adopted a threshold of 90% correct (16 reporting transfer within WM, 9 reporting transfer beyond WM, and 8 reporting no transfer). Finally, 3 experiments adapted task difficulty by changing the ISI length (not considered here).

Feedback – Feedback plays an important role in the process of learning, particularly in complex cognitive tasks and in monitoring goal progress (West, Welch and Thorn, 2001). Feedback is usually delivered based on participants' accuracy and/or

response speed and is typically designed to encourage participants to optimize their performance to achieve better learning and/or greater reward (Simen, Buck, Holmes, Hu, & Cohen, 2009). Feedback can indeed facilitate learning, as demonstrated by cognitive training and perceptual learning research (Abe et al., 2011; Seitz, Nanez, Holloway, Tsushima, & Watanabe, 2006).

Out of the 56 experiments reviewed, 24 experiments employed some type of feedback (11 reporting transfer to untrained WM tasks, 5 reporting transfer beyond WM, and 8 reporting no transfer) while 32 experiments either did not provide feedback or did not explicitly report the use of feedback (16 reporting transfer within WM, 9 reporting transfer beyond WM, and 7 reporting no transfer). Of those experiments employing feedback, it was provided at different times: at the end of each block (N = 8), at the end of each session (N = 9), after each trial (N = 4), however, in most of the experiments, timing was not reported (N = 33). Thus, despite the critical role of feedback in motivation and learning (Burgers, Eden, van Engelenburg, and Buningh, 2015), the majority of studies don't describe whether or what type of feedback was employed.

Intervention length – There is evidence that longer training leads to more learning in terms of more pronounced changes in brain regions involved in WM function (Dahlin, Neely, Larsson, Bäckman, & Nyberg, 2008; Lövdén et al., 2010). Hempel et al. (2004) highlighted the role of visual spatial N-Back training length, showing specific brain activation increases with improved performance after 2 weeks of training, and conversely, activation decreases at the time of consolidation of performance gains after 4 weeks. These

results are consistent with the hypothesis that WM training duration affects training results (Jaeggi et al., 2008; Stepankova et al., 2014), although the appropriate amount of training for a given procedure for a given participant is not well established.

In our sample, of the 56 experiments that measured both transfer to WM and beyond WM, 46 used training equal or longer than 10 sessions (29 reporting transfer within WM, 11 reporting transfer beyond WM, and 6 reporting no transfer), and 10 experiments used fewer than 10 sessions (5 reporting transfer within WM, 2 reporting transfer beyond WM, and 3 reporting no transfer).

Transfer Task Features

In addition to the parameters of the training tasks, it is important to consider the details of the outcome measures. Across 56 experiments, 120 different transfer tasks were employed (see Table 1.2), which speaks to the issue of variability in transfer tasks. The number of outcome measures per study ranged from 1 to as many as 20. Using large test batteries can give rise to participant fatigue and decreased participant engagement (Ackerman & Kanfer, 2009), and it can also lead to issues with multiple comparison. In addition, unexpected cognitive benefits may occur as a function of assessing multiple tasks at once, wherein the transfer battery could act as a form of training (Salthouse and Tucker-Drob, 2008; see also Green et al., 2019; Morrison and Chein, 2011). However, using only one or a few outcome measures can limit opportunity to estimate latent factors. Most of the studies investigated transfer effects using a large variety of tests designed to measure more than one cognitive ability, within and beyond WM. In particular, across all the experiments,

9 focused on just one cognitive function (or task type), 10 experiments focused on two, 9 on three, and 28 on four or more cognitive functions. As follows, we give an overview of how these outcome measures varied across experiments:

Transfer within the domain of WM was assessed with 30 different tasks, including various *simple span* measures (Corsi Block, Digit Span, Grid Span, etc.) and *complex span* tasks (Operation Span, Symmetry Span, etc.), *updating* tasks (N-Back, Running Span, Number Updating, etc.), and *other* types of WM tasks such as delayed match to sample tasks and sequencing tasks. Fourteen experiments did not assess WM according to our classification (denoted as *'N/A'* in Figure 1.1), 23 experiments reported using WM measures that fall under one of the four categories mentioned above, and 19 experiments reported using WM tasks that include at least two of these categories (denoted as *'multiple'* in Figure 1.1). Out of the experiments that used only one WM task type, 5 experiments used simple span tasks, another 4 used complex span tasks, 13 used updating tasks, and 1 experiment used a WM task classified as 'other' (for details see 'WM task type' in Figure 1.1). Out of the 42 experiments that measured WM, 15 experiments reported using only verbal/numerical WM tasks and 3 reported using only visual/spatial WM tasks; however, most used WM tasks that covered both verbal/numerical and visual/spatial domains (N = 24; see 'WM task domain' in Figure 1.1).

In sum, even though they all measure some aspects of WM, these 30 different tasks are likely to measure a number of cognitive skills, a fact often overlooked by extant meta-analyses. While some distinctions have been made in terms of task type (untrained N-Back

vs. WM tasks in Soveri et al., 2017) and task domain (verbal vs. visuospatial WM in Melby-Lervåg & Hulme, 2013; Melby-Lervåg, Redick, & Hulme, 2016; Schwaighofer, Fischer, & Büchner, 2015), such categorization does not capture the full range of cognitive demands imposed by different WM tasks and may even mask improvements in a subgroup of tasks. Performance on N-Back tasks only correlates weakly with performance on complex span tasks (Redick and Lindsey, 2013) therefore it makes sense to consider updating and span tasks separately. Furthermore, even if two research groups use the same task with similar types of stimuli, the tasks may still differ in the choice of timing parameters, instructions, feedback etc., as is often the case with custom-built tasks.

Transfer beyond WM, in particular to fluid intelligence, was assessed with 27 different tasks. Forty-seven out of 56 experiments reported assessing fluid intelligence. These tasks were categorized as: *matrix reasoning tests* (including any type of Raven's matrices or Bochum Matrices Test Advanced - BOMAT)*, spatial visualization tests* (Paper folding, Mental rotation, Card Rotation, Surface Development Test, Form Board, Block Design, Spatial Relations), *deduction tests* (Nonsense Syllogisms, Inferences), *induction tests* (Number Series, Inductive Reasoning PMA-R, Letter Sets, Abstract Reasoning DAT, Verbal Analogies) and *other tests* (Reading Comprehension, Figure Weights). Approximately half of the experiments reported the use of batteries that contain multiple tests (e.g., WASI) or the use of multiple tests that include at least 2 of the categories described above (e.g. matrix reasoning and deduction), which were classified as *Multiple* (N = 24). The remaining 23 experiments included matrix reasoning tests (22 experiments)

and spatial visualization tests (1 experiment) (see 'Fluid intelligence task type' in Figure 1.1, and Table 1.2). Moreover, in terms of 'task domain', fluid intelligence tests were categorized as: figural, verbal, or numerical (Beauducel, Brocke, & Liepmann, 2001). Most experiments (N = 38) reported using tests with figural content, and even though no experiments used only verbal or only numerical tests, 9 experiments reported using a combination of figural/verbal or figural/numerical tests.

While matrix reasoning was the most common type of test used to assess fluid intelligence, which allows for a certain level of comparison across experiments, using just one type of test is not sufficient to estimate fluid intelligence at the latent level. When combined with other fluid intelligence tasks, which vary substantially in terms of the cognitive processes that are required to solve the task (i.e. visuospatial transformation, induction, deduction, attention, working memory), and the degree to which these overlap with the cognitive processes targeted during training, estimating training-related changes in the construct of fluid intelligence across studies becomes challenging.

In addition to the two cognitive domains described above, studies also used other transfer measures representing a wide range of cognitive functions (not reported in Figure 1.1; for further details see Table 1.2). Specifically, 5 different tasks were used to assess long-term memory (LTM), 1 task to assess false memory, 4 different tasks to assess visual search, 5 to assess vocabulary, 6 to assess crystallized/ general intelligence, 3 different tasks for reading, 4 for math, 9 different tasks for processing speed, 4 for decision

making/problem solving, 17 different tasks for attention/cognitive control, 1 for motor learning, and 3 for multitasking (for further details see Table 1.2).

Overall, this diversity of transfer tasks measured across studies raises serious issues of the extent to which the same underlying cognitive outcomes are assessed across studies and thus, limits the interpretation of the extant literature.

Test Reliability. An important factor that might impact transfer is task reliability, especially test-retest reliability (Jaeggi et al., 2014). However, for most of the 120 of tasks used, no reliability measures are reported, and it is unclear whether standard forms or custom forms of the tasks are employed, making it difficult to find information on the reliability in the extant literature. It is not uncommon for WM measures to show weak or inconsistent test-retest reliability (e.g. Jaeggi, Buschkuehl, Perrig, and Meier, 2010), which could mask transfer effects: the lower the reliability, the lower the chances for transfer (Jaeggi et al., 2014). Comparing transfer effects on tasks that differ substantially in their reliability may be misleading if this factor is not taken into account. Unfortunately, only a few fluid intelligence tasks have reliable parallel test versions and the commonly used method of splitting tests in half can reduce the reliability and validity of the tests (Jaeggi et al., 2014). Recent efforts to develop multiple parallel reasoning tests may mitigate these types of problems in future intervention studies (Pahor et al., 2018; Kyllonen et al., 2018). Overall, the diversity of transfer tests and batteries used across studies poses a challenge as these outcome measures vary in their degree of similarity with the trained task, and

furthermore, their reliability and their validity in measuring the factor of interest are often unclear.

Control group

It has also been argued that the type of control group plays a significant role in whether transfer is observed. The impact of control groups related to the degree of similarity between the N-Back training and the control interventions, and/or to the differential participant engagement and motivation, and/or participant expectations (Green, et al., 2019). For example, Tsai et al. (2018) suggested that placebo effects might represent an additional factor that contributes to improvements achieved during cognitive training due to alterations in participant expectations. However, literature on WM training is mixed both in regard to what control conditions are employed, some using active controls and others passive controls, and also the extent to which the control type seems to alter the magnitude of observed transfer (Au et al., 2015). A simple reason for this is that the features and the effects of the control condition are likely to be more nuanced than what can be captured by simple distinction into active or passive controls. Participant recruitment and population, as well as other factors like engagement and self-perceived improvements might considerably contribute to the extent to which expectations may impact training outcomes.

In our sample, 51 experiments included at least one control group: 23 experiments included only an active control group, 15 experiments included only a passive control group, 13 experiments included both. Among the 36 experiments that included an active

control group, 7 experiments used vocabulary or knowledge-based training, 8 used commercial games such as Tetris, Angry Birds, and Bejeweled, 10 used a variant of N-Back training (typically non-adaptive and/or low-difficulty), 8 non-WM training (e.g. processing speed training), and 3 experiments, all belonging to one study, employed alternative WM training (spatial STM). These active control conditions differ in their cognitive and perceptual demands and similarity to the experimental condition, as well as most likely in the induced expectations about performance improvement due to training, again making it difficult to compare results across studies.

Discussion and future directions

Although reports on N-Back training are steadily increasing, the mechanisms of transfer and the factors that might impact them are still unclear. We suggest that this lack of clarity is due to the variety of training procedures implemented and the selection of transfer measures gauging training outcomes. Despite numerous meta-analyses aimed to understand the effectiveness of N-Back training (Au et al., 2015; Soveri, Karlsson, Waris, Grönholm-Nyman, & Laine, 2017; Melby-Lervåg & Hulme, 2013; Melby-Lervåg, Redick, & Hulme, 2016; Schwaighofer, Fischer, & Büchner, 2015), there is still disagreement about the extent of transfer after N-Back training. Here we show that N-Back training studies, while seemingly similar, employ a wide variety of training features, and in addition, they assess transfer effects with a large and diverse selection of outcome measures. To highlight this variety, we characterized some of the factors that might be important for learning, such as type of N-Back, stimulus modalities, task timing, adaptive threshold, feedback and

intervention length (see Table 1.1). Given the small sample size of certain training task features and the extensive variability of methods in the literature, we can only speculate whether these factors are meaningful mediators and moderators. The sheer number of transfer tasks used to assess working memory and other cognitive functions further complicates the matter. At this point, in order to achieve a better understanding of the factors that might interfere with transfer outcomes, we suggest that further training studies and meta-analyses should evaluate more carefully the choice of training features (type of stimuli, ISI, intervention length, etc.), transfer measures (for WM, fluid intelligence, LTM, etc.), the type of control groups, and characteristics within the individuals (educational background, strategies, expectation, etc.) before making inferences. Furthermore, training features, transfer tasks, and individual differences need to be systematically addressed, as the large variability represents a severe issue that limits quantitative conclusions.

We suggest that there are several factors that are leading to this diversity of methods, which we argue limit progress in the field. First, there is the conceptual understanding of WM or fluid intelligence as domain-general processes. This view presumably leads researchers to overlook the importance of domain and task specificity, assuming that it does not matter how a specific exercise or test on WM is given (type, modality, etc.), as all approaches would impact the same cognitive process. Although there is still an ongoing debate about the relationship between specificity of cognitive functions and domain-general processes, emphasis should be given to the fact that all the tests used to investigate these constructs are only partially correlated with the underlying construct.

Thus, different training approaches, even if related to the underlying construct, may lead to distinct transfer outcomes due to task specific learning. The second factor is related to the relative nascence of the field. With any new discovery, it makes sense to conduct studies to address the validity of the results and thus using a variety of methods can be vital to explore the space of possibilities. However, this variance of methods produces the inferential problems in making comparisons across studies.

As a first step to address these issues, researchers should both align training and outcome measures across studies and also conduct large-scale comparative studies. As a field, we need to reach some consensus about the training features that may be most conducive to learning, and thus, worth further study. Moreover, a core set of pre-post measures should be defined both within the WM domain and beyond the WM domain. While studies should necessarily differ in attributes, some uniformity across studies with common tests that have known reliability and stability will allow for comparison with other studies, and researches will still have the option to expand the tests battery based on their particular study goals. This would give more power to meta-analyses to address the question whether WM training is worthwhile (and more importantly, for whom it might work and under which circumstances). We recognize that unifying training and transfer task features may be difficult to achieve in practice and so another approach is to conduct larger scale comparative studies with sample sizes sufficient to directly examine unique combinations of training and transfer. Addressing these issues will elevate our

understanding about what approaches do or do not lead to improvements in untrained tasks, as well as the specific domains that are most susceptible to the effects of WM training.

Another important step is bridging the gap between lab tests of cognitive functions and tests that reflect the use of cognitive functions in daily life. To enter the next stage of maturity in the field, new approaches that facilitate comparisons of different training approaches and outcomes are needed, to address issues of robustness, reproducibility and broader generality of findings outside of a limited set of laboratory conditions. To accomplish this, we need to become aware of which WM processes are differently required in daily life activities, and which training condition would be hypothesized to transfer to these conditions. To whatever extent existing tests of cognitive functions predict cognitive functions in daily life, this relationship may not hold after training on task structures that specifically resemble the cognitive tests. For example, if performance on two tasks is correlated, but they do not rely upon the exact same mechanisms, then a change in one may not predict a change in the other.

In conclusion, we suggest that it is time for WM training research to retool. Methods employed to date have been valuable to identify a broad set of issues that need to be considered in order to understand the true benefits and limitations of WM training. However, to move the field forward, it will be necessary to conduct large-scale studies that are targeted to uncover how particular training features and transfer measures may lead to differential learning and generalization of that learning. Furthermore, individual differences that may moderate these training effects need to be considered, together with a

standard set of reliable outcome measures to better understand the profiles of transfer, and how these are reflected in daily-life activities, going beyond the simple question of whether or not near or far transfer occurs.

Acknowledgments

This research was supported by NIMH R01 MH111742 to ARS and SMJ, NIH/NIA 1K02AG054665 to SMJ, and a research grant to VP from the Belgian Fund for Scientific Research-Flanders (G088314N), by research grants to MMVH from the Financing program (PFV/10/008), an interdisciplinary research project (IDO/12/007), an industrial research fund project (IOF/HB/12/021) and a special research fund project (C24/18/098) of the KU Leuven, the Belgian Fund for Scientific Research – Flanders (G088314N, G0A0914N, G0A4118N), the Interuniversity Attraction Poles Programme – Belgian Science Policy (IUAP P7/11), the Flemish Regional Ministry of Education (Belgium) (GOA 10/019), and the Hercules Foundation (AKUL 043). SMJ has an indirect conflict of interest with the MIND Research Institute whose interests are related to this work.

Method

To identify candidate papers, we searched Google Scholar, Google, and PubMed for relevant research reports in the last decade, between 2008 and 2018. The search terms used were “N-Back training” and “updating training”/ “N-Back training game” and “updating training game”. In Google, citation marks were used to reduce noise in the research. The first run resulted in 12.100 hits in Google Scholar for “N-Back training”, 675.000 for “updating training”, 2.730 for “N-Back training game” and 127.000 for “updating training

game". We found 219 hits in PubMed for "N-Back training", 1501 for "updating training", 6 hits for "N-Back training game" and 9 for "updating training game". In Google the hits were 46.300 for "N-Back training", 71.400 for "updating training", 2.170 for "N-Back training game" and no results found for "updating training game". We screened all hits in the databases (Google Scholar, Pubmed and Google) thereby limiting ourselves to the first 150 ranked ones. For a study to be included at this stage, it needed to meet the following criteria:

1. Cognitive training that included game or no-game version of single or dual N-Back task
2. Studies with at least one training group
3. Sample of healthy adults (mean age range 19-69 years old)
4. N-Back training equal to or longer than 3 sessions
5. Focused on transfer to WM and/or other cognitive domains

Search hits were screened in the mentioned ranking, and papers already evaluated in previous databases were not considered in the following screening. Our inclusion criteria decreased the number of the studies to 45 on Google Scholar, 6 on Pubmed and 0 on Google for "N-Back training", "updating training', 'N-Back training game' and "updating training game". In total, our research resulted in 51 studies (excluding the number of overlapping studies) (Figure 1.2). Of these 51 studies, 4 studies included more than one N-Back training group, which we considered separately, giving rise to a total of 56 experiments.

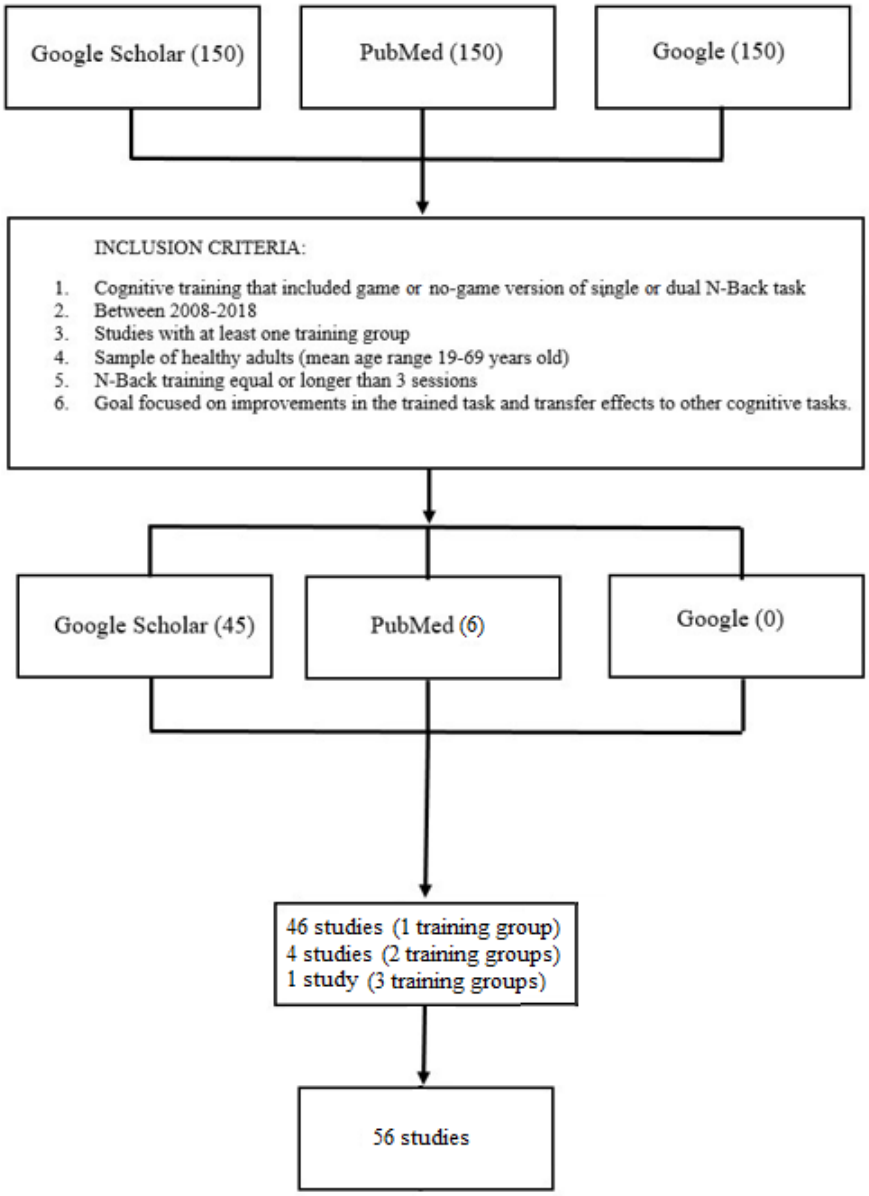

Figure 1. 2. Search for literature and screening process.

References

Abe, M., Schambra, H., Wassermann, E. M., Luckenbaugh, D., Schweighofer, N., & Cohen, L. G. (2011). Reward improves long-term retention of a motor memory through induction of offline memory gains. Current Biology, 21(7), 557-562.

Ackerman, P. L., & Kanfer, R. (2009). Test length and cognitive fatigue: an empirical examination of effects on performance and test-taker reactions. Journal of Experimental Psychology: Applied, 15(2), 163.

*Anguera, J. A., Bernard, J. A., Jaeggi, S. M., Buschkuehl, M., Benson, B. L., Jennett, S., ... & Seidler, R. D. (2012). The effects of working memory resource depletion and training on sensorimotor adaptation. Behavioural Brain Research, 228(1), 107-115.

Ahissar, M., & Hochstein, S. (1997). Task difficulty and the specificity of perceptual learning. Nature, 387(6631), 401.

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: a meta-analysis. Psychonomic bulletin & review, 22(2), 366-377.

Beauducel, A., Brocke, B., & Liepmann, D. (2001). Perspectives on fluid and crystallized intelligence: facets for verbal, numerical, and figural intelligence. Personality and individual differences, 30(6), 977-994.

*Beavon, P. (2012). Improving memory using N-back training.

*Blacker, K. J., Negoita, S., Ewen, J. B., & Courtney, S. M. (2017). N-back versus complex span working memory training. Journal of Cognitive Enhancement, 1(4), 434-454.

Borella, E., Carretti, B., Riboldi, F., & De Beni, R. (2010). Working memory training in older adults: evidence of transfer and maintenance effects. Psychology and aging, 25(4), 767.

Buonomano, D. V., Bramen, J., & Khodadadifar, M. (2009). Influence of the interstimulus interval on temporal processing and learning: testing the state-dependent network model. Philosophical Transactions of the Royal Society of London B: Biological Sciences, 364(1525), 1865-1873.

Buonomano, D.V. & Merzenich, M.M. (1998). Cortical plasticity: from synapses to maps. Annual review of neuroscience, 21(1), 149–186.

Burgers, C., Eden, A., van Engelenburg, M. D., & Buningh, S. (2015). How feedback boosts motivation and play in a brain-training game. Computers in Human Behavior, 48, 94-103.

*Bürki, C. N., Ludwig, C., Chicherio, C., & De Ribaupierre, A. (2014). Individual differences in cognitive plasticity: An investigation of training curves in younger and older adults. Psychological Research, 78(6), 821–835.

*Buschkuehl, M., Hernandez-Garcia, L., Jaeggi, S.M., Bernard, J. A., & Jonides, J. (2014). Neural effects of short-term training on working memory. Cognitive, Affective, & Behavioral Neuroscience, 14(1), 147–160.

Chein, J. M., & Morrison, A. B. (2010). Expanding the mind's workspace: Training and transfer effects with a complex working memory span task. Psychonomic Bulletin & Review, 17(2), 193–199.

*Chooi, W.-T., & Thompson, L. A. (2012). Working memory training does not improve intelligence in healthy young adults. Intelligence, 40(6), 531–542.

*Clark, C. M., Lawlor-Savage, L., & Goghari, V. M. (2017a). Functional brain activation associated with working memory training and transfer. Behavioural brain research, 334, 34-49.

*Clark, C. M., Lawlor-Savage, L., & Goghari, V. M. (2017b). Working memory training in healthy young adults: Support for the null from a randomized comparison to active and passive control groups. PloS one, 12(5), e0177707.

*Clouter, A. (2013). The effects of dual n-back training on the components of working memory and fluid intelligence: An individual differences approach.

*Colom, R., Román, F. J., Abad, F. J., Shih, P. C., Privado, J., Froufe, M., ... & Karama, S. (2013). Adaptive n-back training does not improve fluid intelligence at the construct level: Gains on individual tests suggest that training may enhance visuospatial processing. Intelligence, 41(5), 712-727.

Crottaz-Herbette, S., Anagnoson, R. T., & Menon, V. (2004). Modality effects in verbal working memory: differential prefrontal and parietal responses to auditory and visual stimuli. Neuroimage, 21(1), 340-351.

Dahlin, E., Neely, A. S., Larsson, A., Bäckman, L., & Nyberg, L. (2008). Transfer of learning after updating training mediated by the striatum. Science, 320(5882), 1510-1512.

*Qiu, F., Wei, Q., Zhao, L., Lin, L. (2009). Study on Improving Fluid Intelligence through Cognitive Training System Based on Gabor Stimulus. The 1st International Conference on Information Science and Engineering (ICISE2009).

Fox, M. D., Snyder, A. Z., Vincent, J. L., & Raichle, M. E. (2007). Intrinsic fluctuations within cortical systems account for intertrial variability in human behavior. Neuron, 56(1), 171-184.

Green, C. S., Bavelier, D., Kramer, A. F., Vinogradov, S., Ansorge, U., Ball, K. K., … Witt, C. M. (2019). Improving Methodological Standards in Behavioral Interventions for Cognitive Enhancement. Journal of Cognitive Enhancement.

*Heinzel, S., Lorenz, R. C., Pelz, P., Heinz, A.,Walter, H., Kathmann, N., … & Stelzel, C. (2016). Neural correlates of training and transfer effects in working memory in older adults. Neuroimage, 134, 236– 249.

*Heinzel, S., Rimpel, J., Stelzel, C., & Rapp, M. A. (2017). Transfer Effects to a Multimodal Dual-Task after Working Memory Training and Associated Neural Correlates in Older Adults–A Pilot Study. Frontiers in human neuroscience, 11, 85.

*Heinzel, S., Schulte, S., Onken, J., Duong, Q. L., Riemer, T. G., Heinz, A.,& Rapp, M. A. (2014).Working memory training improvements and gains in non-trained cognitive tasks in young and older adults. Neuropsychology, Development, and Cognition. Section B, Aging, Neuropsychology and Cognition, 21(2), 146–173.

Hempel, A., Giesel, F. L., Garcia Caraballo, N. M., Amann, M., Meyer, H., Wüstenberg, T., ... & Schröder, J. (2004). Plasticity of cortical activation related to working memory during training. American Journal of Psychiatry, 161(4), 745-747.

*Hogrefe, A. B., Studer-Luethi, B., Kodzhabashev, S., & Perrig, W. J. (2017). Mechanisms underlying n-back training: response consistency during training influences training outcome. Journal of Cognitive Enhancement, 1(4), 406-418.

Holmes, J., Gathercole, S. E., & Dunning, D. L. (2009). Adaptive training leads to sustained enhancement of poor working memory in children. Developmental science, 12(4), F9-F15.

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of physiology, 160(1), 106-154.

Hung, S. C., & Seitz, A. R. (2014). Prolonged training at threshold promotes robust retinotopic specificity in perceptual learning. Journal of Neuroscience, 34(25), 8423-8431.

*Hussey, E. K., Harbison, J., Teubner-Rhodes, S. E., Mishler, A., Velnoskey, K., & Novick, J. M. (2017). Memory and language improvements following cognitive control training. Journal of Experimental Psychology: Learning, Memory, and Cognition, 43(1), 23.

*Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. Proceedings of the National Academy of Sciences, 105(19), 6829-6833.

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010 a). The concurrent validity of the N-back task as a working memory measure. Memory, 18(4), 394-412.

*Jaeggi, S. M., Buschkuehl, M., Shah, P., & Jonides, J. (2014). The role of individual differences in cognitive training and transfer. Memory and Cognition, 42(3), 464–480.

Jaeggi, S. M., Seewer, R., Nirkko, A. C., Eckstein, D., Schroth, G., Groner, R., & Gutbrod, K. (2003). Does excessive memory load attenuate activation in the prefrontal cortex? Load-dependent processing in single and dual tasks: functional magnetic resonance imaging study. NeuroImage, 19(2), 210-225.

*Jaeggi, S. M., Studer-Luethi, B., Buschkuehl, M., Su, Y. F., Jonides, J., & Perrig, W. J. (2010 b). The relationship between n-back performance and matrix reasoning—implications for training and transfer. Intelligence, 38(6), 625-635.

Jackson, J. J., Hill, P. L., Payne, B. R., Roberts, B. W., & Stine-Morrow, E. A. (2012). Can an old dog learn (and want to experience) new tricks? Cognitive training increases openness to experience in older adults. Psychology and Aging, 27(2), 286-292.

*Jonasson, C. (2014). Defining boundaries between school and work: teachers and students' attribution of quality to school-based vocational training. Journal of Education and Work, 27(5), 544-563.

Kane, M. J., Hambrick, D. Z., Tuholski, S. W., Wilhelm, O., Payne, T. W., & Engle, R. W. (2004). The generality of working memory capacity: a latent-variable approach to verbal and visuospatial memory span and reasoning. Journal of Experimental Psychology: General, 133(2), 189.

*Katz, B., Jaeggi, S. M., Buschkuehl, M., Shah, P., & Jonides, J. (2018). The effect of monetary compensation on cognitive training outcomes. Learning and Motivation, 63, 77-90.

Klingberg, T. (2010). Training and plasticity of working memory. Trends in cognitive sciences, 14(7), 317-324.

Klingberg, T. (2012). Is working memory capacity fixed?. Journal of Applied Research in Memory and Cognition, 1(3), 194-196.

*Kühn, S., Schmiedek, F., Noack, H., Wenger, E., Bodammer, N. C., Lindenberger, U., & Lövden, M. (2013). The dynamics of change in striatal activity following updating training. Human brain mapping, 34(7), 1530-1541.

*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. Journal of Neuroscience, 33(20), 8705–8715.

*Küper, K., & Karbach, J. (2016). Increased training complexity reduces the effectiveness of brief working memory training: evidence from short-term single and dual n-back training interventions. Journal of Cognitive Psychology, 28(2), 199-208.

Kyllonen, P., Hartman, R., Sprenger, A., Weeks, J., Bertling, M., McGrew, K., ... & Stankov, L. (2018). General fluid/inductive reasoning battery for a high-ability population. Behavior research methods, 1-16.

Laine, M., Fellman, D., Waris, O., & Nyman, T. J. (2018). The early effects of external and internal strategies on working memory updating training. Scientific Reports, 8(1), 4045.

*Lawlor-Savage, L., & Goghari, V. M. (2016). Dual n-back working memory training in healthy adults: A randomized comparison to processing speed training. PloS one, 11(4), e0151817.

*Li, S. C., Schmiedek, F., Huxhold, O., Röcke, C., Smith, J., & Lindenberger, U. (2008). Working memory plasticity in old age: practice gain, transfer, and maintenance. Psychology and aging, 23(4), 731.

*Lilienthal, L., Tamez, E., Shelton, J. T., Myerson, J., & Hale, S. (2013). Dual n-back training increases the capacity of the focus of attention. Psychonomic bulletin & review, 20(1), 135-141.

*Loosli, S. V., Falquez, R., Unterrainer, J. M., Weiller, C., Rahm, B., & Kaller, C. P. (2016). Training of resistance to proactive interference and working memory in older adults: a randomized double-blind study. International psychogeriatrics, 28(3), 453-467.

Lövdén, M., Schaefer, S., Noack, H., Kanowski, M., Kaufmann, J., Tempelmann, C., ... & Düzel, E. (2010). Performance-related increases in hippocampal N-acetylaspartate (NAA) induced by spatial navigation training are restricted to BDNF Val homozygotes. Cerebral cortex, 21(6), 1435-1442.

*Maraver, M. J., Bajo, M. T., & Gomez-Ariza, C. J. (2016). Training on Working Memory and Inhibitory Control in Young Adults. Frontiers in human neuroscience, 10, 588.

*Marček, V. (2015). Effectiveness of n-back cognitive training: quantitative and qualitative aspects (Doctoral dissertation, Masarykova univerzita, Fakulta sociálních studií).

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. Developmental psychology, 49(2), 270.

Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of "far transfer" evidence from a meta-analytic review. Perspectives on Psychological Science, 11(4), 512-534.

*Minear, M., Brasher, F., Guerrero, C. B., Brasher, M., Moore, A., & Sukeena, J. (2016). A simultaneous examination of two forms of working memory training: Evidence for near transfer only. Memory & cognition, 44(7), 1014-1037.

*Mohammed, S., Flores, L., Deveau, J., Hoffing, R. C., Phung, C., Parlett, C. M., ... & Zordan, V. (2017). The benefits and challenges of implementing motivational features to boost cognitive training outcome. Journal of Cognitive Enhancement, 1(4), 491-507.

Morrison, A. B., & Chein, J. M. (2011). Does working memory training work? The promise and challenges of enhancing cognition by training working memory. Psychonomic bulletin & review, 18(1), 46-60.

*Nagle, A., Riener, R., & Wolf, P. (2015). High user control in game design elements increases compliance and in-game performance in a memory training game. Frontiers in psychology, 6, 1774.

Nee, D. E., Brown, J. W., Askren, M. K., Berman, M. G., Demiralp, E., Krawitz, A., & Jonides, J. (2012). A meta-analysis of executive components of working memory. Cerebral cortex, 23(2), 264-282.

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. Human brain mapping, 25(1), 46-59.

Pahor, A., Stavropoulos, T., Jaeggi, S. M., & Seitz, A. R. (2018). Validation of a matrix reasoning task for mobile devices. Behavior research methods, 1-12.

*Preece, D. (2012). The effect of working memory (n-back) training on fluid intelligence.

Ramani, G. B., Jaeggi, S. M., Daubert, E. N., & Buschkuehl, M. (2017). Domain-specific and domain-general training to improve kindergarten children's mathematics. Journal of Numerical Cognition, 3(2), 468-495.

Redick, T. S., & Lindsey, D. R. (2013). Complex span and n-back measures of working memory: a meta-analysis. Psychonomic bulletin & review, 20(6), 1102-1113.

*Redick, T. S., Shipstead, Z., Harrison, T. L., Hicks, K. L., Fried, D. E., Hambrick, D. Z., & Engle, R. W. (2013). No evidence of intelligence improvement after working memory training: a randomized, placebo-controlled study. Journal of Experimental Psychology. General, 142(2), 359–379.

*Rudebeck, S. R., Bor, D., Ormond, A., O'Reilly, J. X. & Lee, A. C. H. (2012). A potential spatial working memory training task to improve both episodic memory and fluid intelligence. PLoS ONE, 7(11): e50431.

*Salminen, T., Frensch, P., Strobach, T., & Schubert, T. (2015). Age-specific differences of dual n-back training. Aging, Neuropsychology, and Cognition, 23(1), 18-39.

*Salminen, T., Strobach, T., & Schubert, T. (2012). On the impacts of working memory training on executive functioning. Frontiers in human neuroscience, 6, 166.

Salthouse, T. A., & Tucker-Drob, E. M. (2008). Implications of short-term retest effects for the interpretation of longitudinal change. Neuropsychology, 22(6), 800.

Schwaighofer, M., Fischer, F., & Bühner, M. (2015). Does working memory training transfer? A meta-analysis including training conditions as moderators. Educational Psychologist, 50(2), 138-166.

*Schwarb, H., Nail, J., & Schumacher, E. H. (2015). Working memory training improves visual short-term memory capacity. Psychological Research, 80(1), 128–148.

*Schweizer, S., Hampshire, A., & Dalgleish, T. (2011). Extending brain-training to the affective domain: increasing cognitive and affective executive control through emotional working memory training. PloS one, 6(9), e24372.

Seitz, A. R., Nanez, J. E., Holloway, S., Tsushima, Y., & Watanabe, T. (2006). Two cases requiring external reinforcement in perceptual learning. Journal of vision, 6(9), 9-9.

Serge, S. R., Priest, H. A., Durlach, P. J., & Johnson, C. I. (2013). The effects of static and adaptive performance feedback in game-based training. Computers in Human Behavior, 29, 1150–1158.

Shah, P., & Miyake, A. (1999). Models of working memory: An introduction. In A. Miyake & P. Shah (Eds.), Models of working memory: Mechanism of active maintenance and executive control (pp. 1–26). New York: Cambridge University Press.

*Shahar, N., & Meiran, N. (2015). Learning to control actions: transfer effects following a procedural cognitive control computerized training. PloS one, 10(3), e0119992.

Simen, P., Contreras, D., Buck, C., Hu, P., Holmes, P., & Cohen, J. D. (2009). Reward rate optimization in two-alternative decision making: empirical tests of theoretical predictions. Journal of Experimental Psychology: Human Perception and Performance, 35(6), 1865.

*Smith, S. P., Stibric, M., & Smithson, D. (2013). Exploring the effectiveness of commercial and custom-built games for cognitive training. Computers in Human Behavior, 29(6), 2388-2393.

*Soveri, A., Antfolk, J., Karlsson, L., Salo, B., & Laine, M. (2017). Working memory training revisited: A multi-level meta-analysis of n-back training studies. Psychonomic bulletin & review, 24(4), 1077-1096.

Soveri, A., Karlsson, E., Waris, O., Grönholm-Nyman, P., & Laine, M. (2017). Pattern of near transfer effects following working memory training with a dual n-back task. Experimental psychology, 64(4), 240.

*Stepankova, H., Lukavsky, J., Buschkuehl, M., Kopecek, M., Ripova, D., & Jaeggi, S. M. (2013). The malleability of working memory and visuospatial skills: A randomized controlled study in older adults. Developmental Psychology, 50(4), 1049–1059.

*Stephenson, C. L., & Halpern, D. F. (2013). Improved matrix reasoning is limited to training on tasks with a visuospatial component. Intelligence, 41(5), 341–357.

Strüber, D., & Polich, J. (2002). P300 and slow wave from oddball and single-stimulus visual tasks: inter-stimulus interval effects. International Journal of psychophysiology, 45(3), 187-196.

*Thompson, T. W., Waskom, M. L., Garel, K. L., Cardenas-Iniguez, C., Reynolds, G. O., Winter, R., & Gabrieli, J. D. (2013). Failure of working memory training to enhance cognition or intelligence. PLoS One, 8(5), e63614.

Tsai, N., Buschkuehl, M., Kamarsu, S., Shah, P., Jonides, J., & Jaeggi, S. M. (2018). (Un) Great Expectations: The Role of Placebo Effects in Cognitive Training. Journal of applied research in memory and cognition, 7(4), 564-573.

*Urbánek, T., & Marček, V. (2015). Investigating the effectiveness of working memory training in the context of Personality Systems Interaction theory. Psychological Research, 80(5), 877–888.

*Vartanian, O., Jobidon, M.-E., Bouak, F., Nakashima, A., Smith, I., Lam, Q., & Cheung, B. (2013).Working memory training is associated with lower prefrontal cortex activation in a divergent thinking task. Neuroscience, 236, 186–194.

von Bastian, C. C., & Eschen, A. (2016). Does working memory training have to be adaptive? Psychological Research, 80(2), 181–194.

von Bastian, C. C., & Oberauer, K. (2014). Effects and mechanisms of working memory training: a review. Psychological Research, 78(6), 803-820.

*Waris, O., Soveri, A., & Laine, M. (2015). Transfer after working memory updating training. PloS one, 10(9), e0138734.

*Zajac-Lamparska, L., & Trempala, J. (2016). Effects of working memory and attentional control training and their transfer onto fluid intelligence in early and late adulthood. Health Psychology Report, 4(1), 41–53.

*Zhao, X., Xu, Y., Fu, J., & Maes, J. H. (2018). Are training and transfer effects of working memory updating training modulated by achievement motivation?. Memory & cognition, 46(3), 398-409.

Supplementary Material

Table S1. 1. Training and individual features that did not show substantial variability across studies (Age) or there was not enough information available to quantify the data (N-Back Strategy, Years of Formal Education, Motivation and Expectancy, Blinding, and Payment).

Age	Due to our selection criteria, all studies included healthy adults. 45 studies focused on middle adulthood (range: 18.4-37.7 years), with the exception of 7 studies that included older adults (range: 65.7-68.9 years).
N-Back Strategy	6 studies assessed participants' strategies during N-Back training in order to assess their role in transfer results. 3 studies reported that a task-specific strategy instruction yielded better results than self-reported strategies in showing transfer to other cognitive tasks (Redick et al., 2013; Thompson et al., 2013; Laine et al., 2018), whereas 2 studies reported that also certain self-reported strategies may mediate learning and transfer (Burki et al., 2014; Minear et al., 2016), and 1 study reported that strategies did not affect transfer (Marcek et al., 2015).
Years of Formal Education	15 studies reported years of formal education (range = 13.8-18.2 years).
Motivation and Expectancy	19 studies investigated the role of motivation and expectancy in training and transfer, but these factors were not considered due to the variability of methods used to evaluate these effects (such as self-report questionnaires, monetary compensation, or feedback).
Blinding	3 were single-blind, 4 were double blind, and 45 did not provide information about blinding.
Payment	26 studies provided information about compensation of research subjects, of which 15 offered low payment (*equal or less than 150$*).

Appendix A

We calculated the effect size by using Hedges'g for two post cognitive tests. Of all the papers included in this study we considered N-Back for near-transfer effects and Raven Progressive Matrix for far-transfer effects (when values for this task was missing in the paper, we used the value of another fluid intelligence test). We considered mean, standard deviation and sample size for each paper and for each group (training and control groups). Next, we calculated Hedges'g for post-tests of WM training and control groups. The assumption was that no group differences existed at pre-test. To check that there were not significant differences, we run a t-test for pre-tests between training and control groups. Then, we ran a t-test by using Hedges'g between training and control groups in order to see if there were any significant differences in Effect Size. Finally, we compared our studies based on our N-Back training features by using a t-test. Regardless, due to small sample size, we were unable to make any strong conclusions regarding the N-back training features.

Table A1. 1. Hedge's g for post training – post control group

N-Back features	Study	Group	M	SD	N	Hedges' g	Group	M	SD	N
SINGLE	Maraver 2016									
AUDIO/SPATIAL			post-test					post-test		
FEEDBACK		**N-Back training group**					**Passive control group**			
SHORT ISI	near	Spatial N-Back	2.43	0.75	32	**3.93**	Spatial N-Back	0.07	0.30	24
SHORT TRAINING SESSIONS	far	RAPM	0.47	0.17	32	**0.29**	RAPM	0.42	0.17	24
SINGLE	Minear 2016									
SPATIAL			post-test					post-test		
NO FEEDBACK		**Spatial N-Back training**					**Non-adaptive N-Back training**			
LONG ISI	near	verbal N-Back	29.3	16.2	31	**0.10**	verbal N-Back	27.7	14.2	27
LONG TRAINING SESSIONS		object N-Back	15.7	18.9	31	**0.21**	object N-Back	19.3	15.4	27
NONFORGIVING	far	Raven	24.1	4.5	31	**0.02**	Raven	24	5.1	27
		Cattell	31.9	3.6	31	**0.24**	Cattell	30.9	4.6	27
GAME/SINGLE	Mohamed 2017									
VISUAL/SPATIAL			post-test					post-test		
FEEDBACK		**non-gamified N-Back training**					**gamified N-Back training**			
LONG TRAINING SESSIONS	near	2-Back	0.84	0.13	46	**0.15**	2-Back	0.82	0.13	67
FORGIVING	far	BOMAT	14.00	3.07	47	**0.01**	BOMAT	13.97	3.43	67
		DAT	12.81	3.29	47	**0.03**	DAT	12.91	3.15	67
		ETS form board	73.60	24.58	47	**0.04**	ETS form board	72.56	22.07	67
SINGLE	Loosli 2015									
VISUAL			post-test					post-test		
NO FEEDBACK		**High interference N-Back training**					**Low interference N-Back training**			
LONG ISI	near	N-Back	0.16	0.17	14	**0.22**	N-Back	0.20	0.19	11

N-Back features	Study	Group	M	SD	N	Hedges' g	Group	M	SD	N
SINGLE	Preece_2012									
SPATIAL			post-test					post-test		
FEEDBACK		**N-Back training**					**Control group**			
LONG ISI	near	N-Back	5.45	1.23	29		N-Back	-	-	29
LONG TRAINING SESSIONS	far	Figure Weights	20.45	4.43	27	**0.03**	Figure Weights	20.33	3.68	27
DUAL	Redick_2013									
AUDIO/SPATIAL			post-test					post-test		
NO FEEDBACK		**N-Back training**					**Passive control group**			
LONG ISI	near	-					-			
LONG TRAINING SESSIONS	far	RAPM	6.25	3.08	24	**0.08**	RAPM	6	3	20
NONFORGIVING		RASPM	16.09	2.61	24	**0.30**	RASPM	16.85	2.35	20
		Cattel	11.38	2.45	24	**0.03**	Cattel	11.45	2.65	20
DUAL	Rudebeck_2012									
VISUAL/SPATIAL			post-test					post-test		
NO FEEDBACK		**HG N-Back training**					**Passive control group**			
LONG ISI	near	N-Back	3.49	0.83	14	**3.05**	N-Back	1.69	0.43	28
LONG TRAINING SESSIONS	far	BOMAT	9.93	2.13	14	**0.91**	BOMAT	7.75	2.53	28
DUAL	Salminen_2012									
AUDIO/SPATIAL			post-test					post-test		
FEEDBACK		**N-Back training**					**Control group**			
LONG ISI	near	N-Back	4.9	1.5	20	**2.22**	N-Back	2.3	0.5	16
LONG TRAINING SESSIONS	far	RAPM	13.7	2.2	20	**0.85**	RAPM	10.9	4.3	16

N-Back features	Study	Group	M	SD	N	Hedges' g	Group	M	SD	N
DUAL	Salminen_2015									
AUDIO/SPATIAL			post-test					post-test		
FEEDBACK		**N-Back training**					**Control group**			
LONG ISI	near	N-Back	2.4	0.7	25	**1.03**	N-Back	1.8	0.4	21
SINGLE	Shahar_2015									
VISUO/SPATIAL			post-test		Sample size			post-test		Sample size
FEEDBACK		**N-Back training**					**Control group**			
SHORT ISI	near	-					-			
LONG TRAINING SESSIONS	far	Fluid intelligence	0.72	0.20	9	**0.31**	Fluid intelligence	0.67	0.12	10
DUAL	Smith_2013									
AUDIO/SPATIAL			post-test					post-test		
LONG ISI		**N-Back training**					**Control group**			
LONG TRAINING SESSIONS	near	-					-			
	far	RPM	11.5	2.99	10	**0.16**	RPM	11.9	1.58	9
DUAL	Soveri_2017									
AUDIO/SPATIAL			post-test					post-test		Sample size
NO FEEDBACK		**N-Back training**					**Active control group**			
LONG ISI	near	1-Back	96.04	4.39	25	**0.11**	1-Back	95.61	3.39	28
LONG TRAINING SESSIONS		2-Back	86.71	9.05	25	**0.60**	2-Back	80.11	12.57	28
DUAL	Stepankova_2013									
VISUAL			post-test					post-test		
FEEDBACK		**N-Back training (10 sessions)**					**control group**			
SHORT ISI	near	N-Back	3.38	0.92	20	**1.35**	N-Back	2.43	0.47	25
LONG TRAINING SESSIONS	far	RPM	20.25	3.77	20	**0.71**	RPM	17.04	5.02	25

N-Back features	Study	Group	M	SD	N	Hedges' g	Group	M	SD	N
DUAL	Stephenson_2013									
AUDIO/SPATIAL			post-test					post-test		
NO FEEDBACK		**N-Back training**					**Control group**			
LONG ISI	near	N-Back	0.48	0.98	28	**0.11**	N-Back	0.38	0.88	26
LONG TRAINING SESSIONS	far	Raven	17.54	4.04	28	**0.91**	Raven	14.15	3.32	26
NONFORGIVING		Cattell	28.75	5.05	28	**0.52**	Cattell	26.26	4.59	26
		WASI	22.04	2.33	28	**0.41**	WASI	21.12	2.14	26
		BETA	22.11	2.01	28	**1.20**	BETA	19.42	2.47	26
DUAL	Thompson_2013									
AUDIO/SPATIAL			post-test				post-test			
NO FEEDBACK		**N-Back training**					**Control group**			
LONG ISI	near	N-Back	2.92	0.67	20		N-Back			19
LONG TRAINING SESSIONS	far	RAPM	13.2	0.67	20	**0.77**	RAPM	12.7	0.62	19
NONFORGIVING		Matrix reasoning	13.4	0.6	20	**0.56**	Matrix reasoning	13.7	0.45	19
SINGLE	Urbanek_2016									
SPATIAL			post-test					post-test		
NO FEEDBACK		**N-Back training**					**Active control group**			
LONG ISI	near	N-Back			37		N-Back			34
SHORT TRAINING SESSIONS	far	RAPM (z-scores)	0.10	0.94	37	**0.17**	RAPM	(-) 0.27	1.07	34
FORGIVING		BOMAT	0.08	1.17	37	**0.04**	BOMAT	0.04	1.11	34
SINGLE	Vartanian_2013									
VISUAL			post-test					post-test		
NO FEEDBACK		**N-Back training**					**N-Back training**			
LONG ISI	near	N-Back			17		N-Back			17

N-Back features	Study	Group	M	SD	N	Hedges' g	Group	M	SD	N
DUAL	Waris_2015									
AUDIO/SPATIAL			post-test					post-test		
NO FEEDBACK		**N-Back training**					**Active control group**			
LONG ISI	near	1-Back	98.5	1.3	15	**0.26**	1-Back	98.1	1.7	16
LONG TRAINING SESSIONS		3-Back	96.2	3.7	15	**0.78**	3-Back	92.2	6.2	16
NONFORGIVING	far	RAPM	16.4	2.8	15	**0.17**	RAPM	15.9	3	16
SINGLE	Zhao_2017									
SPATIAL			post-test					post-test		
FEEDBACK		**N-Back training**					**N-Back training**			
LONG ISI	near	2-Back	4.30	1.87	26	**0.18**	2-Back	3.98	1.76	26
LONG TRAINING SESSIONS	far	Raven	0.73	0.13	26	**0.48**	Raven	0.64	0.23	26
NONFORGIVING										
DUAL	Anguera_2012									
AUDIO/SPATIAL			post-test					post-test		
FEEDBACK		**N-Back training**					**Active control group**			
LONG ISI	near	3-Back	0.74	0.31	22	**0.35**	3-Back	0.60	0.47	22
LONG TRAINING SESSIONS		4-Back	0.63	0.31	22	**0.52**	4-Back	0.47	0.30	22
NONFORGIVING	far	-					-			
SINGLE	Beavon_2012									
SPATIAL			post-test					post-test		
LONG ISI		**N-Back training**					**Active control group**			
LONG TRAINING SESSIONS	near	N-Back	4.81	1.39	26		-			
NONFORGIVING	far	WJ-III Test 7	18.15	4.29	26	**0.07**	WJ-III Test 7	18.43	4	21
		WJ-III Test 9	32.42	5.76	26	**0.05**	WJ-III Test 9	32.71	5.81	21

N-Back features	Study	Group	M	SD	N	Hedges' g	Group	M	SD	N
SINGLE	Burki_2014									
VISUAL			post-test					post-test		
LONG ISI		**N-Back training**					**Control group**			
LONG TRAINING SESSIONS	near	verbal 2-Back	0.96	0.05	22	**0.60**	verbal 2-Back	0.92	0.08	21
		spatial 2-Back	0.93	0.08	22	**0.11**	spatial 2-Back	0.92	0.10	21
	far	Raven	37.41	6.43	22	**0.08**	Raven	36.86	6.55	21
DUAL	Chooi_2012									
AUDIO/SPATIAL			post-test		S			post-test		
LONG ISI		**N-Back training (8 sessions)**					**Active control group**			
LONG TRAINING SESSIONS	near	N-Back					N-Back			
NONFORGIVING	far	RAPM	12.7	2	9	**0.24**	RAPM	13.3	1.91	15
DUAL	Colom_2013									
AUDIO/SPATIAL			post-test					post-test		
FEEDBACK		**N-Back training**					**Control group**			
LONG ISI	near	N-Back					N-Back			
LONG TRAINING SESSIONS	far	RAPM	11.79	2.27	28	**0.41**	RAPM	10.64	3.25	28
		DAT-AR	13.64	3.30	28	**0.08**	DAT-AR	13.36	4	28
		PMA-R	11.82	2.21	28	**0.16**	PMA-R	11.46	2.32	28
SINGLE	Heinzel_2014									
VISUAL			post-test		S			post-test		
NO FEEDBACK		**N-Back training**					**Control group**			
SHORT ISI	near	N-Back	11.93	3.20	15	**4.09**	N-Back	2.27	0.96	15
LONG TRAINING SESSIONS	far	RSPM	24.53	2.90	15	**0.55**	RSPM	23.07	2.34	15
SINGLE	Heinzel_2016									
VISUAL			post-test					post-test		
NO FEEDBACK		**N-Back training**					**Control group**			
SHORT ISI	near	1,2-Back	88.4	1.6	15	**3.66**	N-Back	78.8	3.4	14
LONG TRAINING SESSIONS	far	RSPM	17.73	1.20	15	**1.17**	RSPM	16.43	1	14

N-Back features	Study	Group	M	SD	N	Hedges' g	Group	M	SD	N
SINGLE	Heinzel_2017 (older adults)									
VISUAL			post-test					post-test		
NO FEEDBACK		**N-Back training**					**N-Back training**			
SHORT ISI	near	single visual 2-Back task	0.78	0.17	18	**0.15**	single visual 2-Back task	0.75	0.23	16
LONG TRAINING SESSIONS		single auditory 2-Back task	0.85	0.17	18	**0.00**	single auditory 2-Back task	0.85	0.19	16
FORGIVING		dual 2-Back task	0.36	0.27	18	**0.14**	dual 2-Back task	0.32	0.31	16
		dual 2-Back task	0.40	0.24	18	**0.25**	dual 2-Back task	0.34	0.23	16
SINGLE	Hogrefe_2017									
SPATIAL			post-test					post-test		
FEEDBACK		**N-Back training**					**Control group**			
SHORT ISI	near	N-Back consistency	5.75	1.58	32	**1.08**	N-Back consistency	4.08	1.49	26
LONG TRAINING SESSIONS		N-Back auditory	5.17	1.52	32	**0.83**	N-Back auditory	4.11	0.90	26
NONFORGIVING	far	Fluid intelligence	10.63	2.67	32	**0.12**	Fluid intelligence	10.96	2.63	26
DUAL	Jaeggi_2010									
AUDIO/SPATIAL			post-test					post-test		
FEEDBACK		**Single N-Back training**								
LONG ISI	near	N-Back	0.64	0.18	20	**1.30**	N-Back	0.37	0.22	41
LONG TRAINING SESSIONS	far	RAPM	12.81	2.27	21	**0.44**	RAPM	11.81	2.27	43
NONFORGIVING		BOMAT	13.67	3.17	21	**0.80**	BOMAT	11.44	2.58	43
SINGLE	Jaeggi_2014									
			post-test		Sample size			post-test		
AUDIO(SPATIAL)		**Single N-Back training**					**Active control group**			
NO FEEDBACK	near	N-Back					N-Back			
LONG ISI	far	BOMAT	19.21	2.89	14	**0.59**	BOMAT	17.13	3.82	23
LONG TRAINING SESSIONS		CFT	20.29	2.43	14	**0.24**	CFT	19.57	3.26	23
NONFORGIVING		DST	75.36	11.32	14	**0.31**	DST	71.83	11.51	23

N-Back features	Study	Group	M	SD	N	Hedges' g	Group	M	SD	N
DUAL	Katz_2018									
AUDIO/SPATIAL			post-test					post-test		
NO FEEDBACK		**N-Back training**					**Active control group**			
LONG ISI	near	N-Back					N-Back			
LONG TRAINING SESSIONS	far	BOMAT	16.58	3.52	36	**0.29**	BOMAT	17.63	3.78	27
NONFORGIVING		CFT	19.31	2.96	36	**0.11**	CFT	19.63	3.12	27
DUAL	Kundu_2013									
VISUAL/SPATIAL			post-test					post-test		
FEEDBACK		**N-Back training**					**Control group**			
LONG ISI	near	N-Back					N-Back			
LONG TRAINING SESSIONS	far	RAPM	31	1.73	15	**0.20**	RAPM	30.33	4.51	15
SINGLE	Kuper_2016									
VISUAL			post-test					post-test		
NO FEEDBACK		**Single N-Back training**					**Control group**			
SHORT ISI	near	3-Back	20.8	2.6	18	**1.03**	3-Back	17.2	4.2	18
SHORT TRAINING SESSIONS	far	Fluid intelligene	23.6	6.1	18	**0.43**	Fluid intelligene	20.7	7.3	18
DUAL	Lawlor-Savage_2016									
AUDIO/SPATIAL			post-test					post-test		
FEEDBACK		**N-Back training**					**Active control group**			
LONG ISI	near	N-Back (trained)	3.29	0.65	27		N-Back (trained)			
LONG TRAINING SESSIONS	far	RAPM	11.48	2.98	27	**0.17**	RAPM	11.02	2.34	29
FORGIVING		Cattell	29.04	4.82	27	**0.19**	Cattell	28.13	4.57	29
DUAL	Lilienthal_2012									
AUDIO/SPATIAL			post-test					post-test		
NO FEEDBACK		**N-Back training**					**Control group**			
LONG ISI	near	N-Back	3.56	1.43	13	**1.09**	N-Back	2.56	0.51	26
DUAL	Katz_2018									
AUDIO/SPATIAL			post-test					post-test		

Table A1. 2. Comparison between WM training and control groups

Analysis	**near/far tests**	**p (t-test values)**	**N (number of studies)**
PRE-TESTS DIFFERENCES (MEAN) BETWEEN TRAINING AND CONTROL GROUPS	for N-Back	0.9844	N=29
	for fluid intelligence	0.7026	N=29
HEDGES'G BETWEEN TRAINING AND CONTROL GROUPS	for N-Back	0.0013	N=29
	for fluid intelligence	0.5280	N=29

Table A1. 3. N-back training task features for near/far transfer tests

N-Back training features	**near/far tests**	**p (t-test values)**	**N (number of studies)**
Single vs dual	for N-Back	p=0.7552	N=12
	for fluid intelligence	p=0.28082	N=14
audio vs visual (modality)	for N-Back	p=0.9126	N=10
	for fluid intelligence	p=0.9734	N=7
spatial vs visual (modality)	for N-Back	p=0.0762	N=6
	for fluid intelligence	p=0.1061	N=6
audio vs spatial (modality)	for N-Back	p=0.2653	N=6
	for fluid intelligence	p=0.1005	N=6
Feedback	for N-Back	p=0.8343	N=10
	for fluid intelligence	p=0.2836	N=12
ISI	for N-Back	p=0.1310	N=8
	for fluid intelligence	p=0.8969	N=6
Number of training sessions	for N-Back	p=0.7332	N=3 (sample size too small)
	for fluid intelligence	p=0.0280	N=3 (sample size too small)
Adaptive threshold	for N-Back	p=0.7899	N=7
	for fluid intelligence	p=0.3394	N=7

Chapter 2

This chapter has been previously published as

N-Back related ERPs depend on stimulus type, task structure, pre- processing and lab factors

Mahsa Alizadeh Shalchy[1*], Valentina Pergher[2*], Anja Pahor[1*], Marc M. Van Hulle[2a], and Aaron R. Seitz[1a]

1 University of California, Riverside, Department of Psychology
Riverside, California, USA
2 KU Leuven - University of Leuven, Department of Neurosciences, Laboratory for Neuro- & Psychophysiology, Leuven, Belgium

[*]Contributed equally to the work.
[a]Share senior authorship.

Frontiers in Human Neuroscience (2020)
https://doi.org/10.3389/fnhum.2020.549966

Abstract

The N-Back, a common working memory updating task, is increasingly used in basic and applied psychological research. As such, an increasing number of EEG studies have sought to identify the electrophysiological signatures of N-Back task performance. However, *stimulus type*, *task structure*, pre-processing methods, and differences in laboratory environment, including EEG recording setup employed, greatly vary across studies, which in turn may introduce inconsistencies in the obtained results. Here we address this issue by conducting nine different variations of an N-Back task manipulating *stimulus type* and *task structure*. Furthermore, we explored the effect of the pre-processing method used and differences in laboratory environment. Results reveal significant differences in behavioral and electrophysiological signatures in response to N-Back *stimulus type*, *task structure*, pre-processing method, and laboratory environment. In conclusion, we suggest that experimental factors, analysis pipeline, and laboratory differences, which are often ignored in the literature, need to be accounted for when interpreting findings and making comparisons across studies.

Introduction

Working memory (WM), defined as a limited capacity system responsible for temporary storage and manipulation of relevant information (Baddeley, 2012), has been studied extensively in the last few decades due to the fact that it correlates with a wide range of complex cognitive abilities such as problem solving, reasoning, learning and planning of goal-directed behaviors (Miyake, & Shah, 1999). A considerable number of studies have addressed behavioral and neurophysiological, and underlying hypothetical constructs of WM using both single session (Scharinger, Soutschek, Schubert, & Gerjets, 2015; Scharinger Soutschek, Schubert, & Gerjets, 2017) and repeated practice (Anguera et al., 2012; Buschkuehl, Hernandez-Garcia, Jaeggi, Bernard, & Jonides, 2014; Jaeggi, Buschkuehl, Shah, & Jonides, 2014).

One of the commonly used techniques to probe WM is the N-Back task, a complex task that requires storage, maintenance and manipulation of information (Jaeggi, Buschkuehl, Jonides, & Perrig, 2008; Chen, Mitra, & Schlaghecken, 2008) as well as inhibitory and interference control (Oberauer, 2005; Kane, Conway, Miura, & Colflesh, 2007). The N-Back task has been used in single-session behavioral (Brouwer, Hogervorst, Van Erp, Heffelaar, Zimmerman, & Oostenveld, 2012; Jaeggi, Buschkuehl, Perrig, & Meier, 2010) and neurophysiological (Esposito, Aragri, Piccoli, Tedeschi, Goebel, & Di Salle, 2009; Krause et al., 2000; Scharinger et al., 2017; Pesonen, Hämäläinen, & Krause, 2007) studies as well as in multi-session behavioral (Jaeggi et al., 2008; Jaeggi et al., 2014; Blacker, Negoita, Ewen, & Courtney, 2017; Minear, Brasher, Guerrero, Brasher, Moore,

& Sukeena, 2016) and neurophysiological (Chen, & Mitra, 2009; Dong, Reder, Yao, Liu, & Chen, 2015; Pergher, Wittevrongel, Tournoy, Schoenmakers, & Van Hulle, 2018) training studies, to name a few. Many N-Back studies focus on task difficulty at different N-levels, indicating lower ERP amplitudes for more difficult tasks (Brouwer et al., 2012; Herff, Heger, Fortmann, Hennrich, Putze, & Schultz, 2014; Scharinger et al., 2017; Pergher, Wittevrongel, Tournoy, Schoenmakers & Van Hulle 2019a) and/or *stimulus type*, such as the use of spatial (for instance when the target stimulus occurs in different locations on the screen) vs. verbal (for instance when presented stimulus is word or syllable) stimuli. This indicates that stimulus and load factors play a significant role in modulating P2, N2 and P3 components (Chen et al., 2008; Chen, & Mitra, 2009; Ross, & Segalowitz, 2000; Polich, 2007). However, there are many other task parameters such as stimulus duration, inter-stimulus interval (ISI), feedback, etc. that, although previously explored, are rarely consistent across N-Back studies (for a review see Pergher, Shalchy, Pahor, Van Hulle, Jaeggi, & Seitz, 2019). Different combinations of these parameters may differentially affect electrophysiological signatures associated with task performance and thus limit the interpretation of functional significance of ERP components related to the N-Back task and their comparison across studies.

Here we examine a number of candidate factors that may affect ERP and behavioral signatures during N-Back task performance, not only in terms of task parameters such as *stimulus type* (*words*, *pictures and colors*) and (stimulus duration, ISI, and feedback), but also in terms of different data pre-processing pipelines and laboratory effects, such as

differences in room setup, computer testing stations, as well as EEG hardware and software. While this is true of numerous areas of ERP research, the N-Back is particularly notorious in how it varies across studies (Kane et al., 2007; Owen, McMillan, Laird, & Bullmore, 2005; Mencarelli et al., 2019) and the data presented here is the first to detail the extent of these efforts for a variety of N-Back variations.

Material and Methods

Three datasets involving the N-Back task were included in the current study. Dataset I was collected specifically for the current study and was collected at the University of California – Riverside (UCR), USA. The purpose of this study was to explore the potential factors that affect ERP morphology and behavioral signatures of N-Back task, and to replicate experimental procedures described in two published datasets collected in different labs (Datasets II and III). Dataset II was collected at KU Leuven, Belgium (Pergher et al., 2018) as part of a study that investigated near and far transfer effects, the former involving cognitive sub-processes similar to the one practiced during training, whereas the latter calling upon other mental processes (de Ribaupierre & Lecerf, 2006), after 10 N-Back training sessions using behavioral and EEG recording. Dataset III was collected at the University of Maribor (UM), Slovenia (Pahor et al., 2018) in a study that examined the effects of transcranial alternating current stimulation on working memory performance and EEG responses. Participants in each dataset were healthy young subjects, who reported normal or corrected-to-normal vision, no history of psychiatric or

neurological diseases and were not taking any medication known to interfere with cognitive functioning.

Table 2. 1. Demographics. Means (± Standard Deviations (SDs)) age of participants.

	Participants		
	UCR (Dataset I)	KU Leuven (Dataset II)	UM (Dataset III)
N	36	16	16
Age	19.58±0.97	23.42±0.98	20.56±1.59
Sex	27 F (8M)	9 F (7M)	16 F

Dataset I: UCR

Participants. Thirty-six right-handed adults (27 females and 9 males, mean age = 19.58, SD = 0.97), undergraduate students, were recruited from UCR. The experimental protocol was approved by the Institutional Review Board of UCR and all participants gave their informed consent prior to the experiment. They received course credit and a payment of $10 for participating in two sessions.

Stimuli and task structure. Nine variants of the N-Back task were obtained by crossing 3 *task structures* (see below) with 3 *stimulus types*: *words* (i.e. so, do, up), *picture*s (i.e. apple, fish, bag) and *colors* (i.e. red, green, blue). *task structures* differed in terms of stimulus duration, ISI, response contingency and feedback (see Figure 2. 1) and were

modeled after tasks used in previous studies, as mentioned above: *task* 1 (Pahor, & Jaušovec, 2018), *task* 2 (Pergher et al., 2018), and *task* 3 (Mohammed et al., 2017).

Task 1 had a stimulus duration of 400 ms, ISI of 1600 ms and employed a two-alternative forced choice design for responding to targets and non-targets during the ISI. A white fixation cross appeared during ISI, turning blue when a response was registered or red if no response was detected. *Task* 2 had a stimulus duration of 1000 ms and ISI of 2000 ms. During the ISI, participants viewed a white fixation cross and were instructed to press a button only for target trials. *Task* 3 had a stimulus duration of 2500 ms and ISI of 500 ms. Participants were instructed to respond to targets during stimulus presentation, and were given feedback for correct (green circle around the stimulus) and incorrect responses (red circle around the stimulus). For *task* 1 and *task* 2 no response was allowed during stimulus presentation.

Procedure. Each participant performed four out of the nine N-Back variations across two different sessions conducted on different days, where the same difficulty levels were administered each day, for a total of approximately 90 minutes per session. This ensured that all combination of conditions existed in a within subject design, even though not all participants completed every condition. The assignment of each participant to each N-Back variant was done randomly based on the subject number to ensure an equal number of participants (N = 16) in each variant. Each session consisted of 11 blocks presented in the following order: 1-back practice block, 2-back practice block, four 2-back test blocks, 3-back practice block, and four 3-back experimental blocks. Instructions were provided

prior to each new N-level and 15 second breaks were given between blocks. Practice blocks consisted of sixteen trials during which the participant performed *task* 3 with Color stimuli whereas experimental blocks consisted of N+40 trials (i.e. 2-Back had 42 trials).

The experiment took place in an electrically shielded room with DC lighting. An Apple Mac Mini with OS X 10.6.8 running MATLAB 2007b (Mathworks, Natick, MA, USA) and Psychtoolbox Version 3.0.8 were used to present the task and generate the stimuli (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007). The stimuli were displayed on a 22.5-in. wide Sony Trinitron (Sony Corp., Tokyo, Japan) CRT monitor with a resolution of 1280 × 1024 pixels and a refresh rate of 75 Hz. In addition, to guarantee temporal precision of event-markers with experimental stimuli, a DATAPixx stimulus unit was used (VPixx, Vision Science Solutions, Quebec, Canada) that ensured that triggers were sent precisely at the times of the vertical interrupt of the monitor and button presses.

EEG Recording. EEG was recorded continuously using a Biosemi Active Two system (Biosemi B.V. Amsterdam) operating at a sampling rate of 2048 Hz. Active Two system stored the EEG signal with no high-pass filter and low-pass filtered only by the anti-aliasing filter. Thirty-two active Ag/AgCl electrodes placed according to the 10/20 system (Jasper, 1958) at O1,Oz, O2, , PO3, PO4, P7, P3, Pz, P4, P8, CP5, CP1, CP2, CP6, T7, C3, Cz, C4, T8, FC5, FC1, FC2, FC6, F7, F3, Fz, F4, F8, AF3, AF4, Fp1, Fp2. In addition, six external electrodes were placed on the mastoids for referencing, and to record the horizontal and vertical electro-oculogram (EOG).

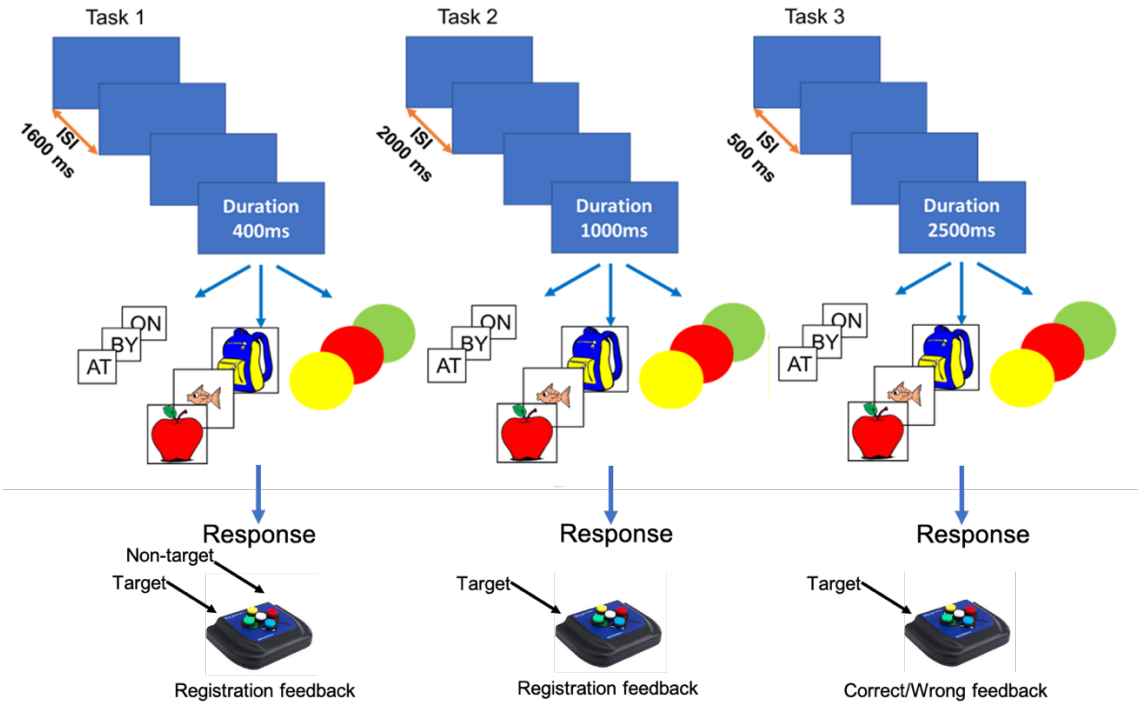

Figure 2. 1. Graphic rendition of N-Back task features for stimulus type, stimulus duration and Inter-stimulus Interval (ISI) for Dataset I.

Dataset II: KU Leuven

Participants. Twenty-three healthy adults (12 females and 11 males, mean age = 24.37, SD = 1.78) were recruited via advertisements and flyers[3]. We randomly selected 16 subjects out of the first two sessions of dataset II to have comparable sample size for cross-laboratory comparison purposes (see Table 1). Prior to starting the experiment, all participants were informed about the experimental procedure and signed an informed

[3] Eight of these participants were included in N-Back training study conducted by Pergher et al. (2018).

consent. They received a payment of 20 euros for participating in two experiment. The study was approved by the UZ KU Leuven ethical committee (S59475).

Stimuli and Procedure. Dataset II had a *task structure* similar to *task* 2 of Dataset I, mentioned above, where each stimulus was presented for 1000 ms, followed by an ISI of 2000 ms. The stimuli were generated using MATLAB 2007b (Mathworks, Natick, MA, USA) and Psychtoolbox Version 3.0.8 (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007) and displayed on a CRT monitor. Participants had to respond only to targets. The stimuli used were pictures (Pergher et al., 2018).

EEG Recording. EEG was recorded at full bandwidth with a SynAmpsRT device (Compumedics, Australia) at a sampling rate of 2000 Hz, using 32 Ag/AgCl electrodes placed at O1, Oz, O2, PO3, P8, P4, Pz, P3, P7, TP9, CP5, CP1, CP2, CP6, TP10, T7, C3, Cz, C4, T8, FC6, FC2, FC1, FC5, F3, Fz, F4, AF3, AF4, Fp1, Fp2. The reference was placed at AFz and the ground at CPz. In addition, four external electrodes around the eyes were used for electro-oculogram recording (EOG) following the instructions of Croft and Barry (2000).

Dataset III: UM

Participants. Seventy-two healthy adults were recruited from the University of Maribor, Slovenia (Pahor, & Jaušovec, 2018[4]), 24 of which were assigned to sham

[4] Dataset 3 only included participants that were in a sham stimulation condition during their first session.

stimulation in the first session (all females, mean age = 20.42, SD = 1.56) and were thus not exposed to any active stimulation. Sixteen of these participants (session 1 data only) were randomly selected for Dataset III (see Table 1). The protocol was approved by the Commission for Ethics in Research at the Faculty of Arts. Participants gave written informed consent and received course credit as compensation.

Stimuli and Procedure. Dataset III had a *task structure* similar to *task* 1 of Dataset I, where each stimulus was shown for 400 ms, followed by an ISI of 1600 ms. The stimuli were generated on STIM2 (Compumedics Neuroscan Systems, Charlotte, NC, USA) and displayed on a CRT monitor. Participants had to respond to both targets and non-targets. Two types of stimuli were used: 2-letter *words* and *colors* (Pahor, & Jaušovec, 2018).

EEG Recording. EEG was recorded over 19 scalp locations based on the 10–20 Electrode Placement System using a Quik-Cap (Quik-Cap Compumedics Neuromedical supplies, Charlotte, NC, USA) with sintered electrodes. The EEG data were recorded using a SynAmps RT system and had a band-pass filter of 0.15–100.0 Hz. The 19 EEG traces were sampled online at 1000 Hz. Vertical eye movements were recorded using two external electrodes placed above and below the left eye and a ground electrode was applied to the forehead. Two ear lobe references (A1 and A2) were used for online referencing, followed by common average re-referencing.

Preprocessing and Analysis

ERP Pre-processing Pipelines. Two pre-processing pipelines were used to analyze Dataset I: pipeline I and pipeline II. For ERP comparison across the three datasets,

only pipeline I was used. The pipelines were chosen as they represented different, but standard, approaches to ERP analysis (Delorme & Makeig, 2004; Groppe, Makeig & Kutas, 2009; Croft, & Barry, 2000).

Pipeline I. Pipeline I was conducted in EEGLAB (Matlab 2015a, MathWorks, Inc.; EEGLAB v. 14.1.1 Delorme, & Makeig, 2004): the data was resampled to 512 Hz and filtered using a Butterworth filter with lower and upper cut-off frequencies of 0.1 Hz and 40 Hz. Electrode recordings were re-referenced to the average of the mastoid recordings (average mastoid reference, TP9 & TP10). Manual inspection was first performed to locate and remove clearly visible disturbances in the data. Epochs were created from 1000 ms before to 2000 ms after stimulus onset, and the pre-onset average was subtracted from the post-onset signal (baseline correction). Independent components analysis (ICA) was used to extract blinking and eye movements within the data. Independent components (ICs) that were identified by the data analyst as ocular artifacts were rejected. Finally, epochs were averaged for each N-Back variant and baseline corrected using 200 ms before stimulus onset.

Pipeline II. Pipeline II was conducted by using Matlab R2016a (Mathworks, Natick, MA, USA). The data was resampled to 1000 Hz and filtered in the 0.1 – 30 Hz range using a zero-phase 4^{th}-order Butterworth filter. All electrodes were re-referenced offline to the average of the two mastoid signals (average mastoid reference, TP9 & TP10; Luck, 2014). Epochs were created from 200 ms before to 1000 ms after stimulus onset, and baseline correction was performed by subtracting the average of the 200 ms pre-stimulus onset

signal from the 1000 ms post-stimulus onset signal. The EOG recorded before and during the experiment was used for correcting the EEG signal for eye artifacts using Croft and Barry's aligned-artifact average (AAA) procedure (Croft, & Barry, 2000). Finally, epochs with EEG signals greater than 50μV were also excluded as they could signify motion artifacts (Wittevrongel & Van Hulle, 2016; Van Vliet et al., 2016). This Pipeline has been developed by the computational neuroscience group at KU Leuven (van Vliet et al., 2014, 2016; Wittevrongel & Van Hulle, 2016) and since then used in dozens of published studies from this group (http://lirias.kuleuven.be/cv?Username=U0013308). The method was developed as it accounts for eye artifacts using an automatic procedure (aligned-artifact average (AAA) procedure in Croft, & Barry, 2000) rather than having to rely on a post-hoc ICA analysis where the data analyst needs to identify which IC's contain those artifacts (as in EEGLab).

Statistical Analysis

To assess the effect of N-Back task variations on behavioral responses (average of correct responses across trials) and ERP morphology (we considered the same three midline location electrodes: Fz, Cz, and Pz investigated by Watter, Geffen & Geffen (2001), we used nonparametric permutation-based tests (Derrick, White, & Toher, 2018; Guo, & Yuan, 2017) as our datasets failed the Shapiro-Wilk test of normality (Shapiro, & Francia, 1972) and the Levene test of equality of variances (Levene, 1960). Specifically, Dataset I utilized a mixed within/between design where each participant performed 4 out of 9 variations. The rationale for using a mixed design was to obtain enough power –16

participants– for each of the 9 variations by recruiting only 36 subjects. Therefore, we used a nonparametric permutation-based test to account for the mixed (within/between) design (Efron & Tibshirani, 1993). The null hypothesis distribution is generated empirically regardless of any assumptions about the data distribution. The observed results were then assessed relative to the empirical null hypothesis distribution (Collingridge, 2013) and the *p* value was calculated by comparing the absolute distance between observed values of two groups to the absolute of the empirical null distribution (Cohen, 2017). The results were considered statistically significant when the *p*-value was less than 0.05. We ran 30.000 iterations for permutation testing of behavioral data and 3.000 for ERP data. We adopted the same statistical tests for the comparison between datasets (UCR, KU Leuven, and UM), and for ERP and behavioral data comparisons respectively. We note that this *p*-value is monotonically relatable to other measures of reliability, such as differences in signal to noise ratio (SNR).

Furthermore, we performed a power analysis for accuracy to ensure that our samples, considering the significant results of Figure 2. 2, were large enough. Here, we reported the comparison between *task* 1 and *task* 3 for *words* that revealed that 14 subjects were sufficient to support a power of 80%, for *colors* that showed that 16 subjects were sufficient to support a power of 80%, and for *pictures* that demonstrated that 22 subjects were sufficient to support a power of 80%. Although the latter did show that a bigger sample size would be necessary, we believe that it does not significantly affect our results. We

used Matlab (sampsizepwr) to conduct the power analysis and small to medium effect sizes supported our analysis (i.e., Cohen's d = 0.25).

ERP Component. We investigated the following ERP components in the 0–800 ms post-stimulus time window:

P100 (P1), a positive deflection with a peak around 100 ms after stimulus presentation. It is distributed over the lateral occipital electrodes and reflects early sensory processing of visual stimuli. P1 latency depends on stimulus contrast, such as luminance or SNR, while its amplitude is modulated by attention (Hillyard, Vogel, & Luck, 1998) and discrimination processes (Vogel & Luck, 2000).

N100 (N1), a negative deflection that peaks around 100 – 200 ms after stimulus onset. It has a distribution over the entire scalp, but it peaks earlier over the frontal regions of the scalp. It has been shown that its amplitude is modulated by attention. Larger amplitude is associated with attended stimuli, while smaller is associated with increasing stimulus presentation frequency (Luck, Heinze, Mangun, & Hillyard, 1990). N1 latency is affected by cognitive processing effort: the bigger the effort, the longer the latency (Callaway, & Halliday, 1982).

P200 (P2), a positive deflection with a peak around 150 – 275 ms after stimulus presentation. It is distributed over the fronto-central and parieto-occipital regions of the scalp, but its maximal is over the frontal area. It is elicited by visual stimuli and modulated by attention (Liu, Zhang, Ma, Li, Yin, and Luo, 2013). Its amplitude is suppressed with

increasing of attentiveness (Kanske, Plitscha, & Kotz, 2011) and more frequent target stimuli (Lu, Williamson, & Kaufman, 1992).

N200 (N2), a negative deflection detected around 200 – 350 ms after stimulus onset. It is distributed over the frontal regions of the scalp and posterior regions in visual attention tasks (Folstein & Van Petten, 2008). N2 component reflects several functions such as stimulus identification, attentional shift and motor response inhibition (Patel & Azzam, 2005).

P300 (P3), a positive deflection with a peak occurring around 250 – 600 ms after stimulus onset. It shows a stronger distribution over the centro-parietal electrodes on the scalp. Its amplitude becomes larger with infrequent target stimuli and decreases with habituation and task difficulty. Its latency is modulated by the difficulty in discriminating the target stimulus (Picton, 1992; Polich and Kok, 1995).

N400 (N4), a negative deflection detected between 400 – 600 ms after stimulus onset. It is typically stronger over centro-parietal regions of the scalp and reflects brain response to semantically meaningful stimuli that can include visual and auditory words, sounds, pictures and faces (Kutas & Federmeier, 2011). N4 amplitude is affected by priming and frequency of the stimulus (Van Petten & Kutas, 1990).

Positive Late Component (PLC), a positive deflection, with a peak occurring around 500-1000 ms after stimulus onset. It is most prominent for posterior scalp channels. The PLC amplitude is modulated by stimulus repetition: suppressed for stimuli that have been already presented, and generally larger for new stimuli (i.e. ‘old-new’ effect), in both long-

and short-term memory paradigms (Danker, 2008; Olichney, 2000). PLC is believed to index top-down allocation of attention in a memory recollection process (Mecklinger, 2000).

Results

Effect of stimulus type and task structure – Dataset I (UCR)

In order to investigate the effect of experimental features (*stimulus type* and *task structure* type), we performed a nonparametric permutation-based analysis on behavioral and electrophysiological data.

Behavioral Results. While *pictures* were associated with the highest accuracy level when holding *task structure* constant (Figure 2. 2A; see Tables S1, S2 and S3 in Supplemental Material for means, standard deviations and statistics per condition, respectfully), there was no statistically significant effect of *stimulus type* on accuracy ($p>.2$ for all conditions other than for *task* 1: *words* vs. *picture*s: $p = .075$). On the other hand, results revealed a robust overall effect of *task structure* (see Figure 2. 2B), showing higher accuracy for *task* 3 vs *task* 1 ($p<.002$ for all *stimulus types*), and for *task* 3 vs *task* 2 for *words* ($p<.001$) and *colors* ($p<.012$) but only a trend for *pictures* ($p=.067$). However, there was no statistically significant difference between *task* 1 and 2 ($p>.4$ for all conditions other for *words* $p = .074$). These results show that while there is a highly significant effect of tasks, especially *task* 3 vs the others, that the choice of stimulus has a lesser effect on task performance.

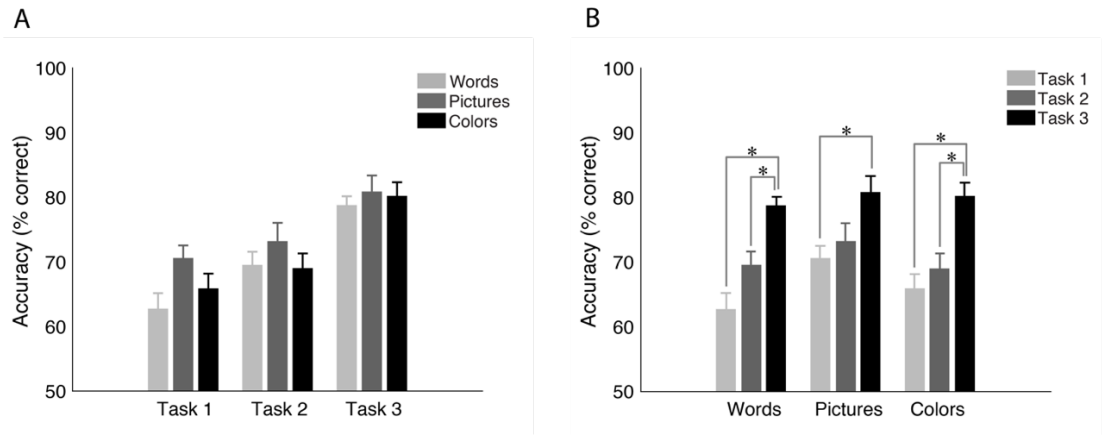

Figure 2. 2. Mean accuracy and SEM for target trials in the UCR dataset. (A) Accuracy as a function of task type. (B) Accuracy as a function of stimulus type. * indicates significance of p<.05

ERP Morphology. Overall, ERP morphologies changed substantially both as a function of *stimulus type* and *task structure*. This can be seen in Figures 2. 3 and 4 for channel Cz, while channels Fz and Pz are shown in Supplemental Material (Figures S2. 1-4). We also presented topographies and reported differences between them in the Supplemental Material (see Figure S2. 9 and Tables S2. 6-7). In the following sections, we highlight some of the significant effects by running permutation tests that demonstrate the extent to which various differences in morphology across the time-course are different as a function of condition. Significant differences discussed below are in regard to shaded regions in graphs that indicate periods in the ERPs where differences are *p*-value of less than or equal to 0.05 for at least 12 consecutive bins with Δt of 1/512 Hz.

Effects as a function of stimulus type. For *task 1*, ERP morphologies differed more frequently for *pictures* compared to *colors* and *words*, as seen in Figure 2. 3. While *pictures* vs *words* differed more frequently in the N1, N2, and P2 components in channels Fz, Cz,

and Pz (the latter only for P2), *pictures* vs *colors* showed differences in the N2, P2, and P3 components in channels Fz and Cz. Additionally, *words* vs *colors* showed differences in the P2 component in channels Fz, Cz, and Pz.

For *task 2*, we found that ERP morphologies differed more frequently for *colors* compared to *pictures* and *words* (see Figure 2. 3). While both *colors* and *words* differed from *pictures* more frequently in the N1 component in channels Cz, Pz, and Fz respectively, *colors* vs *words* and *colors* vs *pictures* showed differences mostly in the P2 component in channels Fz and Cz. Additionally, *words* differed from *pictures* and *colors* in the N2 component for channel Cz, while *colors* compared to *pictures* differed more frequently in P3 component for channels Fz and Cz.

For *task 3*, ERP morphologies differed more frequently for *words* compared to *pictures* (see Figure 2. 3). *words* and *pictures* showed differences in the N2, P2, and P3 components in channel Fz and Cz. Additionally, *words* differed from *colors* in the P3 component in channel Fz and Cz.

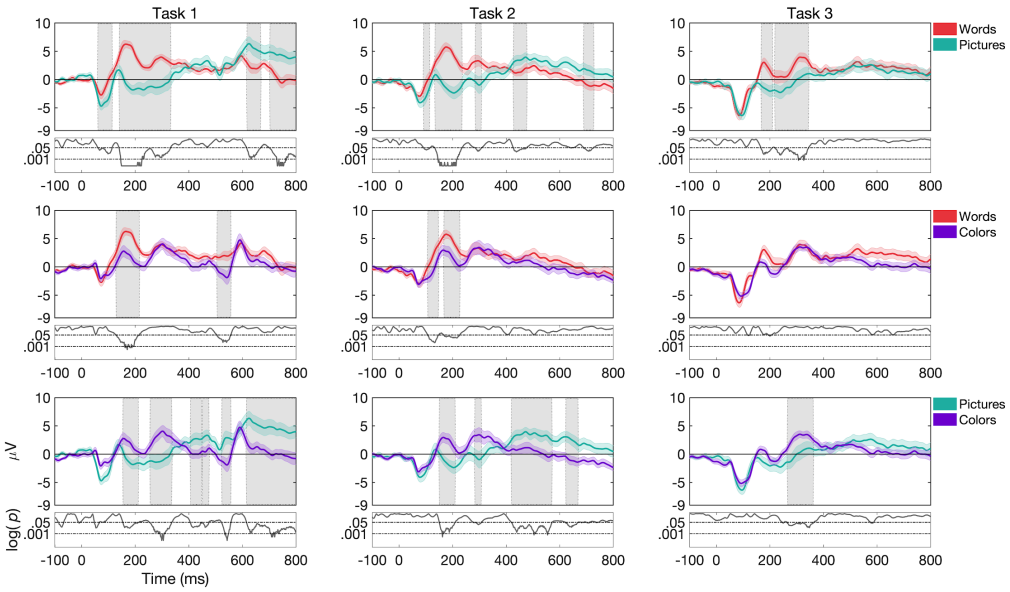

Figure 2. 3. Grand average and SEM of ERP curve for UCR dataset at Cz electrode for target trials during variations of stimulus types (words, pictures and colors). Gray shaded areas indicate significantly different data points (p<.05). P-values that are less than 0.0001 are thresholded to 0.0001 for viewing purposes, as shown by the black curve at the bottom of each graph where log p-values are reported.

Effect as a function of task structure. For *words*, ERP morphologies differed more frequently for *task* 3 compared to *task* 1 and *task* 2 (see Figure 2. 4). While *task* 3 and *task* 1 showed differences in the N1, P1 and P2 components in channels Fz, Cz and Pz, *task* 3 and *task* 2 showed differences in the N1 and P2 components in channels Fz, Cz and Pz. We do note, that in the case of where the stimulus offset occurred at 400 ms, waveforms after 400 ms may have been impacted by a stimulus offset event in addition to other task-related factors.

For *pictures*, ERP morphologies differed more frequently for *task* 1 compared to *task* 3 and *task* 2 (Figure 2. 4). While *task* 1 and *task* 3 showed differences in the N1 component in channels Fz and Cz, *task* 1 and *task* 2 showed differences in the P2 component in channels Fz, Cz and Pz. Additionally, *task* 3 differed from *task* 2 in the N1 component in channels Fz and Cz.

For *colors*, ERP morphologies differed more frequently for *task* 3 compared to *task* 1 and *task* 2 (Figure 2. 4). While *task* 3 and *task* 1 showed differences in the N1 and P2 components in channels Fz, Cz and Pz, *task* 3 and *task* 2 showed differences in the N1 and P2 components in channels Cz and Pz.

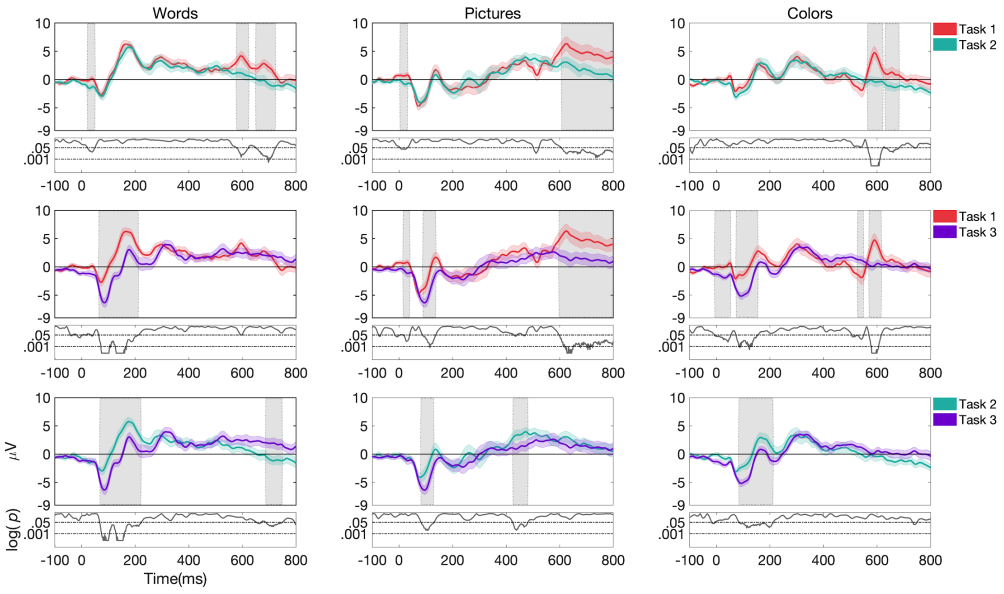

Figure 2. 4. Grand average and SEM of ERP curve for UCR dataset at Cz electrode for target trials during variations of task structure (task 1, task 2 and task 3). Gray shaded areas indicate significantly different data points ($p<.05$). P values that are less than 0.0001 are thresholded to 0.0001 for viewing purposes.

Effects as a function of task load and performance. To understand how other factors may have influenced the ERPs, we also examined effects of memory load and performance on ERP waveforms (see Figure 2. 5). In regard to N-back load (N = 2, N = 3), the main effect of load is shown in Figure 2. 5-A with this effect of load being significant (p<0.05) for all the components mentioned in this paper except for P1 (see Table S2. 4 for stats). However, this effect was largely independent of *task*, and *stimulus* types (see Figure S2. 5 and Table S2. 4 for break-down of ERPs and stats across the different *task and stimulus* conditions). Likewise, we also observed differences in the ERPs as a function of metrics of performance (Figure 2. 5-B); hits (correctly responded targets), misses (incorrectly responded targets), correct rejections (correctly responded non-targets), and false alarms (incorrectly responded non-targets). There is a significant main effect of performance (p<0.001) for all the components, except for P1. However, again, this effect was largely independent of *task*, and *stimulus* types (see Figure S2. 6 and Table S2. 5 for ERPs for break-down of ERPs and stats across the different *stimulus* and *task* conditions).

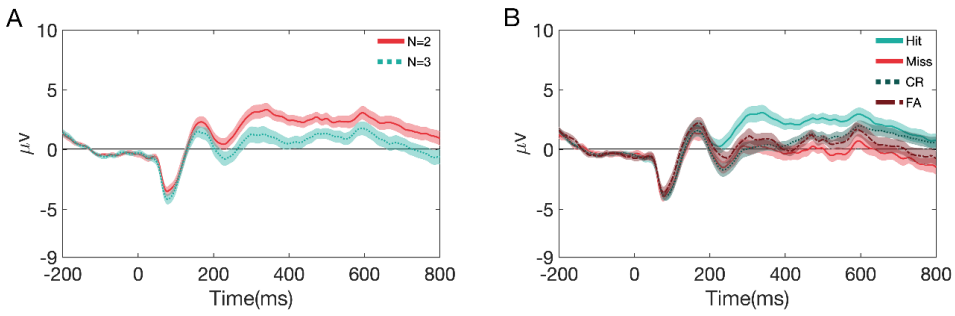

Figure 2. 5. Grand average and SEM of ERP curve for UCR dataset at Cz electrode as a function of N-back load (A) and performance metrics (B).

Comparison between Pre-processing Pipelines in Dataset I (UCR)

We next examined the extent to which differences in analysis pipelines used across labs resulted in changes in estimated ERP morphologies. Interestingly, early ERP components are relatively preserved across the pipelines, but that later ERP components showed significant differences between pipeline I and pipeline II (see Figure 2. 6). Further, these differences showed some interaction with task and stimulus. For example, the effect of pipeline was found in all variations in *task structure* 1 (for channels Fz, Cz and Pz). Moreover, the Word N-Back variation with *task structure* 1 showed significant differences in P3 components between the two pipelines. For *task structure* 2 and *words*, Cz showed significant difference in N2 and P3 components. For *task structure* 2 and *pictures*, Fz revealed significant differences in PLC, and Cz in P3 and PLC. For *task structure* 2 and *color* stimulus, Fz showed significant differences in P3, N4 and PLC signatures and Cz in P3 and PLC. For t*ask structure* 3 and *words*, Fz showed significant differences in N2, P3, N4 and PLC components. Further, Cz showed differences for N2 and P3 and Pz for PLC. For *task structure* 3 and *pictures*, Fz and Cz showed significant difference in PLC. Finally, for *task structure* 3 and *colors*, Fz showed significant differences in P3, N4 and PLC components and Cz showed significant difference in PLC.

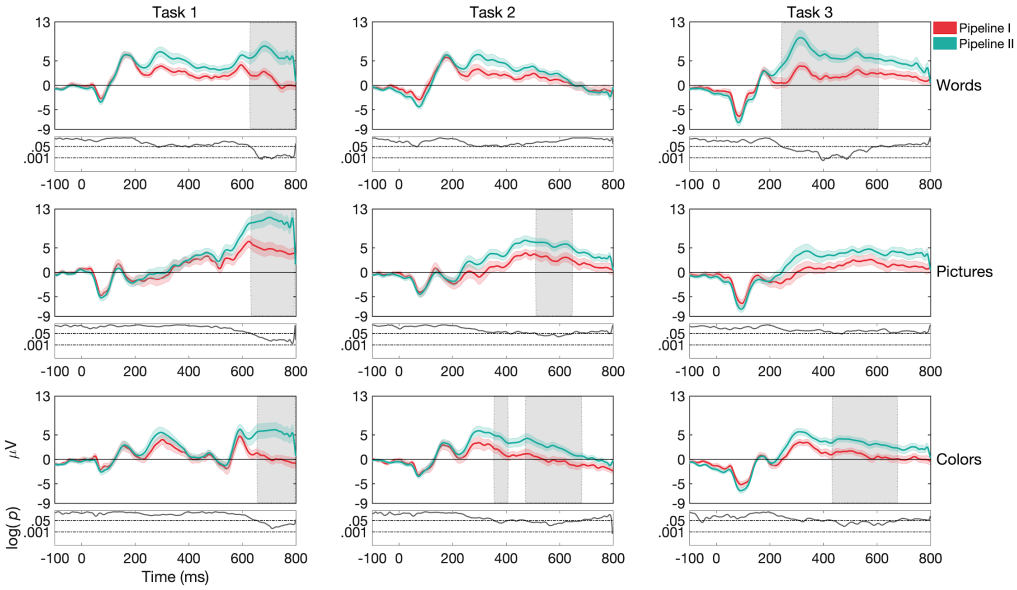

Figure 2. 6. Grand average and SEM of ERP curve at Cz electrode for target trials for different pipelines (Pipeline I vs. Pipeline II) for the UCR dataset (see Supplemental Material for Fz and Pz, Figures S2. 7-8). Gray shaded areas indicate significantly different data points ($p<.05$). P values that are less than 0.0001 are thresholded to 0.0001 for viewing purposes. Data in Pipeline II was up-sampled to 512 to make the comparison possible.

Due to the fact that Pipelines I and II differ in several ways ranging from analysis toolbox, eye artifacts removal to reference electrodes etc., there are too many candidate parameters to be causally related to a specific difference in an ERP component. Nevertheless, these results are interesting as they highlight how the use of different pre-processing pipelines commonly used in the EEG literature can affect ERP morphology at an aggregate level, and in particular the choice of pipeline can impact the extent to which one correctly/incorrectly determines differences between conditions. While it would be interesting to unveil possible causal relations between these differences in the pipeline, the goal of the present study is to illuminate the impacts of common methodological

differences between studies rather than to fully explain such differences, which would require a larger study. Furthermore, considering the few existing studies in literature (Jiang, Bian, & Tian, 2019; Yao et al., 2019; Dong et al., 2019) that demonstrated a significant role played by pre-processing factors, we think it is likely that the eye artifacts removal method and reference electrodes might have greatest impacts in our pipelines on the resulting ERPs. Still, we note that our analysis of pipeline is merely illustrative of how the pipelines used in the previously published versions of these datasets give rise to different ERP morphologies and that a complete characterization of how pipeline elements effect the signal and/or SNR (Robbins, Touryan, Mullen, Kothe, & Bigdely-Shamlo, 2020) is beyond the scope of the present manuscript.

Laboratory Effects

Another potential aspect of variation is experimental location resulting in behavioral and ERP morphology differences. Specifically, we refer to different laboratories in order to explore differences in several characteristics such as lab settings, stimuli, tasks, subject pools, subject instructions, processing pipelines, and so on. Using pipeline I, we compared *task* 2 (*pictures* only, N = 16) as used in Dataset I (UCR) and Dataset II (KU Leuven), as well as *task* 1 (*words* only, N = 16) which was used in Dataset I (UCR) and Dataset III (UM). We did not compare Datasets II and III as the stimuli were different: *pictures* vs. *words* respectively, whereas Dataset I included both *words* and *pictures* and could therefore be compared to both datasets. We note, that while this analysis is far from comprehensive and it would be ideal to collect data on identical procedures across the labs,

however, this is at least illuminative of other, unexplained, variance that can be expected from different labs conducting similar research but not coordinating on the exact details of the studies, which is typical of the extant literature.

Dataset I vs. Dataset II (UCR vs KU Leuven). Behavioral results for *task 2* showed a significantly higher accuracy in Dataset II compared to Dataset I ($p < .001$) (Figure 2. 7-A) and ERP morphology outcomes revealed larger ERP amplitudes in Dataset II compared to Dataset I (Figure 2. 8). Namely, significant differences between Dataset I and Dataset II ($p < .05$) were found in P1, N1, P2, N2 and P3 components, in channels Fz and Cz.

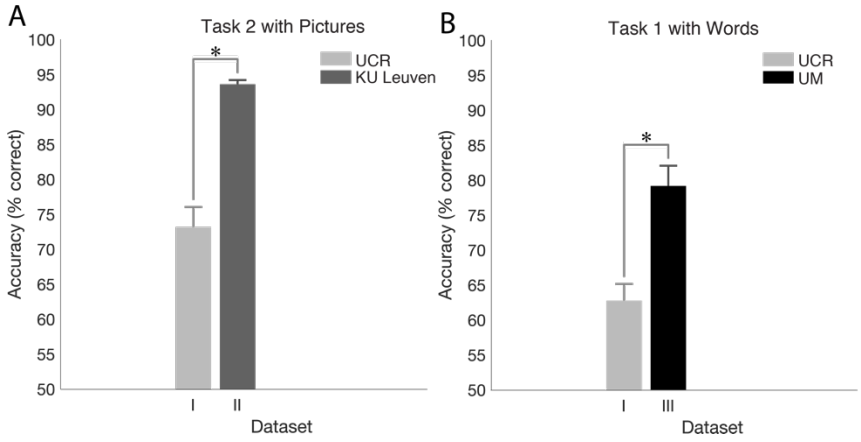

Figure 2. 7. Cross- laboratory accuracy comparison. (A) Accuracy for N-Back task 2 with pictures in dataset I (UCR) and in dataset II (Ku-Leuven). (B) Accuracy for N-Back task 1 with words in dataset I (UCR) and in dataset III (UM).

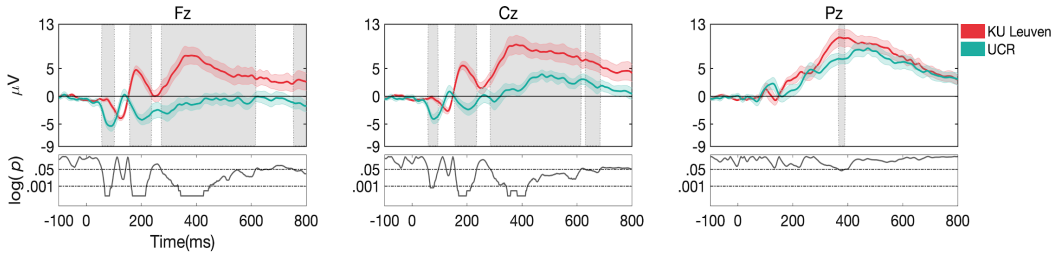

Figure 2. 8. ERP responses during task 2, only for target stimuli recorded at different laboratories. Gray shaded areas show significant differences at p<.05. Both datasets were pre-processed with pipeline I.

Dataset I vs. Dataset III (UCR vs UM). For *task 1*, higher accuracy was observed in Dataset III compared to Dataset I ($p < .001$) (Figure 2. 7-B), and ERP morphology (Figure 2. 9) indicated significant differences in P1, N1, P2, N2 and P3 components, in channels Fz, Cz and Pz.

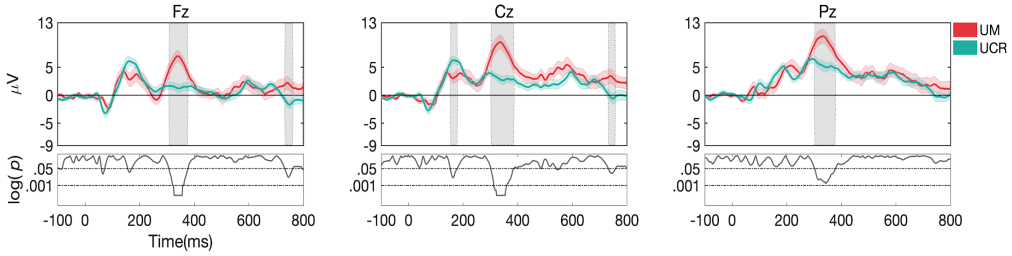

Figure 2. 9. ERP responses during task 1 (mean and standard deviation of targets) recorded at different laboratories. Gray shaded areas show significant differences at p<.05. Both datasets were pre-processed with pipeline I.

Discussion

The goal of the present study was to fill a gap in the extant literature by illuminating the extent to which common procedural differences related to N-back task variants, EEG recording setups, and preprocessing pipelines affect behavioral and electrophysiological

correlates of performance. To address this, we compared variants of the N-Back task used in 3 laboratories, 2 in Europe (Pahor, & Jaušovec, 2018; Pergher et al., 2018) and 1 in the US where the behavioral and EEG datasets were replicated. Our findings suggest that *stimulus type*, *task structure*, pre-processing pipeline and lab factors contribute to differences in behavioral and neurophysiological responses on the N-Back task.

Given the fact that most meta-analyses overlook differences in the N-Back task adopted in each study (Glahn et al., 2005; Redick, & Lindsey, 2013; Brunoni, & Vanderhasselt, 2014; Heishman, Kleykamp, & Singleton, 2010), we characterized some of those factors that might affect cognitive task outcomes. First we examined *task structure* and showed differences in accuracy level between tasks (*task* 1, *task* 2, *task* 3), revealing higher accuracy for *task* 3 compared to the other two, perhaps due to having the longest stimulus duration (2500 ms) thereby supporting the process for encoding of information that is facilitated when stimulus duration is longer. Indeed, Kunimi (2016) showed that increasing stimulus duration (from 500 ms to 5000 ms) improves memory performance during retention of visuospatial information, whereas Fox, Snyder, Vincent, & Raichle (2007) showed that longer ISI was associated with increased accuracy level. We also investigated *stimulus type* and observed better performance for *pictures* of objects compared to *words* and *colors*. In contrast, Nystrom et al. (2000) reported higher accuracy for letters compared to shapes.

Another important aspect when considering the following factors such as *task structure* and *stimulus type* is their impact on ERP morphology. To highlight this variance,

we examined differences in several ERP components, named N1, P2, N2 and P3 for both factors, as previous studies suggested ERP component modulation in response to WM experimental features, particularly for *stimulus type*, and observed their spatial distribution. Mecklinger and Pfeifer (1996) reported that the encoding of object features was associated with modulation of P2 component, whereas Ruchkin, Johnson, Grafman, Canoune, and Ritter (1992) showed variations of N2 and P3 components for visuo-spatial stimuli compared to phonological stimuli, indicating that visuo-spatial stimuli were processed more quickly than phonological ones. Moreover, Rossion, Joyce, Cottrell, and Tarr (2003) observed N1 modulation in response to faces and objects compared to objects. Thus while it is clear in the literature that both task and stimulus should influence ERPs in systematic ways, to date this has been largely overlooked in the literature examining ERP signatures of working memory tasks such as the N-Back.

In addition to *stimulus type* and *task structure*, we suggest that different experimental laboratories and pre-processing procedures might also affect accuracy and ERP morphology. Seemingly arbitrary procedures are employed by different laboratories, in terms of environment and equipment, as well as data pre-processed and analyzed by different pipelines, which have been shown to produce different findings (Busch, Herrmann, Müller, Lenz, & Gruber, 2006; for review, see Zimmer, Cohen, Foley, Guynn, Engelkamp, & Kormi-Nouri, 2001). Here we show that the same N-Back task performed in two laboratories produces different behavioral and ERP morphology results. However, we suggest interpreting these results carefully, as participants' individual differences and

EEG and analysis operator skills may also have affected these results (Jaeggi et al., 2014). Green et al. (2018) observed that reward, motivation and/or participant expectations, such as differences in task performance, researcher instructions, etc., could also count as factors for behavioral differences when comparing performances between different laboratories. Moreover, we highlight the impact of pre-processing pipelines on ERP data, supporting the recommendations provided by Smith, & Kutas (2015) regarding the power of EEG data pipeline, including baseline correction, artifact rejection and the filtering procedure (Acunzo, MacKenzie, & van Rossum., 2012) on ERP analysis. In line with the goal of this study, we did not associate a specific step of the pre-processing pipeline procedure to an ERP component or cognitive process since we aimed to show at a more general level the impact of stimuli, task and laboratory environment on both accuracy and ERP responses.

Our study presents several limitations. We considered only accuracy during N-Back performance, due to the fact that the three Datasets and the related tasks had different response requirements, and so it would have been very complex to compare them. As Dataset I utilized a mixed within/between design, individual differences might have affected ERP signatures attributed to laboratory effects. Indeed, a recent review paper highlighted the variety of features that may impact N-Back performance, including both task and individual features (Pergher et al., 2019). The samples compared here were of similar age and had a similar educational level (undergraduates), and in Datasets I and II, a similar distribution of gender. While Dataset III only consisted of data collected from females, a recent study by Pliatsikas et al. (2019) demonstrated that gender, age, and

education level affect response accuracy after a single N-Back training session in healthy older individuals. Since the present study consists of N-Back performance across 1 or 2 sessions in young subjects, we do expect these variables to have a moderate effect on behavioral and electrophysiological results. Nevertheless, there might be other individual difference factors such as motivation, personality, and working memory capacity (Dong et al., 2015) that were not accounted for but could have affected the results. Future studies will need to examine whether these individual differences, along with other factors such as time-of-day and environment, affect N-Back task performance and ERP signatures. Moreover, further studies should also consider the choice of words, pictures and colors, as they may play an important role in affecting behavioral and ERP responses due to different colors and shape used, and familiarity with the objects presented. Finally, since Dataset III represents a sham condition in a brain stimulation study (Pahor & Jaušovec, 2018) it is possible that placebo effects could have affected performance. Since we only retained data collected in session 1, i.e. prior to exposure to active stimulation, it is unlikely that these effects are large. Still, we suggest that while more work can be done to clarify the effects presented here, and that other differences still exist in the extant literature, that the present work is informative of how some of the most common differences in the N-Back between studies can impact observed behavioral and physiological measures.

In conclusion, the present data sets help clarify the extent to which common N-Back task variations in terms of *stimulus type*, *task structure*, and laboratory and processing pipeline give rise to differences in behavioral and physiological outcomes. While future

research is needed to help us understand the mechanisms that underly these observed differences, the present work can help readers appreciate effect sizes to be expected related to the many variations considered here. We note that while, in general, it is well acknowledged any difference between studies can have an impact, the significance of these variations in the case of the N-Back have been largely overlooked, thus limiting understanding of their role in affecting accuracy and ERP morphology and of potential important information related to the mechanisms that regulate WM processes. We suggest that for the field to move forward, experimental features, analysis pipeline, and laboratory differences need to be taken into consideration when interpreting findings and making comparisons across studies.

Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Author Contributions

V.P., M.A.S., and A.S. conceived the presented idea and developed the theory. M.A.S., V.P., and A.P. carried out the experiment, M.A.S. and V.P performed the computations. All authors revised the findings and wrote the manuscript.

Funding

This research was supported by NIMH R01 MH111742 to ARS, a research grant to VP from a special research fund project (C24/18/098) of the KU Leuven, and by research

grants to MMVH from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 857375, the Financing Program (PFV/10/008) and the special research fund of the KU Leuven (C24/18/098), the Belgian Fund for Scientific Research -- Flanders (G088314N, G0A0914N, G0A4118N), the Interuniversity Attraction Poles Programme -- Belgian Science Policy (IUAP P7/11), and the Hercules Foundation (AKUL 043).

References

Acunzo, D. J., MacKenzie, G., & van Rossum, M. C. (2012). Systematic biases in early ERP and ERF components as a result of high-pass filtering. Journal of neuroscience methods, 209(1), 212-218.

Anguera, J. A., Bernard, J. A., Jaeggi, S. M., Buschkuehl, M., Benson, B. L., Jennett, S., ... & Seidler, R. D. (2012). The effects of working memory resource depletion and training on sensorimotor adaptation. Behavioural Brain Research, 228(1), 107-115.

Baddeley, A. (2012). Working memory: theories, models, and controversies. Annual review of psychology, 63, 1-29.

Blacker, K. J., Negoita, S., Ewen, J. B., and Courtney, S. M. (2017). N-back versus complex span working memory training. Journal of Cognitive Enhancement, 1(4), 434-454.

Brainard, D. H. (1997). The psychophysics toolbox. Spatial vision, 10(4), 433-436.

Brouwer, A. M., Hogervorst, M. A., Van Erp, J. B., Heffelaar, T., Zimmerman, P. H., & Oostenveld, R. (2012). Estimating workload using EEG spectral power and ERPs in the n-back task. Journal of neural engineering, 9(4), 045008.

Brunoni, A. R., & Vanderhasselt, M. A. (2014). Working memory improvement with non-invasive brain stimulation of the dorsolateral prefrontal cortex: a systematic review and meta-analysis. Brain and cognition, 86, 1-9.

Busch, N. A., Herrmann, C. S., Müller, M. M., Lenz, D., & Gruber, T. (2006). A cross-laboratory study of event-related gamma activity in a standard object recognition paradigm. Neuroimage, 33(4), 1169-1177.

Buschkuehl, M., Hernandez-Garcia, L., Jaeggi, S. M., Bernard, J. A., & Jonides, J. (2014). Neural effects of short-term training on working memory. Cognitive, Affective, & Behavioral Neuroscience, 14(1), 147-160.

Callaway, E., & Halliday, R. (1982). The effect of attentional effort on visual evoked potential N1 latency. Psychiatry research, 7(3), 299-308.

Chen, Y. N., & Mitra, S. (2009). The spatial-verbal difference in the n-back task: an ERP study. Acta Neurol Taiwan, 18(3), 170-179.

Chen, Y. N., Mitra, S., & Schlaghecken, F. (2008). Sub-processes of working memory in the N-back task: an investigation using ERPs. Clinical Neurophysiology, 119(7), 1546-1559.

Cohen, M. X. (2017). MATLAB for brain and cognitive scientists. MIT Press.

Colagiuri, B., Schenk, L. A., Kessler, M. D., Dorsey, S. G., & Colloca, L. (2015). The placebo effect: from concepts to genes. Neuroscience, 307, 171-190.

Collingridge, D. S. (2013). A primer on quantized data analysis and permutation testing. Journal of Mixed Methods Research, 7(1), 81-97.

Croft, R. J., & Barry, R. J. (2000). Removal of ocular artifact from the EEG: a review. Neurophysiologie Clinique/Clinical Neurophysiology, 30(1), 5-19.

Danker, J. F., Hwang, G. M., Gauthier, L., Geller, A., Kahana, M. J., & Sekuler, R. (2008). Characterizing the ERP Old–New effect in a short-term memory task. Psychophysiology, 45(5), 784-793.

De Ribaupierre, A., & Lecerf, T. (2006). Relationships between working memory and intelligence from a developmental perspective: Convergent evidence from a neo-Piagetian and a psychometric approach. European Journal of Cognitive Psychology, 18(1), 109-137.

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.

Derrick, B., White, P., & Toher, D. (2018). Parametric and non-parametric tests for the comparison of two samples which both include paired and unpaired observations. Jounal of Modern Applied Statistical Methods.

Dong, L., Liu, X., Zhao, L., Lai, Y., Gong, D., Liu, T., & Yao, D. (2019). A Comparative Study of Different EEG Reference Choices for Event-Related Potentials Extracted by Independent Component Analysis. Frontiers in neuroscience, 13, 1068.

Dong, S., Reder, L. M., Yao, Y., Liu, Y., & Chen, F. (2015). Individual differences in working memory capacity are reflected in different ERP and EEG patterns to task difficulty. brain research, 1616, 146-156.

Esposito, F., Aragri, A., Piccoli, T., Tedeschi, G., Goebel, R., & Di Salle, F. (2009). Distributed analysis of simultaneous EEG-fMRI time-series: modeling and interpretation issues. Magnetic resonance imaging, 27(8), 1120-1130.

Folstein, J. R., & Van Petten, C. (2008). Influence of cognitive control and mismatch on the N2 component of the ERP: a review. Psychophysiology, 45(1), 152-170.

Fox, M. D., Snyder, A. Z., Vincent, J. L., & Raichle, M. E. (2007). Intrinsic fluctuations within cortical systems account for intertrial variability in human behavior. Neuron, 56(1), 171-184.

Glahn, D. C., Ragland, J. D., Abramoff, A., Barrett, J., Laird, A. R., Bearden, C. E., & Velligan, D. I. (2005). Beyond hypofrontality: A quantitative meta-analysis of functional neuroimaging studies of working memory in schizophrenia. Human brain mapping, 25(1), 60-69.

Green, C. S., Bavelier, D., Kramer, A. F., Vinogradov, S., Ansorge, U., Ball, K. K., ... & Facoetti, A. (2019). Improving methodological standards in behavioral interventions for cognitive enhancement. Journal of Cognitive Enhancement, 3(1), 2-29.

Guo, B., & Yuan, Y. (2017). A comparative review of methods for comparing means using partially paired data. Statistical methods in medical research, 26(3), 1323-1340.

Heishman, S. J., Kleykamp, B. A., & Singleton, E. G. (2010). Meta-analysis of the acute effects of nicotine and smoking on human performance. Psychopharmacology, 210(4), 453-469.

Herff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., & Schultz, T. (2014). Mental workload during n-back task—quantified in the prefrontal cortex using fNIRS. Frontiers in human neuroscience, 7, 935.

Hillyard, S. A., Vogel, E. K., & Luck, S. J. (1998). Sensory gain control (amplification) as a mechanism of selective attention: electrophysiological and neuroimaging evidence. Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences, 353(1373), 1257-1270.

Jaeggi, S. M., Buschkuehl, M., Perrig, W. J., & Meier, B. (2010). The concurrent validity of the N-back task as a working memory measure. Memory, 18(4), 394-412.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. Proceedings of the National Academy of Sciences, 105(19), 6829-6833.

Jaeggi, S. M., Buschkuehl, M., Shah, P., & Jonides, J. (2014). The role of individual differences in cognitive training and transfer. Memory & cognition, 42(3), 464-480.

Jasper, H. H. (1958). The ten-twenty electrode system of the international federation. Electroencephalogr. Clin. Neurophys. 10, 371–375.

Jiang, X., Bian, G. B., & Tian, Z. (2019). Removal of artifacts from EEG signals: a review. Sensors, 19(5), 987.

Kane, M. J., Conway, A. R. A., Miura, T. K., & Colflesh, G. J. (2007). Working memory, attention control, and the N-back task: A question of construct validity. Journal of Experimental Psychology. Learning, Memory, and Cognition, 33(3): 615–622.

Kanske, P., Plitschka, J., & Kotz, S. A. (2011). Attentional orienting towards emotion: P2 and N400 ERP effects. Neuropsychologia, 49(11), 3121-3129.

Kleiner, M., Brainard, D., & Pelli, D. (2007). What's new in Psychtoolbox-3?.

Krause, C. M., Sillanmäki, L., Koivisto, M., Saarela, C., Häggqvist, A., Laine, M., & Hämäläinen, H. (2000). The effects of memory load on event-related EEG desynchronization and synchronization. Clinical neurophysiology, 111(11), 2071-2078.

Kunimi, M. (2016). Effects of age, gender, and stimulus presentation period on visual short-term memory. Journal of women & aging, 28(1), 24-33.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: finding meaning in the N400 component of the event-related brain potential (ERP). Annual review of psychology, 62, 621-647.

Levene, H. (1960). Levene test for equality of variances. Contributions to probability and statistics, 278-292.

Liu, Y., Zhang, D., Ma, J., Li, D., Yin, H., & Luo, Y. (2013). The attention modulation on timing: an event-related potential study. PloS one, 8(6).

Luck, S. J. (2014). An introduction to the event-related potential technique. MIT press.

Luck, S. J., Heinze, H. J., Mangun, G. R., & Hillyard, S. A. (1990). Visual event-related potentials index focused attention within bilateral stimulus arrays. II. Functional dissociation of P1 and N1 components. Electroencephalography and clinical neurophysiology, 75(6), 528-542.

Lu, Z. L., Williamson, S. J., & Kaufman, L. (1992). Behavioral lifetime of human auditory sensory memory predicted by physiological measures. Science, 258(5088), 1668-1670.

Mecklinger, A. (2000). Interfacing mind and brain: A neurocognitive model of recognition memory. Psychophysiology, 37(5), 565-582.

Mecklinger, A., & Pfeifer, E. (1996). Event-related potentials reveal topographical and temporal distinct neuronal activation patterns for spatial and object working memory. Cognitive Brain Research, 4(3), 211-224.

Mencarelli, L., Neri, F., Momi, D., Menardi, A., Rossi, S., Rossi, A., & Santarnecchi, E. (2019). Stimuli, presentation modality, and load-specific brain activity patterns during n-back task. Human brain mapping, 40(13), 3810-3831.

Minear, M., Brasher, F., Guerrero, C. B., Brasher, M., Moore, A., & Sukeena, J. (2016). A simultaneous examination of two forms of working memory training: Evidence for near transfer only. Memory & cognition 44(7), 1014-1037.

Miyake, A., & Shah, P. (Eds.). (1999). Models of working memory: Mechanisms of active maintenance and executive control. Cambridge University Press.

Mohammed, S., Flores, L., Deveau, J., Hoffing, R. C., Phung, C., Parlett, C. M., ... & Zordan, V. (2017). The benefits and challenges of implementing motivational features to boost cognitive training outcome. Journal of Cognitive Enhancement, 1(4), 491-507.

Nystrom, L. E., Braver, T. S., Sabb, F. W., Delgado, M. R., Noll, D. C., & Cohen, J. D. (2000). Working memory for letters, shapes, and locations: fMRI evidence against stimulus-based regional organization in human prefrontal cortex. Neuroimage, 11(5), 424-446.

Oberauer, K. 2005. Binding and inhibition in working memory: Individual and age differences in short-term recognition. Journal of Experimental Psychology: General, 134(3): 368–387.

Olichney, J. M., Van Petten, C., Paller, K. A., Salmon, D. P., Iragui, V. J., & Kutas, M. (2000). Word repetition in amnesia: Electrophysiological measures of impaired and spared memory. Brain, 123(9), 1948-1963.

Owen, A. M., McMillan, K. M., Laird, A. R., & Bullmore, E. (2005). N-back working memory paradigm: A meta-analysis of normative functional neuroimaging studies. Human brain mapping, 25(1), 46-59.

Pahor, A., & Jaušovec, N. (2018). The effects of theta and gamma tACS on working memory and electrophysiology. Frontiers in human neuroscience, 11, 651.

Patel, S. H., & Azzam, P. N. (2005). Characterization of N200 and P300: selected studies of the event-related potential. International journal of medical sciences, 2(4), 147.

Pergher, V., Shalchy, M. A., Pahor, A., Van Hulle, M. M., Jaeggi, S. M., & Seitz, A. R. (2019). Divergent research methods limit understanding of working memory training. Journal of Cognitive Enhancement, 1-21.

Pergher, V., Wittevrongel, B., Tournoy, J., Schoenmakers, B., & Van Hulle, M. M. (2019b). Mental workload of young and older adults gauged with ERPs and spectral power during N-Back task performance. Biological psychology, 146, 107726.

Pergher, V., Wittevrongel, B., Tournoy, J., Schoenmakers, B., & Van Hulle, M. M. (2018a). N-back training and transfer effects revealed by behavioral responses and EEG. Brain and behavior, 8(11), e01136.

Pesonen, M., Hämäläinen, H., & Krause, C. M. (2007). Brain oscillatory 4–30 Hz responses during a visual n-back memory task with varying memory load. Brain research, 1138, 171-177.

Picton, T. W. (1992). The P300 wave of the human event-related potential. Journal of clinical neurophysiology, 9(4), 456-479.

Pliatsikas, C., Veríssimo, J., Babcock, L., Pullman, M. Y., Glei, D. A., Weinstein, M., ... & Ullman, M. T. (2019). Working memory in older adults declines with age, but is modulated by sex and education. Quarterly Journal of Experimental Psychology, 72(6), 1308-1327.

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. Clinical neurophysiology, 118(10), 2128-2148.

Polich, J., & Kok, A. (1995). Cognitive and biological determinants of P300: an integrative review. Biological psychology, 41(2), 103-146.

Redick, T. S., & Lindsey, D. R. (2013). Complex span and n-back measures of working memory: a meta-analysis. Psychonomic bulletin & review, 20(6), 1102-1113.

Robbins, K. A., Touryan, J., Mullen, T., Kothe, C., & Bigdely-Shamlo, N. (2020). How Sensitive are EEG Results to Preprocessing Methods: A Benchmarking Study. IEEE Transactions on Neural Systems and Rehabilitation Engineering, 28(5), 1081-1090.

Ross, P., & Segalowitz, S. J. (2000). An EEG coherence test of the frontal dorsal versus ventral hypothesis in N-back working memory. Brain and cognition, 43(1-3), 375-379.

Rossion, B., Joyce, C. A., Cottrell, G. W., & Tarr, M. J. (2003). Early lateralization and orientation tuning for face, word, and object processing in the visual cortex. Neuroimage, 20(3), 1609-1624.

Ruchkin, D. S., Johnson Jr, R., Grafman, J., Canoune, H., & Ritter, W. (1992). Distinctions and similarities among working memory processes: An event-related potential study. Cognitive Brain Research, 1(1), 53-66.

Scharinger, C., Soutschek, A., Schubert, T., & Gerjets, P. (2015). When flanker meets the n-back: What EEG and pupil dilation data reveal about the interplay between the two central-executive working memory functions inhibition and updating. Psychophysiology, 52(10), 1293-1304.

Scharinger, C., Soutschek, A., Schubert, T., & Gerjets, P. (2017). Comparison of the working memory load in n-back and working memory span tasks by means of eeg frequency band power and p300 amplitude. Frontiers in human neuroscience, 11, 6.

Shapiro, S. S., & Francia, R. S. (1972). An approximate analysis of variance test for normality. Journal of the American Statistical Association, 67(337), 215-216.

Smith, N. J., & Kutas, M. (2015). Regression-based estimation of ERP waveforms: II. Nonlinear effects, overlap correction, and practical considerations. Psychophysiology, 52(2), 169-181.

Efron, B., & RJ, T. (1993). An introduction to the bootstrap. Chapman and Hall, New York, NY. Farrell, J., Johnston, M. and Twynam, D.(1998),‘‘Volunteer motivation, satisfaction, and management at an elite sporting competition’’, Journal of Sport Management, 12, 288-300.

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequencyinevent-related brainpotentials. Memory & cognition, 18(4), 380-393.

Van Vliet, M., Chumerin, N., De Deyne, S., Wiersema, J. R., Fias, W., Storms, G., & Van Hulle, M. M. (2015). Single-trial erp component analysis using a spatiotemporal lcmv beamformer. IEEE Transactions on Biomedical Engineering, 63(1), 55-66.

Vogel, E. K., & Luck, S. J. (2000). The visual N1 component as an index of a discrimination process. Psychophysiology, vol. 37, no. 2, pp. 190-203, 2000.

Watter, S., Geffen, G. M., & Geffen, L. B. (2001). The n-back as a dual-task: P300 morphology under divided attention. Psychophysiology, 38(6), 998-1003.

Wittevrongel, B., & Van Hulle, M. M. (2016). Faster p300 classifier training using spatiotemporal beamforming. International journal of neural systems, 26(03), 1650014.

Yao, D., Qin, Y., Hu, S., Dong, L., Vega, M. L. B., & Sosa, P. A. V. (2019). Which reference should we use for EEG and ERP practice?. Brain topography, 1-20.

Zimmer, H. D., Cohen, R. L., Foley, M. A., Guynn, M. J., Engelkamp, J., & Kormi-Nouri, R. (2001). Memory for action: A distinct form of episodic memory?. Oxford University Press on Demand.

Supplementary Material

Table S2. 1. Mean and SD of accuracy (%) in dataset I (UCR dataset)

	Task 1	Task 2	Task 3
words	*M* = 62.76, *SD* = 14.23	*M* = 69.53, *SD* = 12.14	*M* = 78.78, *SD* = 8.20
pictures	*M* = 70.52, *SD* = 11.78	*M* = 73.18, *SD* = 17.11	*M* = 80.86, *SD* = 14.95
colors	*M* = 65.88, *SD* = 13.65	*M* = 69.01, *SD* = 13.67	*M* = 80.21, *SD* = 12.55

Table S2. 2. *p*-values for Stimulus-wise comparison for accuracy (%) in dataset I (UCR dataset)

	Task 1	Task 2	Task 3
words vs pictures	.075	.232	.589
words vs colors	.493	.893	.718
pictures vs colors	.199	.210	.851

Table S2. 3. Table S1: p-values for Task-wise comparison for accuracy (%) in dataset I (UCR dataset)

	Words	Pictures	Colors
task 1 vs task 2	.074	.452	.441
task 1 vs task 3	< .001	.002	< .001
task 2 vs task 3	< .001	.067	.012

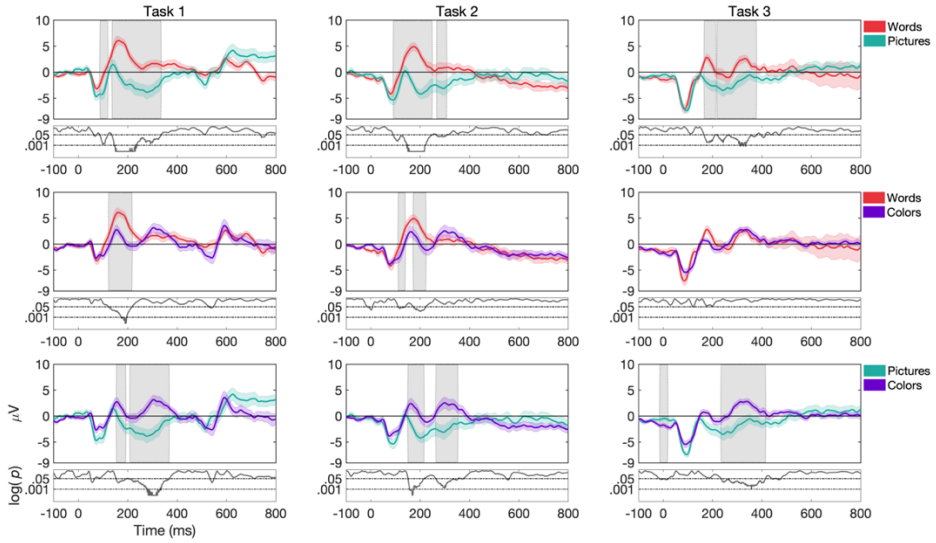

Figure S2. 1. Grand average and SEM of ERP curve at Fz electrode for target trials during variations of stimulus types (words, pictures and colors). The gray shades show p <.05

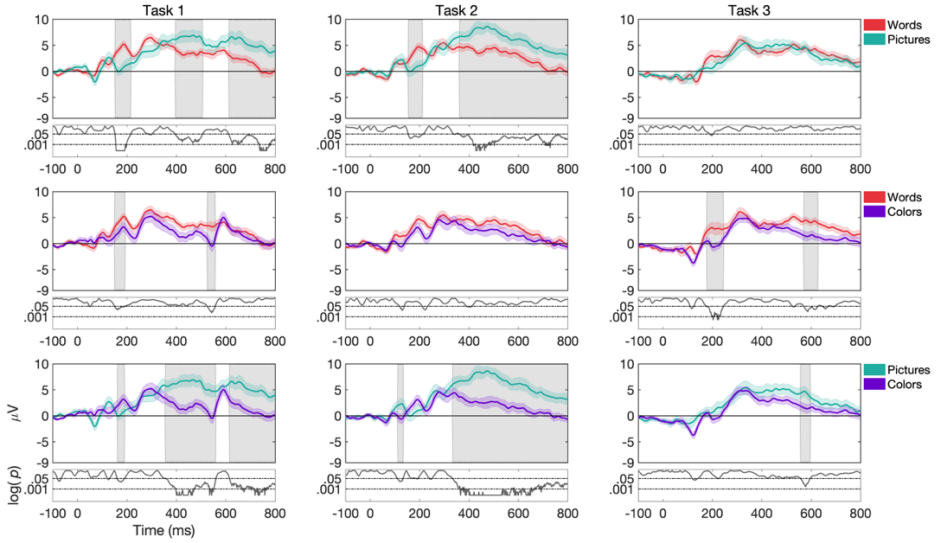

Figure S2. 2. Grand average and SEM of ERP curve at Pz electrode for target trials during variations of stimulus types (words, pictures and colors). The gray shades show p <.05

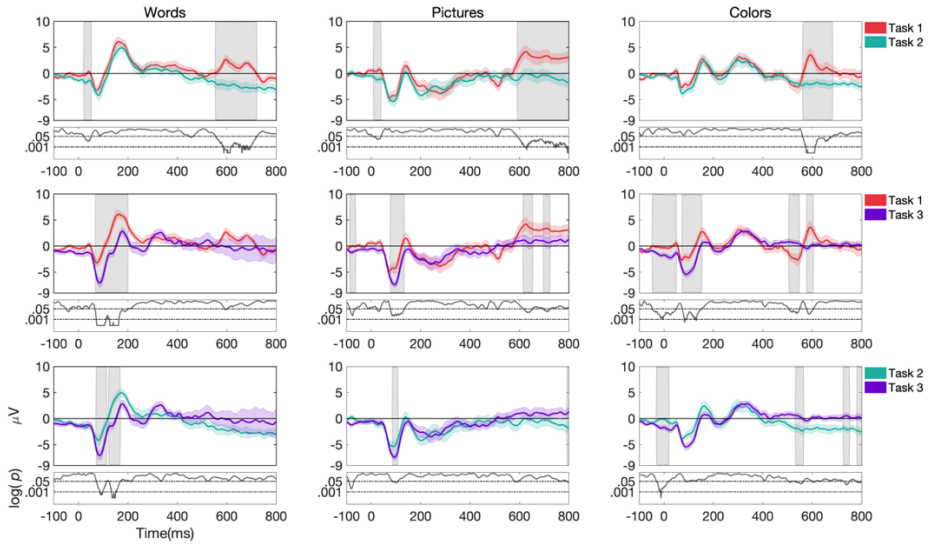

Figure S2. 3. Grand average and SEM of ERP curve at Fz electrode for target trials during variations of task structure types (task 1, task 2, task 3). The gray shades show p <0.5

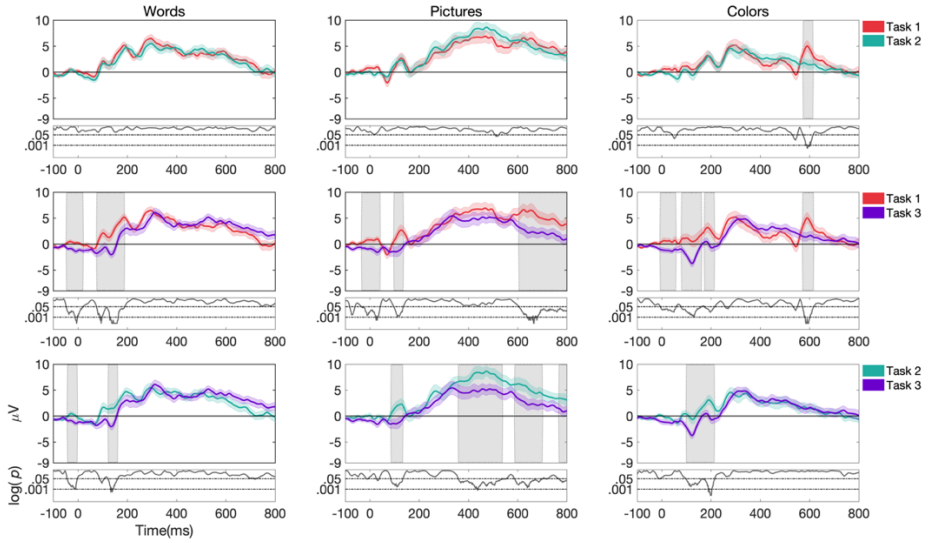

Figure S2. 4. Grand average and SEM of ERP curve at Pz electrode for target trials during variations of task structure types (task 1, task 2, task 3). The gray shades show p <0.5

Table S2. 4. Mixed ANOVA statistics for main and interaction effects of N-back load (N = 2, N = 3)

	P1	N1	P2	N2	P3	N4	PLC
task	<0.001	<0.001	<0.005	0.488	0.929	0.984	0.017
stimulus	0.163	<0.001	<0.001	<0.001	0.362	0.008	<0.001
load	0.181	0.031	<0.001	<0.001	<0.001	<0.001	<0.001
task x stimulus	0.685	0.335	0.519	0.933	0.843	0.561	0.373
load x task	0.248	0.760	0.296	0.090	0.051	0.031	0.002
load x stimulus	0.762	0.867	0.576	0.468	0.971	0.661	0.883
load x task x stimulus	0.295	0.669	0.898	0.794	0.721	0.557	0.720

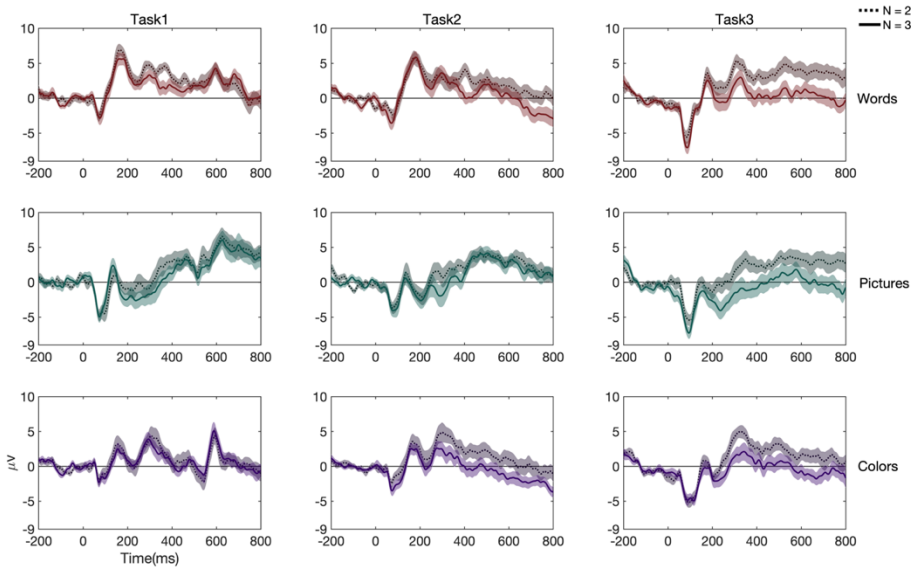

Figure S2. 5. Grand average and SEM of ERP curve at Cz electrode for N-back load (N = 2, N = 3) across task structures and stimulus types

Table S2. 5. Mixed ANOVA statistics for main and interaction effects of performance metrics (hits, misses, correct rejection, and false alarm)

	P1	N1	P2	N2	P3	N4	PLC
task	<0.001	<0.001	<0.001	0.029	0.030	0.049	<0.001
stimulus	0.006	<0.001	<0.001	<0.001	0.148	0.015	<0.001
performance	0.712	0.001	<0.001	<0.001	<0.001	<0.001	<0.001
task x stimulus	0.965	0.767	0.517	0.986	0.83	0.559	0.741
performance x task	0.739	0.236	0.432	0.121	0.031	0.07	0.029
performance x stimulus	0.376	0.708	0.209	0.025	0.004	0.011	0.003
performance x task x stimulus	0.946	0.368	0.780	0.600	0.53	0.49	0.652

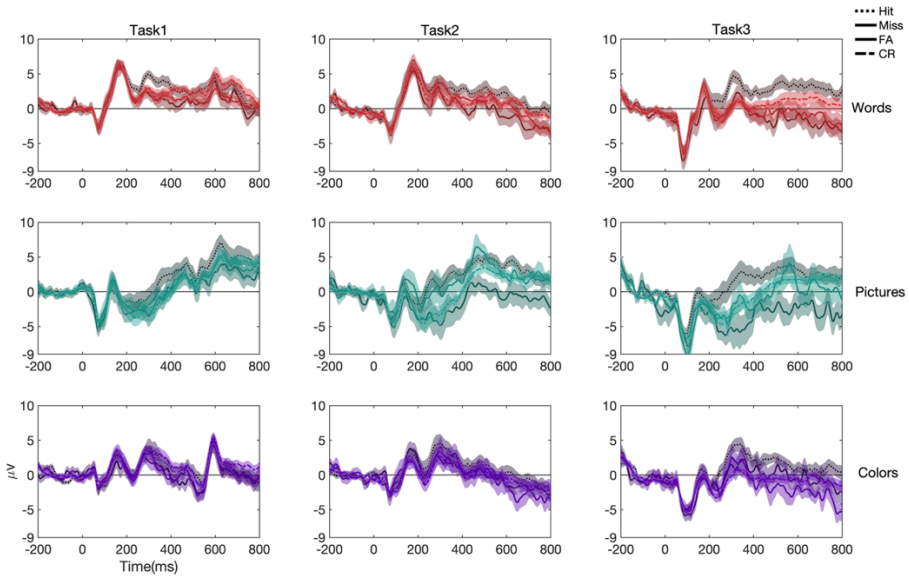

Figure S2. 6. Grand average and SEM of ERP curve at Cz electrode for performance metrics (hit, miss, correct rejection, and false alarm) across task structures and stimulus types

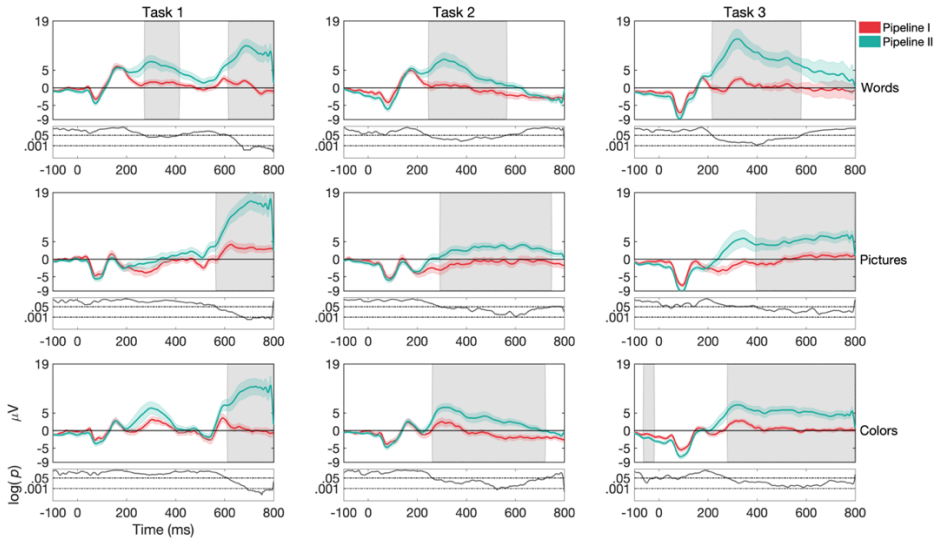

Figure S2. 7. Grand average and SEM of ERP curve at Fz electrode for target trials for different pipelines (Pipeline I vs. Pipeline II) for dataset I (UCR dataset).

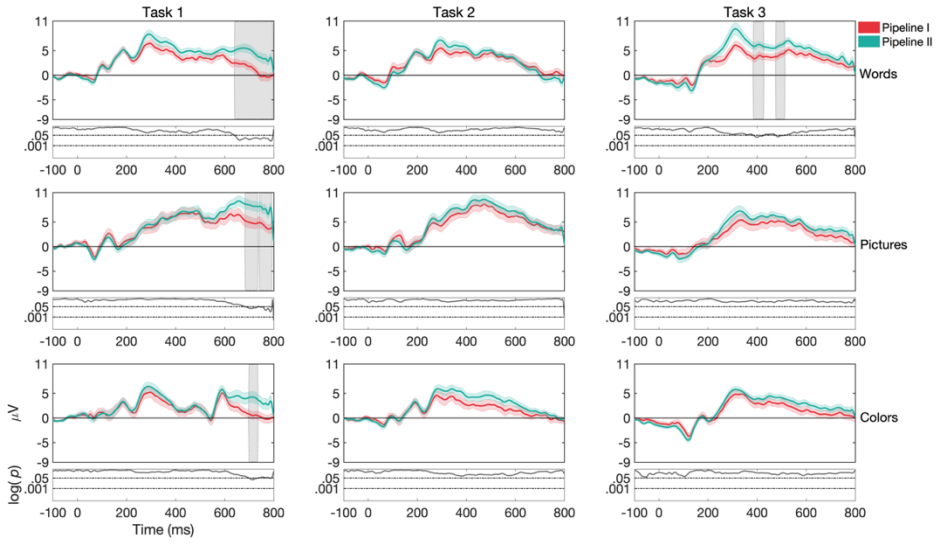

Figure S2. 8. Grand average and SEM of ERP curve at Pz electrode for target trials for different pipelines (Pipeline I vs. Pipeline II) for dataset I (UCR dataset)

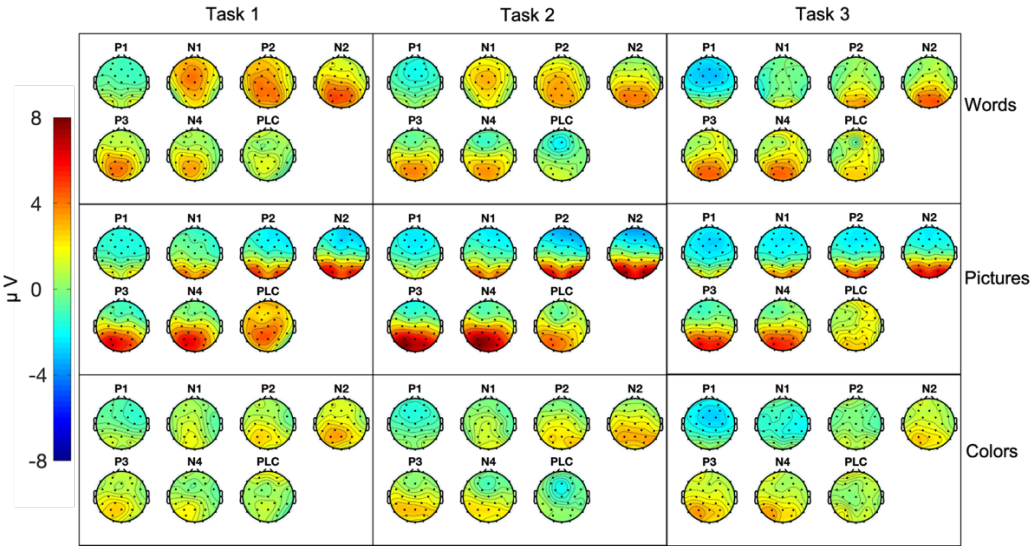

Figure S2. 9. Topographical maps for dataset I (UCR dataset). The time window for each component is as follows: P1 = [0 100], N1 = [100 200], P2 = [150 275], N2 = [200 350], P3 = [250 600], N4 = [400 600], PLC = [500 1000].

Table S2. 6. *p*-values for different components for effect of condition x electrodes using Kruskal Wallis test

		task 1	**task 2**	**task 3**
words vs pictures	**P1**	.471	<.001	.991
	N1	<.001	<.001	<.001
	P2	<.001	<.001	<.001
	N2	<.001	<.001	<.001
	P3	<.001	<.001	<.001
	N4	<.001	.235	.055
	PLC	.126	.998	.999

words vs colors	**P1**	.896	.999	.999
	N1	.053	.698	.008
	P2	.094	.991	.002
	N2	.005	.995	.001
	P3	.214	.999	.024
	N4	.042	.972	.028
	PLC	.822	.991	.018
pictures vs colors	**P1**	<.001	.064	.159
	N1	.007	<.001	<.001
	P2	<.001	<.001	<.001
	N2	<.001	<.001	<.001
	P3	<.001	<.001	<.001
	N4	.003	<.001	<.001
	PLC	.318	.440	.031

Table S2. 7. *p*-values for different components for effect of condition x electrodes using Kruskal Wallis test

		words	**pictures**	**colors**
task 1 vs task 2	**P1**	.998	.249	.979
	N1	.999	.758	.914
	P2	.999	.097	.598

	N2	.982	<.001	.101
	P3	.179	<.001	.316
	N4	.002	<.001	.601
	PLC	.009	<.001	.092
task 1 vs task 3	**P1**	.001	.011	.232
	N1	<.001	.011	.323
	P2	.048	.040	.854
	N2	.785	.263	.999
	P3	.897	.528	.999
	N4	.752	.579	.993
	PLC	.215	.029	.699
task 2 vs task 3	**P1**	.796	.973	.877
	N1	.499	.999	.944
	P2	.638	.999	.144
	N2	.568	.999	.058
	P3	.265	.999	.744
	N4	.610	.987	.942
	PLC	.339	.982	.854

Chapter 3

Modulation of Behavior and Auditory processing by an estimate of arousal activity in a modified auditory oddball task: a multi-measure study of combined behavior, pupillometry and fMRI

Abstract

Even when presented with identical stimuli, our decisions differ from moment to moment. One main source of this perceptual and behavioral variability is an individual's state of arousal –level of alertness and excitability. Arousal is suggested to function on a spectrum: The low extreme consists of sleepiness and inattentiveness while the high extreme consists of panic and stress-induced behavior. Optimal performance can be found in the middle of the spectrum. The locus coeruleus (LC) plays a significant role in cortical arousal via wide, non-specific norepinephrine (NE) projections. However, the interplay between arousal and LC-NE activity in humans has been challenging to investigate due to the lack of direct and robust measures of activity. Here, we examine the potential of an experimental paradigm –a modified auditory oddball task– to measure the variability of perceptual decision-making across multiple levels of observation such as behavioral performance, pupillometry, and neural activity. These rich measures may collectively yield mutual information regarding cortical arousal and LC-NE activity. Our experimental paradigm successfully enabled extractions of stimulus-response functions –the relationship between an auditory stimulus and performance/activity– in all three outcome measures: behavioral measures, pupillometry and neural auditory Blood oxygenation level dependent (BOLD) activity. These findings suggest the viability of our approach in estimating activity of cortical arousal and LC-NE in the future.

Introduction

Our world is an unpredictable place that presents us with a great amount of information and decision-making scenarios, which change from moment to moment. The ability to detect and respond to this information is crucial for survival, especially in uncertain and challenging environments. Interestingly, this process of perceptual decision-making is variable even in the simplest controlled environments. For example, when making decisions for nominally identical stimuli, individuals may vary in response time or behavioral choices from one decision to another (Gold & Shadlen, 2007). One main source of this perceptual and behavioral variability is an individual's state of arousal –level of alertness and excitability. Arousal is thought to affect brain activity and behavior in response to these environmental contingencies and is linked to other functions such as sleep-wake cycle, attention, motivation, anxiety, and stress (Sara & Bouret, 2012; Sara, 2009). Decreased arousal results in inattentive behavior, drowsiness and, in extreme cases, sleep. Increased arousal, which can happen due to a sudden appearance of an environmentally salient stimulus or event, can facilitate behavioral performance. However, it can also result in distractibility and anxiety if extreme (Berridge & Waterhouse, 2003). A classic observation in this regard is the Yerkes-Dodson inverse-U relationship between arousal and performance on various perceptual/ perceptual-motor tasks, which shows that there is a middle ground state of arousal associated with optimal performance in a given task (Yerkes & Dodson, 1908). Recent studies have suggested the locus coeruleus-norepinephrine (LC-NE) neuromodulatory system plays a significant role in regulating

arousal and alertness, and by doing so optimizes performance according to task and environment requirements (Aston-Jones & Cohen, 2005).

LC is a cluster of neurons in the pons of the brainstem that serves as the main source of cortical NE in the brain. The LC-NE plays a critical role in many key cognitive functions including perception, attention, memory and learning besides arousal (Aston- Jones & Cohen, 2005; Sara, 2009). LC neurons fire in two distinct modes: tonic –spontaneous– and phasic –task-relevant– which have different patterns of NE release and behavioral manifestation. Phasic LC activity is a brief, high-frequency (10-20 Hz) burst of action potentials elicited by salient or unexpected stimuli, attentiveness, and response-related signals (Devilbiss & Waterhouse, 2011). The release of NE in this mode enhances stimulus processing by increasing neuronal responsivity (gain) in task-related regions (Aston-Jones & Cohen 2005; Foote, Aston-Jones & Bloom, 1980; Berridge & Waterhouse, 2003). Tonic LC activity consists of low-frequency baseline fluctuations (.1-5 Hz) related to arousal levels and results in impaired attentional performance (Vazey et al., 2018). These two modes of activity are not mutually exclusive, and the balance between phasic and tonic LC activity produces an optimal level of performance (Jepma & Nieuwenhuis, 2010) in a manner that mirrors the inverted-U relationship between the arousal/tonic-LC and task engagement/phasic-LC (Aston-Jones & Cohen, 2005). Aston-Jones and Cohen (2005) developed the Adaptive Gain Theory to explain the role of the LC-NE system in behavioral and neural levels: the phasic mode of LC activity is driven by task-related decision processes, and it provides the target neurons with a transient gain that facilitates task-

related behavior (exploitation). On the other hand, the tonic mode of LC activity produces a more enduring and less discriminative gain increase in the target neurons. Although this is not beneficial for the task at hand, it will facilitate the disengagement of performance from this task and look for other valuable opportunities (exploration).

Much of the information on the role of LC-NE system on cognitive processes and behavior comes from pharmacological studies and intracranial recordings on animal models as well as computational models (Foote et al., 1975; Aston-Jones, Rajkowski, Kubiak, & Alexinsky, 1994; Manunta & Edeline, 2004; Edelin, Manunta, & Hennevin, 2011; Servan-Schreiber, Printz, & Cohen, 1990; Usher, Cohen, Servan-Schreiber, Rajkowski, & Aston-Jones, 1999). Advances in the field of neuroimaging have yielded new opportunities to characterize LC-NE function in humans using markers such as the Blood-Oxygen-Level-Dependent (BOLD) signal of functional magnetic resonance imaging (fMRI). However, this effort is restricted as the LC's small size and location near the fourth ventricle makes the extracted BOLD signal particularly sensitive to physiological noise such as respiration and heart pulses (Szabadi, 2013). Another marker for LC-NE activity and cortical arousal state is the non-luminance mediated pupil diameter which has shown strong correlations with LC activity directly recorded in animals (Rajkowski, Kubiak & Aston-Jones, 1993; McGinley et al., 2015) and links with LC-BOLD in humans (Murphy et al., 2014; Alnæs, et al., 2014; deGee et al., 2017). Particularly, pupillometry seems to reflect both the tonic and phasic aspect of the LC-NE function. In target detection experiments in monkeys, pre-trial baseline pupils had large

values during tonic LC activity and relatively small values during phasic LC activity (Rajkowski et al., 1993). Moreover, there is increasing evidence that rapid large pupil dilation tracks task-related activity and task processing (Richer & Beatty, 1987; Einhäuser et al., 2008). Together, these studies suggest that similar to the reciprocal relationship between LC tonic and LC phasic, baseline pupil diameter and task evoked pupil diameter change should be inversely correlated.

In the oddball target detection task, subjects are presented with sequences of repetitive stimuli that are infrequently interrupted by a deviant or "odd" stimulus. Detection of the odd stimulus reliably evokes transient activity which was initially measured by event-related potential (ERP) research (Squires et al., 1975; Huettel & McCarthy, 2004). Using an auditory oddball task, Gilzenrat et al. (2010) calculated the one-second pre-stimulus pupil diameter as the baseline (tonic) pupil measure as well as the highest deviation from the baseline in the 2.5 s following the tone onset as the task-relevant (phasic) pupil measure. The authors showed that the increase in the baseline pupil was associated with degraded task performance (indexed by reaction time and phasic pupil diameter), whereas reduced baseline diameter (but increased task-evoked pupil dilations) was associated with improved task performance. Murphy et al. (2011) used an auditory oddball task and grouped measures of reaction time (reaction time and reaction time coefficient of variation) based on baseline pupil diameter quantiles. In contrast to Gilzenrat et al. (2010) they failed to observe any significant difference in reaction time based on baseline pupil diameter. However, similar to Gilzenrat (2010), the authors showed an inverse relationship

between tonic and phasic pupil diameters. Together, these studies suggested that despite the challenges in accessing LC-NE activity and lack of direct recordings in humans, neurophysiological markers alongside task engagement measures (indexed by behavioral accuracy and reaction time) can serve as estimates of cortical arousal function. However, this approach requires designing an experimental paradigm that is sensitive enough to characterize how LC engagement may change stimulus encoding, processing and behavior.

One way to evaluate this is to extract stimulus-response functions across multiple measures of a perceptual decision-making task such as behavioral performance, pupillometry and fMRI BOLD. In principle, we can fully characterize a system's (individual's) stimulus-response function by probing the system with all the possible inputs or stimuli and measuring all the corresponding outputs or responses. As in practice this is not possible, one approach is to probe the system with a proper subset of stimuli, record the responses, and quantify this stimulus-response function using mathematical tools. Then by using this function, a feature of a stimulus (e.g., luminance for visual stimulus or frequency for auditory stimulus) can be mapped onto a response to generate a psychometric function. If the measured response is the pupillometry dilation response or the fMRI-BOLD change of a particular brain region, the functions are called pupillometric and neurometric respectively. Estimating and understanding the stimulus-response functions in various stages of perceptual processing is important for two reasons: 1) It give us a quantifiable method to describe behavioral performance and neurophysiology and predict these values even for stimuli that were not used in the estimation. 2) We can compare the stimulus-

response functions to each other and explain the sensory-driven (i.e., A1-BOLD) or arousal-driven (pupillometry) aspects of behavior. For example, by comparing the neurometric function to psychometric function, we can understand whether the neural voxels of the sensory region (i.e., A1) constrain perceptual accuracy.

In the current study, we took the first step towards developing a tool to properly study LC-NE activity, with the long-term goal of investigating the relationship between behavioral performance, pupillometry, A1-BOLD and LC-BOLD. This tool was tested during salient events in the context of a modified auditory oddball task. The oddball paradigm is a popular paradigm commonly used to investigate cognitive processes in event-related studies (Rajkowski, Kubiak, & Aston-Jones 1994; Huettel & McCarthy, 2004; Stevens et al., 2000; Linden et al. 1999). This paradigm requires detecting the odd stimulus embedded in a series of frequent stimuli. Notably, the auditory oddball paradigm detection phase is associated with physiological indices such as pupil diameter, EEG P300 of cognitive processing (Nieuwenhuis et al., 2005) and neural activity in primary auditory cortex (Chen et al., 2015; Walz et al., 2015). In this study, we systematically manipulated the stimulus novelty of the oddball task using a range of frequent offsets while keeping the probability of oddball to frequent stimuli 2:8. We extracted the stimulus-response functions for behavioral performance, pupillometry and neural response. We hypothesized that the used parameters for oddball offsets were sensitive to capture variance and changes in LC-NE levels resulting in quantifiable functions rather than a single-point measure. Further, we tested the robustness of our experimental paradigm and extracted stimulus-response

functions by applying a physical squeeze ball stressor (Mather et al., 2020; Nielsen & Mather, 2015) to see whether the squeeze factor alters the stimulus-response functions by enhancing or disrupting the response to the attended tone. Mather and colleagues (2020) used an isometric squeeze ball and showed that squeezing the ball for a few seconds upregulates the tonic LC and arousal. Therefore, we might expect that the squeeze stressor modulates any of the extracted stimulus-response functions. Finally, we provided insight and recommendations regarding how to further enhance our experimental paradigm, neuroimaging data acquisition and analyses to benefit future LC studies.

Methods

Participants

Thirty healthy students from the University of California, Riverside (UCR) participated in this study [M_{age} = 24.5, SD_{age} = 4.4 years; 17 females]. All the participants were right-handed except for one who was ambidextrous. One additional student participated but their data were excluded from all analyses due to a diagnosis of attention deficit hyperactivity disorder (ADHD), which is an exclusionary factor for this study. Exclusion criteria included being left-handed, a history of cognitive disorder or impairment, use of psychoactive medication or failure to pass the MRI screening interview which conducted via email. The experimental protocol was approved by the Institutional Review Board of UCR, and all participants gave their informed consent before participating in the experiment. The data was collected at UCR's Center for Advanced Neuroimaging (CAN).

Behavioral Protocols

Each participant underwent a resting-state fMRI scan. Then they engaged in either squeezing the stress ball (experimental condition) or holding the stress ball (control condition) and rested after each condition. Next, participants performed three blocks of the auditory oddball task followed by a short version of the squeeze/control-rest sequence. Finally, they performed three more blocks of the auditory oddball task. Figure 3. 1 shows the structure of each of these block types. During rest, participants were instructed to relax, think of nothing, and maintain fixation for 5 min at a centrally presented bullseye. Prior to the start of the squeeze/control-rest sequence block, participants were given an isometric ball of normal stiffness and they either held the ball or squeezed it as hard as possible with the dominant hand. We used a within-subject design meaning that each participant engaged in both the squeeze and hold (control) condition. We counterbalanced the order by randomly assigning participants to either perform the squeeze first or control first. The original squeeze/control-rest block had 5 cycles of squeeze/control alternating with rest. While the duration for squeeze/control was always 18 s, the duration for rest was as follows: 2 min, 2 min, 5 min, 1 min, 1 min. For the short version, we only had 2 cycles of squeeze/control-rest with the resting duration as follows: 1 min, 1 min. The original block took 12.5 min, and the shortened block took 2.6 min making the total time of 15.1 min.

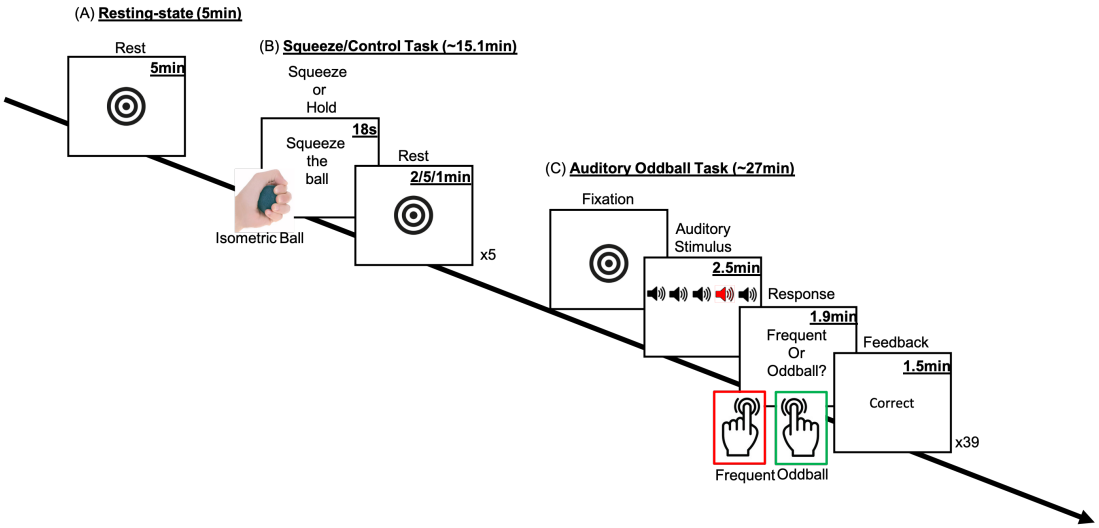

Figure 3. 1. Experimental design showing the (A) Resting-state task during which participants fixated on the bullseyes (B) squeeze/control- rest sequence: The participants performed this twice, once before the oddball task and a shortened version of it after the three blocks of oddball. (C) Oddball task

During the oddball blocks, subjects listened to a series of tones that were either frequent or oddball. A frequent trial contained a sequence of five consecutive tones of the same frequency (1000 Hz), while an oddball trial consisted of five consecutive tones with one odd, embedded tone (1004, 1008, 1016, 1032, 1064, or 1128 Hz). The oddball tone was in either the 2nd, 3rd or 4th tone position. Duration of the auditory stimulus was 2.5 min and duration of each tone was 0.1 min with ISI of 0.4 min. Subjects indicated their choice on each trial ("oddball or frequent?") using two 2-button MRI-compatible response boxes, namely responsepixx (VPixx technologies, Vision Science Solutions, Quebec, Canada). The response window was 1.9 min. After the response window, participants were given visual feedback on the screen in forms of "correct" or "wrong" text. In addition to frequent and oddball trials, we had trials of no tone –blank– to serve as a baseline for % signal change. During blank trials, participants were shown instructions on the screen for which

buttons to press. Each trial began with a fixation point, followed by 2.5 min of auditory stimulus presentation through the MRI-compatible headphones, 1.9 min of response and 1.5 min of feedback. We added pseudorandom jitter of 1 TR at the end of each trial. Each block consisted of 39 trials with 4 trials per oddball level (24 oddball trials in total), 9 frequent trials and 6 blank trials and lasted around 4.5 mins. We had a total of six runs that brought our oddball task time to 27 min. The experiment was generated using MATLAB 2015b, Psychophysics Toolbox, version 3 (Brainard, 1997; Pelli, 1997).

Magnetic Resonance Imaging and Preprocessing

Imaging data were acquired on a 3-Tesla Siemens Prisma MRI scanner (Prisma, Siemens Healthineers, Malvern, PA) equipped with a 64-channel head-coil at the Center for Advanced Neuroimaging at University of California, Riverside. Participants laid supine in the scanner with the head stabilized with foam pads to minimize head movements. For anatomic T1 image: TR/TE/TI=2400/2.72/1060 ms, flip angle = 8 degrees, FOV = 256 × 240 mm^2, voxel size = 0.8 × 0.8 × 0.8 mm3, 208 slices, GRAPPA = 2. For resting-state fMRI and task fMRI: TR/TE = 2000/32 ms, FOV = 224 × 196 mm^2, matrix size = 112 × 98, slice thickness 3 mm, 52 slices, flip angle = 69 degrees, multiband factor = 2, GRAPPA = 2, bandwidth = 1440 Hz/Px, phase encoding direction was AP. Susceptibility distortion correction used a spin echo EPI with spatial parameters matching those of the fMRI. The only differences were as follows: 1) flip angle was 90 degrees for the SE EPI, 2) no multiband or GRAPPA was used in either of AP or PA acquisition.

Preprocessing of the images was performed using Freesurfer (Fischl et al., 2002; http://surfer.nmr.mgh.harvard.edu) and FMRIB Software Library (FSL; Woolrich et al., 2009; https://fsl.fmrib.ox.ac.uk/fsl/fslwiki/). Automated gray matter parcellation was completed using the FreeSurfer software (recon-all). This step was used both for defining our region of interest (ROI) as well as to skull-strip the anatomical images. All individual T1 images were registered to the. A structural auditory cortex ROI for each subject was then constructed combining the G_temp_sup-G_T_transv, G_temp_sup-Lateral, and S_temporal_transverse FreeSurfer labels from Desikan-Killiany Atlas across left and right hemispheres. The structural A1 mask as well as the structural T1 warped into the standard Montreal neurological institute (MNI) space using flirt. Further we applied slice-time correction using a MATLAB (Mathworks, Natick, MA) code. Then we performed motion-correction (mcflirt) and susceptibility distortion correction (topup). Finally using the FEAT tool we applied the temporal filter (high pass, cut-off = 100s) and spatial smoothing with Gaussian kernel (full-width half-maximum; FWHM= 5mm). We applied two general linear models (GLM) to define two functional masks: 1) We applied a GLM model with onset of sound (aggregate of oddballs and frequent trials) and onset of blank trials as two regressors. The purpose of this mask was to locate sound-responsive voxels; 2) We applied a GLM model with onset of oddballs (aggregate of all oddball trials) and onset of frequent trials as two regressors. The purpose of this mask was to locate oddball-responsive voxels. The final hybrid mask was the overlap of the two functional masks within the structural ROI of primary auditory cortex (A1). Figure 3. 2 presents a schematic of applied preprocessing to

preprocess the EPI images and create final hybrid mask that we later used to extract A1-BOLD timeseries. For the LC localization we used the mask described previously elsewhere (Langley, et al., 2020).

Pupillometry and Processing

Pupil diameter was recorded continuously using a binocular eye-tracker (TRACKPixx; VPixx Technologies) with the sampling-rate of 2000 Hz. The recorded data had a column that showed blinks. For preprocessing, we used ET-remove semi-automated program (github.com/EmotionCognitionLab/ET-remove-artifacts) to interpolate the blinks. Then a trained user visually inspected the data to either correct or remove the improperly interpolated segments. For the oddball pupillometry data, we segmented the data based on the event types with -1 second before the sound start and +5s after the sound start. Further, we calculated the trial-by-trial baseline and change in pupil diameter to represent the tonic and phasic pupillary activity, respectively. We defined the tonic (baseline) pupil diameter as the average before 1-s period prior to tone onset. We defined the phasic pupil activity induced by the tone presentation in each trial by calculating the difference between the highest divergence of pupil diameter from the baseline within the 2.5 s (Glizenrat et al., 2010).

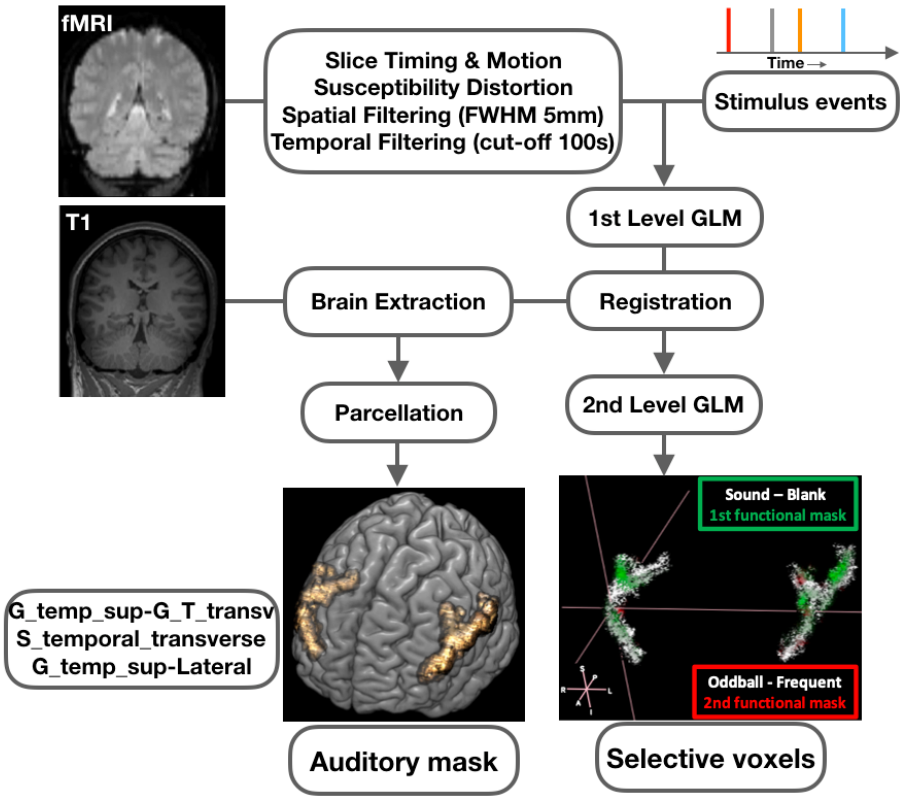

Figure 3. 2. Pipeline for fMRI data pre-processing and fitting an iterative GLM model using the corresponding stimulus events. The hybrid mask is a result of structural [Primary Auditory cortex] x functional mask1 x functional mask2. The functional mask is a result of [sound-blank] x [oddball-frequent]. The final hybrid mask was binarized and generated for each individual subject.

Measures

Our study utilized multi-measure experimentation in which we collected behavioral measures as well as pupillometry and neural data as follows:

Reaction Time. For each trial, reaction time was measured as the time between the response cue ("Is the sound oddball or frequent) and the participant's button press providing an estimate of the oddball detection decision time. As the response window was open only after the offset of the sound stimulus, the estimated reaction time was a delayed

response time. This is thought to vary both by oddball and oddball saliency and may be impacted by the arousal or LC activity where, similar to the relationship portrayed by Yerkes-Dodson curve, slower responses are in the left, fast in the middle and more variables towards the right. Grand-average psychometric stimulus-response function was acquired by averaging the reaction time values across trials, sessions and across participants.

Accuracy. For each trial, accuracy was considered as the binary outcome of whether a trial was "Frequent" or "Oddball" where correct decisions were recorded as 1 and incorrect decisions were recorded as 0. This provides an estimate of decision correctness. This is thought to vary both by oddball and oddball saliency and may be impacted by the LC where more accurate responses happen in the middle of the Yerkes-Dodson curve. Grand-average psychometric stimulus-response function was acquired by averaging the accuracy values across trials, sessions and across participants.

Pupillometry response. For each trial, we averaged the pupil diameter for left and right eyes and for vertical and horizontal axes to increase the quality of the signal and then examined the pupil segments by locking the signal to oddball onset from .1s prior to 5 seconds after. We considered the baseline pupil diameter during .1s as tonic pupil activity and baseline-corrected peak pupil between 0.75 and 1s as phasic pupil activity. Phasic pupil response is thought to covary with phasic LC-NE activity and tonic pupil response is thought to covary with tonic LC-NE activity (Gilzenrat et al., 2010). Grand-average

pupillometric stimulus-response function was acquired by averaging the phasic pupil dilation values across trials, sessions and across participants.

A1-BOLD. For each trial, we extracted the average BOLD signal from hybrid A1 mask during 1TR prior to the sound start time and 9TRs after. We calculated the BOLD signal 1TR prior to sound start as tonic A1 activity and baseline-corrected A1-BOLD at 2 and 3 TR as the phasic activity. Phasic A1-BOLD is considered as the transient change in auditory activity from a baseline activity in response to a sound stimulus. Tonic A1-BOLD is considered as the sustained auditory activity. A1-BOLD may be modulated by LC activity and cortical arousal. Grand-average neurometric stimulus-response function was acquired by averaging the phasic A1-BOLD values across trials and across participants.

LC-BOLD. For each trial, we extracted the average BOLD signal from LC mask (Langley et al., 2020) during 1TR prior to the sound start time and 9TRs after. We calculated the BOLD signal 1TR prior to sound start as tonic LC activity and baseline-corrected peak LC-BOLD at 5 TR after sound start as the phasic activity. Phasic LC-BOLD is thought as the proxy of transient change in LC-NE activity from a baseline activity whereas tonic LC-BOLD is thought as the sustained LC-NE activity. LC-BOLD may be modulated by perceptual stimulus (i.e., saliency of the stimulus) or cognitive processing (i.e., performance monitoring and error related activity). LC-BOLD is extremely susceptible to noise due to its small size and location. Grand-average neurometric stimulus-response function was acquired by averaging the phasic LC-BOLD values across trials and across participants for 10-second time point.

Statistical Analysis

Before running any analyses, we have conducted data cleaning in multiple steps. In step A of the data cleaning, we excluded 3 participants: 1 participant excluded due to later diagnosis of ADHD, 2 participants excluded as their hybrid A1 mask did not result in any active voxels. In step B we excluded the trials and all their corresponding measures where the subject failed to register a response. In step C we excluded the trials and all their corresponding measures where the participants responses were faster than 100ms. In step D, we excluded the trials that had missing values in oddball-locked pupillometry segments. In step E we excluded the trials with missing value in BOLD signal segments. In step F we excluded the trials with no sound (blank). Figure 3. 3 illustrates all the applied trial rejection steps and the final trial counts.

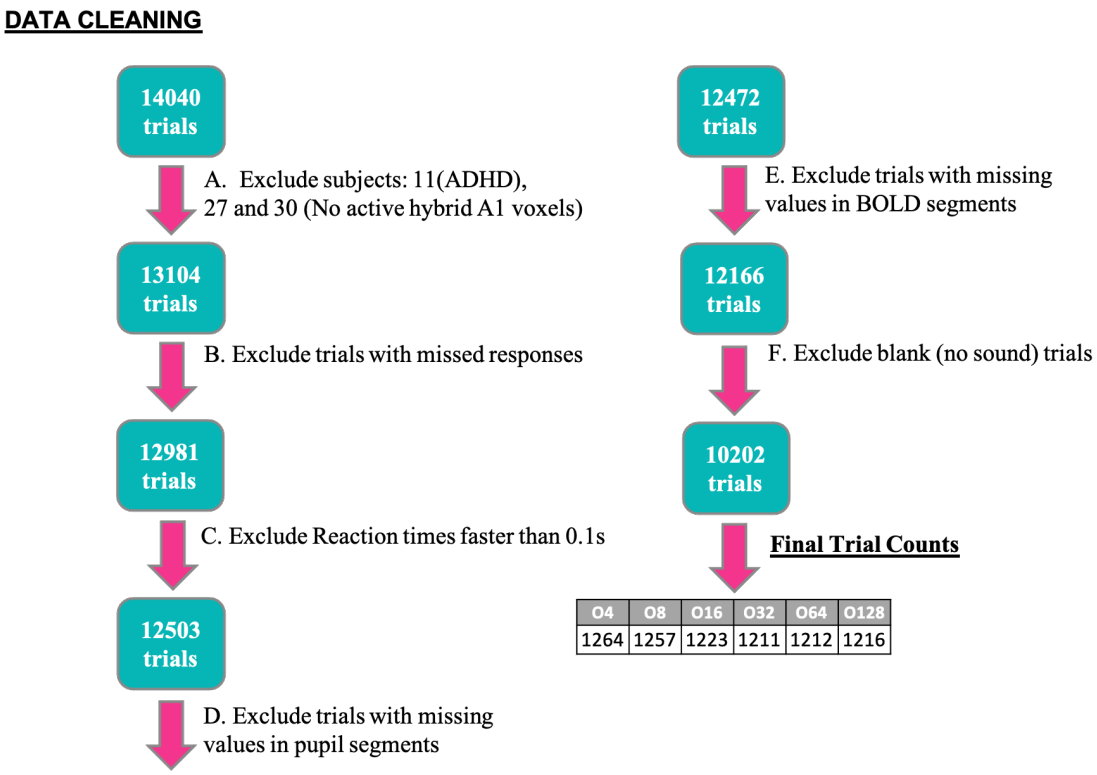

Figure 3. 3. Data cleaning. Before doing any six steps of A to F were implemented.Final trial counts showed no significant difference between the trial numbers of Oddball 1004 (O4), Oddball 1008 (O8), Oddball 1016 (O16), Oddball 1032 (O32), Oddball 1064 (O64), Oddball 1128 (O128).

All the analysis were performed after the data cleaning procedure. As a first step we examined the extent to which the paradigm gave rise to responses in pupil, primary auditory cortex, and LC. Second, we assessed the effect of handgrip manipulation by applying non-parametric permutation tests (Cohen, 2014) to behavioral as well as neurophysiological data. Third, we extracted stimulus-response functions: relating the auditory stimuli, different oddball levels in this case, to the evoked behavioral response, pupillometry response and neural brain responses. For psychometric function, performance at each oddball level quantified as average of percent correct responses across subjects and across sessions. For reaction time trend, we followed a similar analysis on reaction times.

For pupillometric function, pupillometry response at each oddball level was quantified as the average of pupillometry dilation responses between .75-1 s across subjects and across sessions. For neurometric function in primary auditory cortex, A1-BOLD response at each oddball level quantified as average of A1-BOLD between 2-3 seconds across subjects and across sessions. For neurometric function in LC, the stimulus-locked grand-average signal did not show a clear peak and latency, however, we have considered the average data point at 10 seconds at each oddball level. We used repeated measure analysis of variance (ANOVA) with seven levels (Frequent, Oddball 1004, O Oddball 1008, Oddball 1016, Oddball 1032, Oddball 1064, and Oddball 1128). See the supplementary material (Figures S3. 1, S3. 2, S3. 3 and S3. 4) for stimulus-response functions of control and squeeze sessions. For the ANOVA tests, if the sphericity test was violated, we used Greenhouse-Geisser-corrected degrees of freedom based on significant Mauchley's Test of Sphericity, $p < .05$. Finally, we compared the extracted stimulus-response functions across subjects. To do this we normalized the reaction time, pupillary and fMRI-BOLD data between 0 and 1. We characterized the stimulus-response functions, by fitting a linear function (α. x + b) and estimating the slope (α). The linear fits can account for particularly poor fits of more complex models such as Weibull and logistic types. We performed correlations on the fitted slopes to determine the relationship between different stimulus-response functions across subjects.

Results

As a reminder, all the analysis and depicted results in this section were performed after the data cleaning steps (see Methods).

Quantifying Task-Evoked Neurophysiological Responses

As a first step, we examine the extent to which the auditory oddball paradigm gave rise to responses in pupil diameter, BOLD signal of the primary auditory cortex, and LC BOLD activity.

Pupillometry Data. To track phasic arousal, we measured the pupil response, a proxy of LC-NE activity, immediately after the oddball onset. We obtained the pupil time course for each frequency level by averaging across the trials of that frequency level (see Methods). Oddball tones evoked greater dilatory responses in the pupil diameter around .75-1 s after initiation of the tone for both control and squeeze sessions (Figure 3. 4-A, 3. 4-B). As seen in the graph the higher frequencies (saliency) showed larger changes than the lower frequencies, with the frequent as well as oddball tones at 4 and 8 Hz showing a different pattern than the other tones. There is a second peak around 3.5-4.5 s which according to its latency is likely related to the response (button press) or processing feedback (correct or wrong).

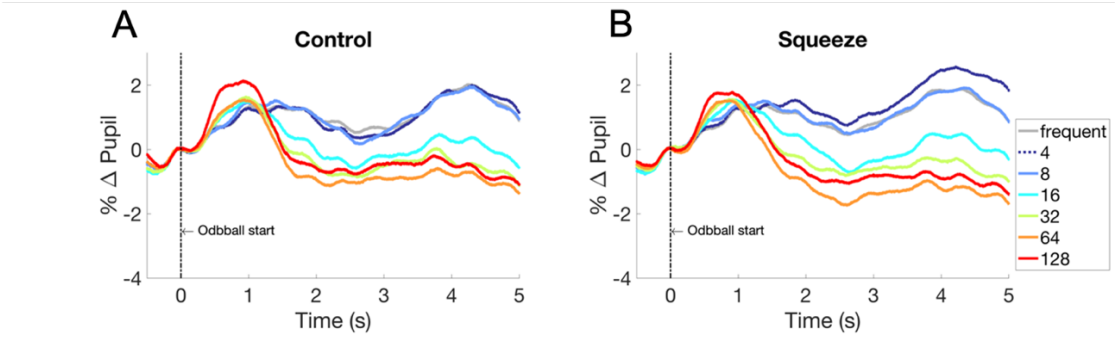

Figure 3. 4. Grand-average pupillary dilation response. (A) Control session (B) Squeeze session

3.1.2. Neural Data (fMRI-BOLD)

Here, we investigated the effect of stimulus saliency on BOLD activity of the primary auditory cortex (A1), a sensory region activated during auditory perception, and LC, a region associated with cortical arousal. Similar to the pupillometry data, oddball tones evoked substantial BOLD change responses in 4-6 s (2-3 TR) in auditory cortex voxels for both control and squeeze sessions based on the oddball saliency (Figure 3. 5-A, 5-B). After the first peak, the BOLD signal dropped at ~10 s, followed by a small second peak. According to latency, the first peak was associated with stimulus processing and the second peak was due to response (button press) or processing feedback. We considered the -1TR as the baseline (in arbitrary units). Visual inspection of the A1-BOLD signals indicated that despite the baseline correction, some variability between frequencies remained during baseline period of -1TR. This remaining noise maybe due to pre-processing steps and scanner noise.

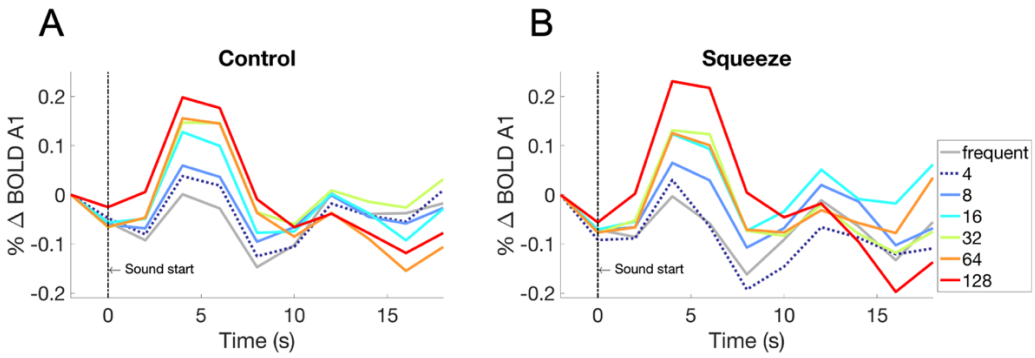

Figure 3. 5. A1-BOLD epochs locked to sound start time. (A) During Control session (B) During Squeeze session

Stimulus-locked BOLD responses of LC did not result in a clear peak for oddball tones; however, the BOLD time course at 10 s seemed to harbor a synchronous activity manifested as local peaks in the control condition (Figure 3. 6-A). Similar to A1-BOLD, we considered the -1TR as the baseline (in arbitrary units). Visual inspection of the LC-BOLD after baseline correction indicated that despite the baseline correction, substantial variability remained at the sound start time especially for the squeeze session (Figure 3. 6-B). This observation indicated the existence of substantial noise in the LC-BOLD measure which can be due to scanner noise, physiological noise such as cardiac cycles, or pre-processing noise such as smoothing (Turker et al., 2021).

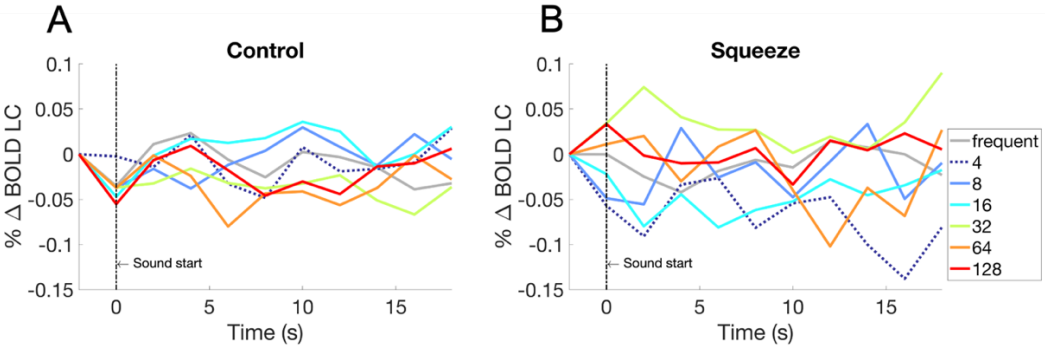

Figure 3. 6. LC-BOLD epochs locked to sound start time. (A) During Control session (B) During Squeeze session

Effect of Hand Grip Manipulation

We applied a non-parametric permutation test (Cohen, 2014) to behavioral as well as neurophysiological data to assess the effect of handgrip manipulation. Permutation-based non-parametric tests are based on fewer assumptions, known as distribution-free and they are amenable to multiple comparisons. Therefore, these tests are suitable specially for neurophysiological data. For signals, we summarized the time course of the signal as follows: a peak at an average of .75-1s for pupil dilation, a peak at an average of 4-6 s in A1-BOLD, and no peak for but the activity in the LC BOLD time course at 10 s.

Table 1 shows the *p*-values. Unlike Mather et al. (2020) who found that squeezing an isometric stress ball resulted in faster oddball detection speed and increased phasic arousal (estimated by pupil responses) to oddballs, we did not find any statistically significant difference between control and squeeze sessions.

Table 3. 1. Results of non-parametric permutation test to compare the behavioral and neurophysiological values of control session to squeeze

TRIAL TYPE	**ACCURACY MEAN *(SD)***		***P*-VALUE**	**REACTION TIME MEAN (SD)**		***P*-VALUE**
	Control	**Squeeze**		**Control**	**Squeeze**	
Frequent	74.84 (16.28)	75.40 (15.05)	.454	495.54 (136.66)	453.61 (126.48)	.885
O4	33.58 (15.09)	33.05 (14.83)	.550	475.32 (143.04)	495.24 (162.06)	.311

O8	56.59 (22.05)	58.53 (20.87)	.371	473.22 (158.87)	437.45 (139.46)	.820
O16	84.93 (18.19)	84.77 (15.57)	.523	400.05 (123.89)	384.91 (131.26)	.679
O32	96.53 (6.11)	94.70 (7.64)	.828	382.22 (121.37)	345.33 (108.41)	.885
O64	97.71 (4.51)	99.31 (2.09)	.053	370.37 (112.56)	342.83 (86.92)	.840
O128	99.51 (1.93)	99.09 (2.49)	.769	346.17 (82.29)	333.43 (68.35)	.728

TRIAL TYPE	**PUPIL MEAN (SD)**		***P*-VALUE**
	Control	**Squeeze**	
Frequent	1.19 (1.01)	1.13 (1.12)	.578
O4	1.17 (1.45)	1.14 (1.51)	.518
O8	1.37 (1.49)	1.25 (1.56)	.613
O16	1.49 (1.41)	1.54 (1.74)	.456
O32	1.64 (1.78)	1.64 (1.53)	.502
O64	1.56 (1.54)	1.62 (1.92)	.448
O128	2.21 (1.84)	1.86 (1.46)	.772

TRIAL TYPE	A1-BOLD MEAN (SD)		*P*-VALUE	LC-BOLD MEAN (SD)		*P*-VALUE
	Control	**Squeeze**		**Control**	**Squeeze**	
Frequent	-0.01 (0.09)	-0.02 (0.05)	.805	0.01 (0.09)	-0.01 (0.11)	.750
O4	0.03 (0.13)	-0.01 (0.12)	.910	0.01 (0.13)	-0.04 (0.12)	.956
O8	0.04 (0.07)	0.04 (0.11)	.509	0.03 (0.18)	-0.03 (0.16)	.943
O16	0.11 (0.14)	0.11 (0.11)	.520	0.03 (0.16)	-0.04 (0.14)	.978
O32	0.14 (0.12)	0.13 (0.10)	.660	-0.02 (0.16)	0.00 (0.17)	.277
O64	0.15 (0.14)	0.11 (0.13)	.839	-0.03 (0.13)	-0.03 (0.16)	.458
O128	0.19 (0.10)	0.22 (0.25)	.264	-0.02 (0.15)	-0.02 (0.16)	.533

As there were no statistically significant differences between control and squeeze session in our measures of interest, we aggregated these two sessions for the remaining results.

Stimulus-Response Functions

Here, we investigated the relationship between stimulus levels (frequency levels: frequent: 1000 Hz, oddballs: 1004, 1008, 1016, 1032, 1064 and 1128 Hz) and behavioral, pupillometry and fMRI-BOLD responses.

Psychometric Function. We first determined how participants' behavioral responses (i.e., percentage of correct responses, reaction time in ms) acted as a function of stimulus frequency during the oddball task**.** This was done by averaging the behavioral responses for each stimulus frequency level (see Methods). Our goal was to determine the effect of stimulus saliency on behavioral performance. Figures 3. 7-A and 3. 7-B show

stimulus-response functions for accuracy and reaction time. Error bars represent standard error of the mean (SEM) across trials.

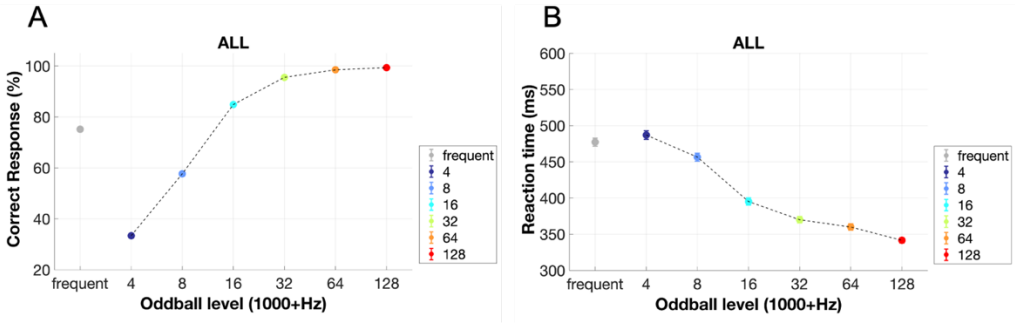

Figure 3. 7. Relationship between behavioral responses and auditory stimuli in oddball detection task. (A)Psychometric function and (B) reaction time trend. Accuracy increased and reaction time decreased as a function of oddball saliency.

There was a main significant effect of stimulus frequency level on accuracy ($F(2.34,63.22) = 136.46$, $p < .001$, $\eta^2 = .83$) and reaction time ($F(2.60,70.44) = 33.24$, $p < .001$, $\eta^2 = .55$). As expected, these analyses suggested that accuracy increased and reaction time decreased with increasing oddball frequency. These results indicated that increasing the oddball saliency led to higher behavioral performance in the form of more correct responses and faster responses. Our results demonstrated that we have utilized the appropriate range of stimulus frequency to capture variability in the behavioral performance for both reaction accuracy and reaction time.

Pupillometric Function. Having tested the behavioral responses as a function of the stimulus frequency, we proceeded to investigate the effect of stimulus frequency on pupil dilation response while participants performed the auditory oddball task. This was done by averaging the pupil diameter dilation responses for each stimulus frequency level (see Methods). Our goal was to determine the effect of stimulus saliency on pupil dilation

response. Figure 3. 8 shows pupillometric function obtained by plotting the pupillometry dilation response as a function of auditory stimulus frequency level. Error bars represent standard error of the mean (SEM) across trials.

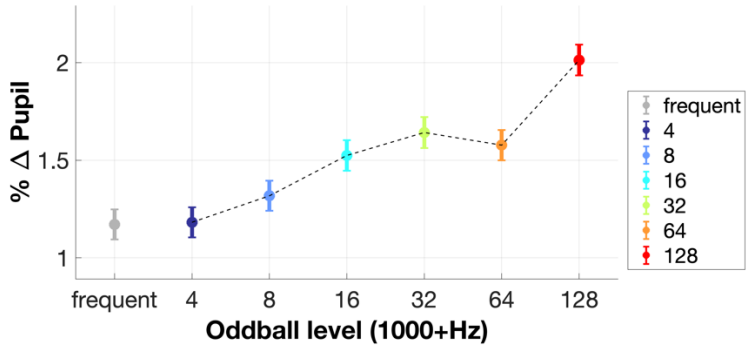

Figure 3. 8. Relationship between pupillometry dilation response and auditory stimuli in oddball detection task. Pupillometry dilation response increased and reaction time decreased as a function of oddball saliency.

There was a main significant effect of stimulus frequency level on pupillometry dilation response ($F(3.67,99.13) = 5.42$, $p < .001$, $\eta^2 = .16$). This result suggested that pupillometry dilation responses increased with the increasing oddball saliency. Since phasic pupil dilation is a proxy of phasic arousal activity, the resulting pupillometric functions indicated that the oddball paradigm was able to capture the variability in phasic arousal.

Neurometric Functions. Next, we investigated how the effect of stimulus saliency was reflected on the neural level. In the context of our perceptual decision-making task, A1 and LC were chosen as the task-relevant regions. Figure 3. 9 shows neurometric function obtained by plotting the average BOLD response within the hybrid A1 region (A1

mask) as a function of auditory stimulus frequency level. Error bars represent standard error of the mean (SEM) across trials.

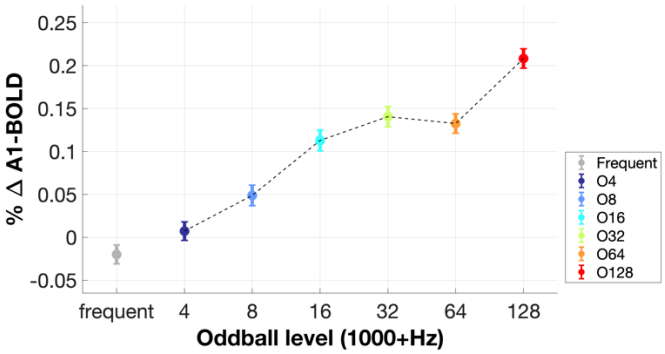

Figure 3. 9. Relationship between A1-BOLD and auditory stimuli in oddball detection task. A1-BOLD increased as a function of oddball saliency.

There was a main significant effect of stimulus frequency level on A1-BOLD % change ($F(3.45,93.27) = 20.32$, $p < .001$, $\eta^2 = .42$). Unlike the neurometric function for A1-BOLD which showed a monotonic signal increase as a function of oddball saliency, LC-BOLD failed to demonstrate any significant effect of oddball saliency ($F(6,162) = 0.40$, $p = .876$). The neurometric function for LC is shown in Figure 3.10. However, caution must be taken when interpreting the LC-BOLD signal change as there were many potential noise factors that may distort the extracted signal (Tucker et al., 2021). Examples of noise factors including but not limited to image noise, scanner noise, motion, cardiac pulsation and the respiratory cycle (Keilholz et al., 2017).

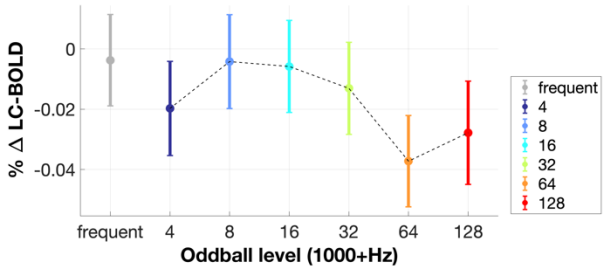

Figure 3. 10. Relationship between LC-BOLD and auditory stimuli in the oddball detection task.

Taken together, this analysis established the relationship between neuronal responses and the stimulus frequency. While auditory processing response (estimated using A1-BOLD) increased with the level of oddball saliency, we did not capture any significant variability in LC-BOLD response.

Comparisons of Stimulus-Response Functions Across Measures. The extracted stimulus-response functions (except the neurometric function of LC) indicated that stimulus frequency/saliency was represented in measures of behavioral performance, pupillometry, and fMRI-BOLD. That is: all the responses increased as a function of stimulus saliency. Therefore, we compared the different stimulus-response functions to each other to decide whether the neurophysiological signals carry information that might be associated with psychophysical behavior. To do this, we characterized the stimulus-response functions for each participant with fitting a linear model (see Figure 3.11) across different measures of behavior,

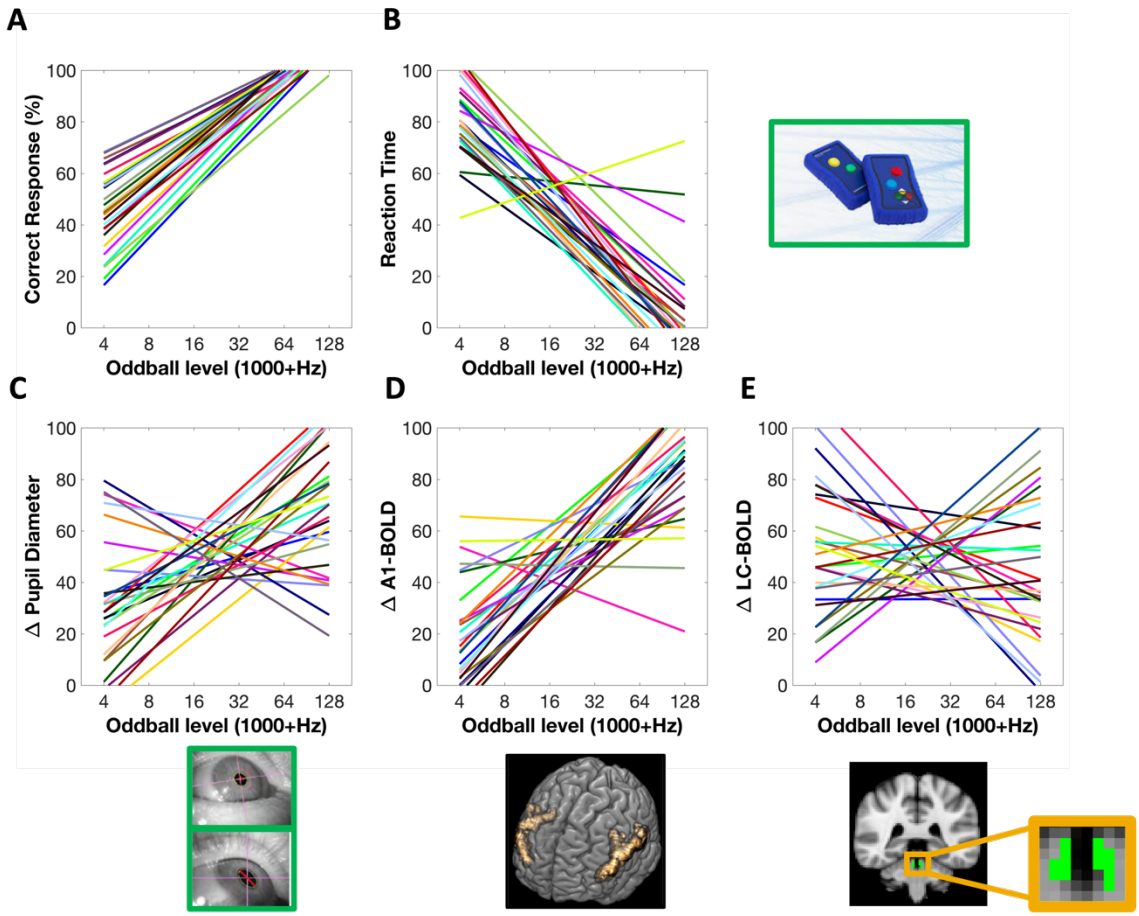

Figure 3. 11. Fitted lines across subjects before subject exclusion. (A) Accuracy (B) Normalized Reaction Time (C) Normalized pupil diameter change (D) Normalized A1-BOLD change and (E) Normalized LC-BOLD change. Each line shows a fit for a participant

pupillometry and BOLD signals. After characterizing the stimulus-response function by extracting slope values, we correlated the different slopes. Figure 3. 12 shows the scatter plots for these associations before subject exclusion. Across participants, greater accuracy slope value was associated with faster reaction time slope ($r(26) = -0.49$, $p = 0.01$). Greater A1-BOLD slope was associated with faster reaction time slope ($r(26) = -0.56$, $p < 0.01$) as well as higher accuracy slope ($r(26) = 0.41$, $p < 0.02$).

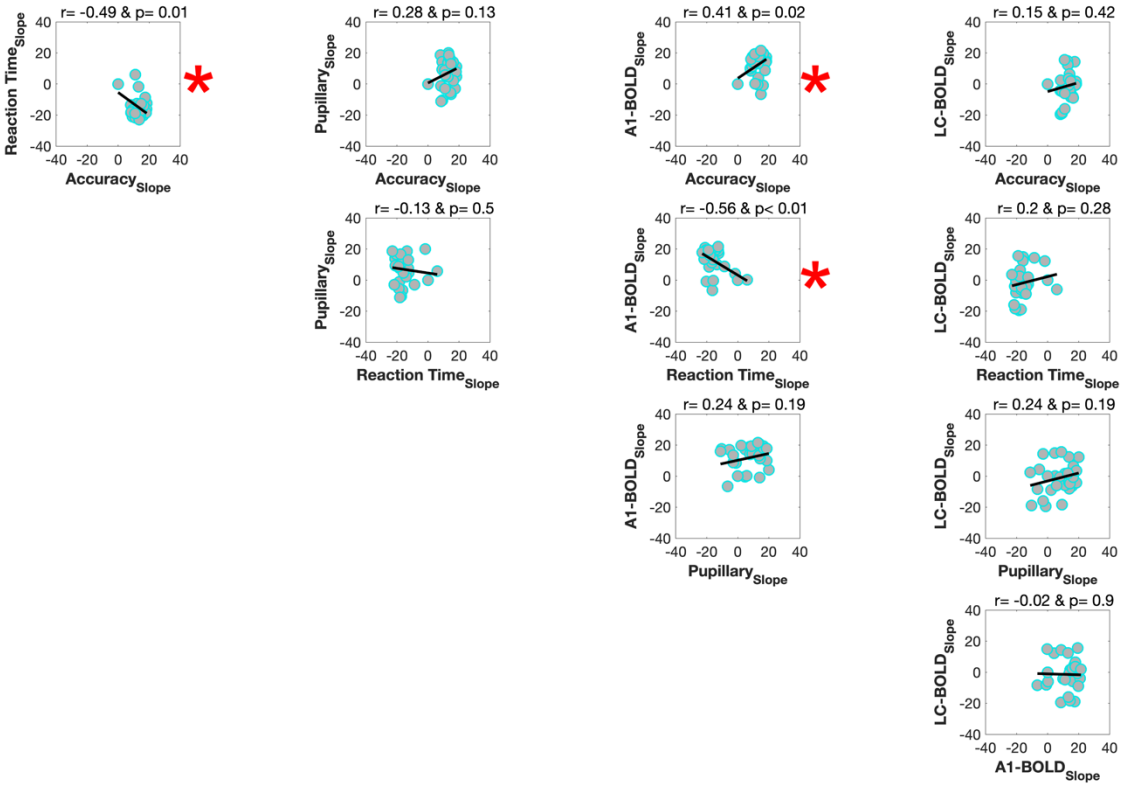

Figure 3. 12. Associations between the estimated slope of different measures (before subject exclusion). Behavioral performance (Accuracy and Reaction Time), pupillary response, A1-BOLD and LC-BOLD

These correlations remained significant when we excluded 10 subjects with positive slope for reaction time and negative slopes for pupillometry and A1-BOLD fits. As a result, a positive correlation between slope of pupillary response and accuracy ($r(16) = 0.47$, $p = 0.03$) and a negative correlation with slope of reaction time ($r(16) = -0.52$, $p = 0.01$) emerged. These findings are significant as pupillary dilation slope is a measure of phasic arousal or phasic LC-NE activity (Aston-Jones & Cohen, 2005; Murphy et al., 2011). As phasic arousal facilitates task-relevant activity (exploitation), it is associated with better performance which is indexed here by faster reaction time slope and higher accuracy slope.

Moreover, we found a trend for a association between pupillary slope and A1-BOLD slope ($r(16) = 0.41$, $p = 0.07$), indicating that phasic arousal promotes auditory processing by increasing A1-BOLD signal slope. Slope of LC-BOLD; however, failed to show any significant associations with any of the measures. Line fits and scatter plots of the associations are shown in the supplementary material Figure S3. 5, Figure S3. 6 respectively. Measures of goodness of fits (R^2) can be found in the supplementary material Table S3.1.

Discussion

Decades of research have revealed that by projecting NE, LC plays a significant role in cortical arousal, perception, attention, decision-making, learning, memory and cognition in general (Aston-Jones & Cohen, 2005; Murphy et al., 2011; Angela & Dayan, 2005; Berridge & Waterhouse, 2003; Clayton et al., 2004; Cohen Hoffing & Seitz, 2015; Silvetti et al., 2013; Sara, 2009). Studies, majorly in animal models, have demonstrated that LC neurons exhibit two distinct modes of activity: phasic(task-relevant), and tonic (spontaneous) which differ in NE releasing properties and behavioral manifestation (Devilbiss & Waterhouse, 2011; Vazey et al., 2018). These two modes of activity are not mutually exclusive, and the balance between phasic and tonic LC activity produces an optimal level of performance (Jepma & Nieuwenhuis, 2010) in a manner that mirrors the inverted-U relationship between the arousal/tonic-LC and task engagement/phasic-LC (Aston-Jones & Cohen, 2005). A key challenge in understanding LC function in humans is the lack of direct measures. Although advances in neuroimaging have opened new

horizons to investigate the mechanisms of the human brain non-invasively, limitations of studying LC persists. Small size and near-ventricle location of the LC structure makes the extracted LC-BOLD signal susceptible to image noise, scanner noise, motion, cardiac pulsation, and respiratory cycles (Keilholz et al., 2017). Therefore, there is an increasing need to establish methods to estimate LC-NE and arousal functions. Several studies have used pupillometry and P300 Event-related potential as electrophysiological markers of using an auditory oddball task (Nieuwenhuis eta l., 2005; Murphy et al., 2011). For example, results of Gilzenrat et al. (2010) and Murphy et al. (2011) suggest that integration of information across different neuropsychological markers and task engagement measures (indexed by behavioral accuracy and reaction time) can serve to estimate the cortical arousal function). However, this approach first requires designing an experimental paradigm that is sensitive enough to characterize how LC engagement may change stimulus encoding, processing and behavior.

Thus, in this study we conducted a multi-measure experiment and took the first step to develop an experimental tool to advance future non-invasive studies of the LC-NE system in humans. We examined behavioral performance, pupillometry response and neural fMRI response in the context of a modified auditory oddball task with multiple oddball levels. We extracted the psychometric function and reaction time trend by examining the behavioral accuracy and reaction time as a function of oddball strength. As predicted, our results showed that decision accuracy and response time improved as oddball saliency increased, providing proof of concept that at the level of behavioral performance,

the oddball detection task captures variability and enables extraction of psychometric function. Similarly, we extracted the pupillometric function that revealed the relationship between pupillary dilation response and oddball saliency. Our findings showed that stronger oddballs elicited larger pupil dilation and the pupillometric function captured the variability of responses across the group of recruited subjects. Finally, we extracted the neurometric functions in the task-relevant areas, the primary auditory cortex and LC, by examining BOLD responses in these areas as a function of oddball saliency. The neurometric function in primary A1 showed that auditory processing increased as a function of oddball saliency, suggesting the viability of our oddball detection task in capturing variance in the neural domain. Interpreting the neurometric function in LC should be approached with caution due to the substantial noise in the baseline LC-BOLD activity which can be due to imaging or preprocessing noise (Keilholz et al., 2017; Turker et al., 2021)

These results are significant as, to our knowledge, our paradigm is the first that enables extraction of stimulus-response functions in different levels of task processing, behavior, pupillometry and neural levels, which may together be used as markers of LC-NE activity. Moreover, the monotonic increase of accuracy (and decrease of reaction time), pupil dilation and A1-BOLD as a function of oddball saliency possibly reflect the additive influence of stimulus novelty and strength. We also applied the isometric handgrip manipulation previously used in several studies (Mather et al., 2020; Nielsen & Mather, 2015) which suggested that squeezing an isometric stress ball was sufficient to engage LC

activity in older adults and female participants; however, our findings suggest that this physiological stress manipulation did not induce any significant changes in LC activity (evaluated via pupil diameter response; see Figure S3) when tested in younger adults. Beside pupillometry response, this manipulation did not result in any statistically significant differences between other stimulus-response functions of behavior, A1-BOLD and LC-BOLD, (see Figures S1, S2, and S4). This showed that the extracted stimulus-response functions were robust to a squeeze stressor. For neurometric function of LC however, the handgrip session resulted in substantial noise in baseline LC activity. One explanation for our results is the significant individual differences in cognitive abilities, stress response and LC-NE system function (LoTemplio et al., 2021; Wood et al., 2017). After demonstrating that the effect of stimulus frequency was reflected across various measures, we characterized the stimulus response functions using slope of linear fits and further investigated the relationships. This approach allowed us to summarize the captured variability of the responses as a function of stimulus level, in a single data point known as slope. Similar to above-mentioned work, we showed that phasic pupillometry slope, a proxy of phasic arousal, was associated with improved task performance. Higher phasic pupil slope was associated with higher accuracy slope and faster reaction time slope as well as higher slope for auditory processing. However, LC-BOLD signal did not show any significant relationship with any of the task related measures.

Our study presents several limitations that should be addressed in the future research. As mentioned earlier LC-BOLD is susceptible to cardiac and respiratory noise

signals and correcting for the physiological noise is vital for the quality of LC-BOLD signal (Liu et al., 2017) which we were unable to do due to lack of such recordings. Therefore, substantial noise in LC signal has remained which makes the interpretation of LC neurometric function and estimated slopes unreliable. Here we used a handgrip manipulation as a stressor to modulate LC signal; however, this manipulation did not result in significant differences between the extracted stimulus-response functions. Therefore, future research should examine other methods of eliciting stress in a laboratory setting such as the cold pressor test (Marmon & Enoka, 2010; Schwabe & Schächinger, 2018) delivering small electric pulses (Stark et al., 2006; Oyarzún et al., 2012), presentation of emotionally arousing pictures (International Affective Picture System, IAPS; Lang et al., 2008) and sounds (International Affective Digitized Sound System, IADS; Bradley & Lang, 2007) , besides manual compression of an isometric squeeze ball (Hartwich et al., 2010). Also, recent research has suggested that using a manual compression concurrent with the task at hand can effectively modulate the arousal (Park et al., 2021). Of note, we aimed to go beyond the effect of normal LC activity on perceptual decision-making by manipulating LC activity using primary sensory stressors and investigate the human electrophysiology, behavior and extracted stimulus-response functions during the above-mentioned auditory oddball task which was put to hold due to COVID pandemic. Moreover, future studies should increase the sample size by increasing the number of the trials to increase the statistical power. This is particularly important for neurophysiological recordings that contain extraneous noise which can be decreased by increasing the number

of data points. Finally, although the grand average measures effectively decrease noise, to draw a dynamic picture of the moment-by-moment LC activity and its markers, single-measure variables should be used. Thus, an effective denoising procedure should be implemented to obtain reliable single-trial measures of pupillometry and BOLD data. For example, Quiroga and Garcia (2003) consider averaged electrophysiological signals as a denoising template and gather wavelet coefficients by applying wavelet transformation to this template. The retrieved coefficients were later applied to single-trial measures.

Conclusion

Our results demonstrated the viability of our modified auditory oddball paradigm in characterizing stimulus-response functions across a number of measures: psychometric, pupillometric and neurometric. We also introduced a stressor (handgrip) condition, aimed at modulating the described stimulus-response functions. The effects of this manipulation were subtle. After establishing the appropriate tools for experimenting and capturing the variability across multiple levels of processing, the next steps in our project will be to further optimize our experimental paradigm, neuroimage data acquisition parameters and use optimal pre-processing and processing methods.

References

Alnæs, D., Sneve, M. H., Espeseth, T., Endestad, T., van de Pavert, S. H. P., & Laeng, B. (2014). Pupil size signals mental effort deployed during multiple object tracking and predicts brain activity in the dorsal attention network and the locus coeruleus. *Journal of vision*, *14*(4), 1-1.

Angela, J. Y., & Dayan, P. (2005). Uncertainty, neuromodulation, and attention. *Neuron*, *46*(4), 681-692.

Aston-Jones, G., Rajkowski, J., Kubiak, P., & Alexinsky, T. (1994). Locus coeruleus neurons in monkey are selectively activated by attended cues in a vigilance task. *Journal of Neuroscience*, *14*(7), 4467-4480.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.*, *28*, 403-450.

Berridge, C. W., & Waterhouse, B. D. (2003). The locus coeruleus–noradrenergic system: modulation of behavioral state and state-dependent cognitive processes. *Brain research reviews*, *42*(1), 33-84.

Bradley, M. M., & Lang, P. J. (2007). The International Affective Digitized Sounds (; IADS-2): Affective ratings of sounds and instruction manual. *University of Florida, Gainesville, FL, Tech. Rep. B-3*.

Brainard DH (1997) The psychophysics toolbox. Spat Vis 10:433–436. Cavanagh P, Alvarez GA (2005) Tracking multiple targets with multifocal attention. Trends Cogn Sci 9:349–354.

Chen, I. W., Helmchen, F., & Lütcke, H. (2015). Specific early and late oddball-evoked responses in excitatory and inhibitory neurons of mouse auditory cortex. *Journal of Neuroscience*, *35*(36), 12560-12573.

Clayton, E. C., Rajkowski, J., Cohen, J. D., & Aston-Jones, G. (2004). Phasic activation of monkey locus ceruleus neurons by simple decisions in a forced-choice task. *Journal of Neuroscience*, *24*(44), 9914-9920.

Cohen, M. X. (2014). *Analyzing neural time series data: theory and practice*. MIT press.

Cohen Hoffing, R., & Seitz, A. R. (2015). Pupillometry as a glimpse into the neurochemical basis of human memory encoding. *Journal of cognitive neuroscience*, *27*(4), 765-774.

de Gee, J. W., Colizoli, O., Kloosterman, N. A., Knapen, T., Nieuwenhuis, S., & Donner, T. H. (2017). Dynamic modulation of decision biases by brainstem arousal systems. *Elife*, *6*, e23232.

Devilbiss, D. M., & Waterhouse, B. D. (2011). Phasic and tonic patterns of locus coeruleus output differentially modulate sensory network function in the awake rat. *Journal of neurophysiology*, *105*(1), 69-87.

Edeline, J. M., Manunta, Y., & Hennevin, E. (2011). Induction of selective plasticity in the frequency tuning of auditory cortex and auditory thalamus neurons by locus coeruleus stimulation. *Hearing research*, *274*(1-2), 75-84.

Einhäuser, W., Stout, J., Koch, C., & Carter, O. (2008). Pupil dilation reflects perceptual selection and predicts subsequent stability in perceptual rivalry. *Proceedings of the National Academy of Sciences*, *105*(5), 1704-1709.

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., ... & Dale, A. M. (2002). Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. *Neuron*, *33*(3), 341-355.

Foote, S. L., Freedman, R., & Oliver, A. P. (1975). Effects of putative neurotransmitters on neuronal activity in monkey auditory cortex. *Brain research*, *86*(2), 229-242.

Foote, S. L., Aston-Jones, G., & Bloom, F. E. (1980). Impulse activity of locus coeruleus neurons in awake rats and monkeys is a function of sensory stimulation and arousal. *Proceedings of the National Academy of Sciences*, *77*(5), 3033-3037.

Gilzenrat, M. S., Nieuwenhuis, S., Jepma, M., & Cohen, J. D. (2010). Pupil diameter tracks changes in control state predicted by the adaptive gain theory of locus coeruleus function. *Cognitive, Affective, & Behavioral Neuroscience*, *10*(2), 252-269.

Gold, J. I., & Shadlen, M. N. (2007). The neural basis of decision making. *Annual review of neuroscience*, *30*.

Hartwich, D., Fowler, K. L., Wynn, L. J., & Fisher, J. P. (2010). Differential responses to sympathetic stimulation in the cerebral and brachial circulations during rhythmic handgrip exercise in humans. *Experimental physiology*, *95*(11), 1089-1097.

Huettel, S. A., & McCarthy, G. (2004). What is odd in the oddball task?: Prefrontal cortex is activated by dynamic changes in response strategy. *Neuropsychologia*, *42*(3), 379-386.

Jepma, M., & Nieuwenhuis, S. (2011). Pupil diameter predicts changes in the exploration–exploitation trade-off: Evidence for the adaptive gain theory. *Journal of cognitive neuroscience*, *23*(7), 1587-1596.

Keilholz, S. D., Pan, W. J., Billings, J., Nezafati, M., & Shakil, S. (2017). Noise and non-neuronal contributions to the BOLD signal: applications to and insights from animal studies. *Neuroimage*, *154*, 267-281.

Lang, P.J., Bradley, M.M., & Cuthbert, B.N. (2008). International affective picture system (IAPS): Affective ratings of pictures and instruction manual. Technical Report A-8. University of Florida, Gainesville, FL.

Langley, J., Hussain, S., Flores, J.J., Bennett, I.J., Hu, X. (2020b) Characterization of age-related microstructural changes in locus coeruleus and substantia nigra pars compacta. Neurobiol Aging, 87:89-97.

Liao, H. I., Yoneya, M., Kidani, S., Kashino, M., & Furukawa, S. (2016). Human pupillary dilation response to deviant auditory stimuli: Effects of stimulus properties and voluntary attention. *Frontiers in Neuroscience*, *10*, 43.

Linden, D. E., Prvulovic, D., Formisano, E., Völlinger, M., Zanella, F. E., Goebel, R., & Dierks, T. (1999). The functional neuroanatomy of target detection: an fMRI study of visual and auditory oddball tasks. *Cerebral cortex*, *9*(8), 815-823.

Liu, K. Y., Marijatta, F., Hämmerer, D., Acosta-Cabronero, J., Düzel, E., & Howard, R. J. (2017). Magnetic resonance imaging of the human locus coeruleus: a systematic review. *Neuroscience & Biobehavioral Reviews*, *83*, 325-355.

LoTemplio, S., Silcox, J., Federmeier, K. D., & Payne, B. R. (2021). Inter-and intra-individual coupling between pupillary, electrophysiological, and behavioral responses in a visual oddball task. *Psychophysiology*, *58*(4), e13758.

Manunta, Y., & Edeline, J. M. (2004). Noradrenergic induction of selective plasticity in the frequency tuning of auditory cortex neurons. *Journal of neurophysiology*, *92*(3), 1445-1463.

Marmon, A. R., & Enoka, R. M. (2010). Comparison of the influence of two stressors on steadiness during index finger abduction. *Physiology & behavior*, *99*(4), 515-520.

Mather, M., Huang, R., Clewett, D., Nielsen, S. E., Velasco, R., Tu, K., ... & Kennedy, B. L. (2020). Isometric exercise facilitates attention to salient events in women via the noradrenergic system. *Neuroimage*, *210*, 116560.

McGinley, M. J., Vinck, M., Reimer, J., Batista-Brito, R., Zagha, E., Cadwell, C. R., ... & McCormick, D. A. (2015). Waking state: rapid variations modulate neural and behavioral responses. *Neuron*, *87*(6), 1143-1161.

Murphy, P. R., Robertson, I. H., Balsters, J. H., & O'connell, R. G. (2011). Pupillometry and P3 index the locus coeruleus–noradrenergic arousal function in humans. *Psychophysiology*, *48*(11), 1532-1543.

Murphy, P. R., O'connell, R. G., O'sullivan, M., Robertson, I. H., & Balsters, J. H. (2014). Pupil diameter covaries with BOLD activity in human locus coeruleus. *Human brain mapping*, *35*(8), 4140-4154.

Nieuwenhuis, S., Aston-Jones, G., & Cohen, J. D. (2005). Decision making, the P3, and the locus coeruleus--norepinephrine system. *Psychological bulletin*, *131*(4), 510.

O'Connell, R. G., Dockree, P. M., Robertson, I. H., Bellgrove, M. A., Foxe, J. J., & Kelly, S. P. (2009). Uncovering the neural signature of lapsing attention: electrophysiological signals predict errors up to 20 s before they occur. Journal of Neuroscience, 29(26), 8604-8611.

Park, H. B., Ahn, S., & Zhang, W. (2021). Visual search under physical effort is faster but more vulnerable to distractor interference. *Cognitive Research: Principles and Implications*, *6*(1), 1-14.

Pelli, D. G., & Vision, S. (1997). The VideoToolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision*, *10*, 437-442.

Polich, J. (2007). Updating P300: an integrative theory of P3a and P3b. *Clinical neurophysiology*, *118*(10), 2128-2148.

Quiroga, R. Q., & Garcia, H. (2003). Single-trial event-related potentials with wavelet denoising. *clinical neurophysiology*, *114*(2), 376-390.

Oyarzún, Javiera P., Diana Lopez-Barroso, Lluís Fuentemilla, David Cucurell, Carmen Pedraza, Antoni Rodriguez-Fornells, and Ruth de Diego-Balaguer. "Updating fearful memories with extinction training during reconsolidation: a human study using auditory aversive stimuli." *PloS one* 7, no. 6 (2012): e38849.

Rajkowski J, Kubiak P, Aston-Jones G. 1993. Correlations between locus coeruleus (LC) neural activity, pupil diameter and behavior in monkey support a role of LC in attention. Soc. Neurosc. Abstr. 19:974.

Rajkowski, J., Kubiak, P., & Aston-Jones, G. (1994). Locus coeruleus activity in monkey: phasic and tonic changes are associated with altered vigilance. *Brain research bulletin*, *35*(5-6), 607-616.

Richer, F., & Beatty, J. (1987). Contrasting effects of response uncertainty on the task-evoked pupillary response and reaction time. *Psychophysiology*, *24*(3), 258-262.

Sara, S. J., & Bouret, S. (2012). Orienting and reorienting: the locus coeruleus mediates cognition through arousal. *Neuron*, *76*(1), 130-141.

Sara, S. J. (2009). The locus coeruleus and noradrenergic modulation of cognition. *Nature reviews neuroscience*, *10*(3), 211-223.

Servan-Schreiber, D., Printz, H., & Cohen, J. D. (1990). A network model of catecholamine effects: gain, signal-to-noise ratio, and behavior. *Science*, *249*(4971), 892-895.

Schwabe, L., & Schächinger, H. (2018). Ten years of research with the Socially Evaluated Cold Pressor Test: Data from the past and guidelines for the future. *Psychoneuroendocrinology*, *92*, 155-161.

Silvetti, M., Seurinck, R., van Bochove, M., & Verguts, T. (2013). The influence of the noradrenergic system on optimal control of neural plasticity. *Frontiers in behavioral neuroscience*, *7*, 160.

Squires, N. K., Squires, K. C., & Hillyard, S. A. (1975). Two varieties of long-latency positive waves evoked by unpredictable auditory stimuli in man. *Electroencephalography and clinical neurophysiology*, *38*(4), 387-401.

Stark, R., Wolf, O. T., Tabbert, K., Kagerer, S., Zimmermann, M., Kirsch, P., ... & Vaitl, D. (2006). Influence of the stress hormone cortisol on fear conditioning in humans: evidence for sex differences in the response of the prefrontal cortex. *Neuroimage*, *32*(3), 1290-1298.

Szabadi, E. (2013). Functional neuroanatomy of the central noradrenergic system. *Journal of psychopharmacology*, *27*(8), 659-693.

Turker, H. B., Riley, E., Luh, W. M., Colcombe, S. J., & Swallow, K. M. (2021). Estimates of locus coeruleus function with functional magnetic resonance imaging are influenced by localization approaches and the use of multi-echo data. *NeuroImage*, *236*, 118047.

Usher, M., Cohen, J. D., Servan-Schreiber, D., Rajkowski, J., & Aston-Jones, G. (1999). The role of locus coeruleus in the regulation of cognitive performance. *Science*, *283*(5401), 549-554.

Vazey, E. M., Moorman, D. E., & Aston-Jones, G. (2018). Phasic locus coeruleus activity regulates cortical encoding of salience information. *Proceedings of the National Academy of Sciences*, *115*(40), E9439-E9448.

Walz, J. M., Goldman, R. I., Carapezza, M., Muraskin, J., Brown, T. R., & Sajda, P. (2015). Prestimulus EEG alpha oscillations modulate task-related fMRI BOLD responses to auditory stimuli. *NeuroImage*, *113*, 153-163.

Wood, C. S., Valentino, R. J., & Wood, S. K. (2017). Individual differences in the locus coeruleus-norepinephrine system: relevance to stress-induced cardiovascular vulnerability. *Physiology & behavior*, *172*, 40-48.

Woolrich, M. W., Jbabdi, S., Patenaude, B., Chappell, M., Makni, S., Behrens, T., ... & Smith, S. M. (2009). Bayesian analysis of neuroimaging data in FSL. *Neuroimage*, *45*(1), S173-S186.

Yerkes, R. M., & Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Punishment: Issues and experiments*, 27-41.

Supplementary Material

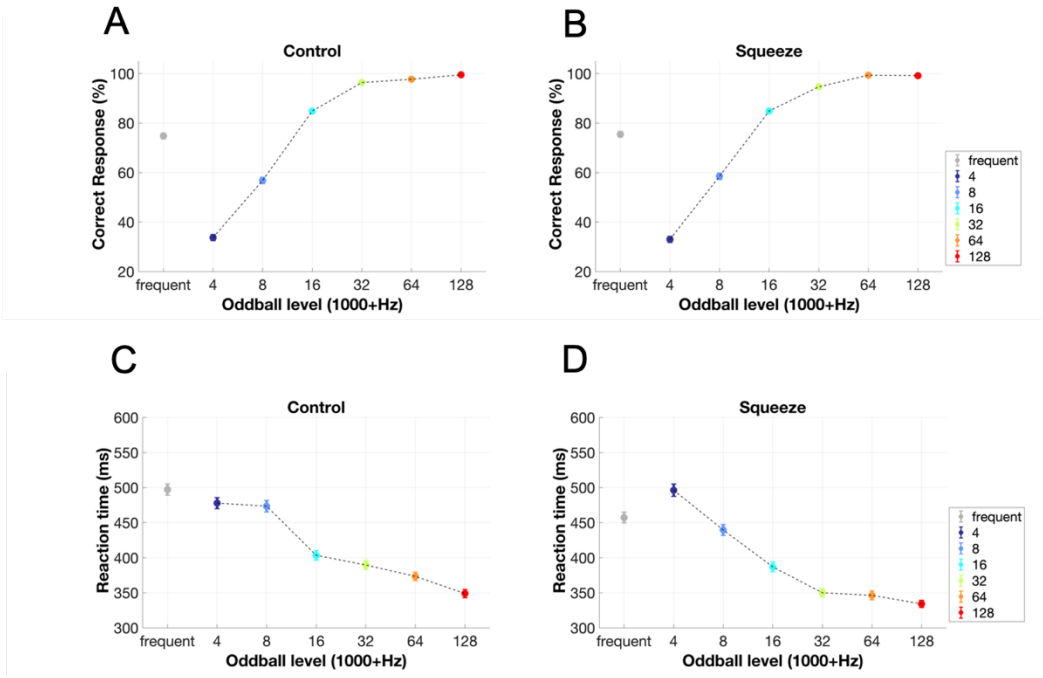

Figure S3. 1. Psychometric functions and Reaction time trend divided based on handgrip manipulation. (A) Psychometric function for control (B) Psychometric function for squeeze (C) Reaction time trend for control (D) Reaction time trend for squeeze

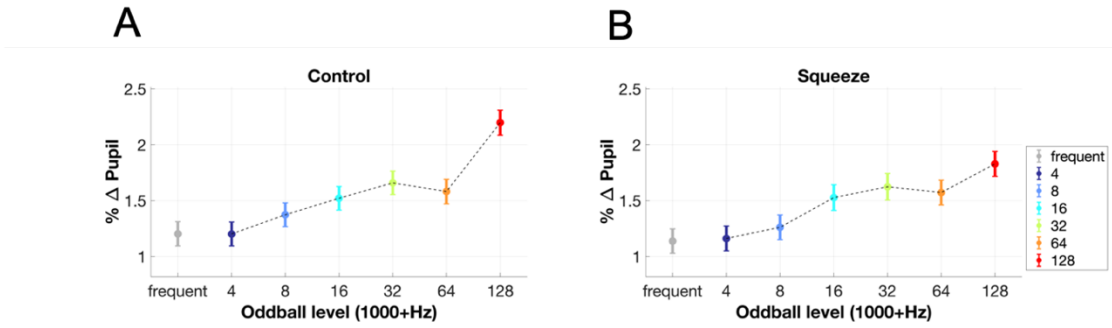

Figure S3. 2. Pupillometric function is divided based on handgrip manipulation. (A) Pupillometric function for control (B) Pupillometric function for squeeze

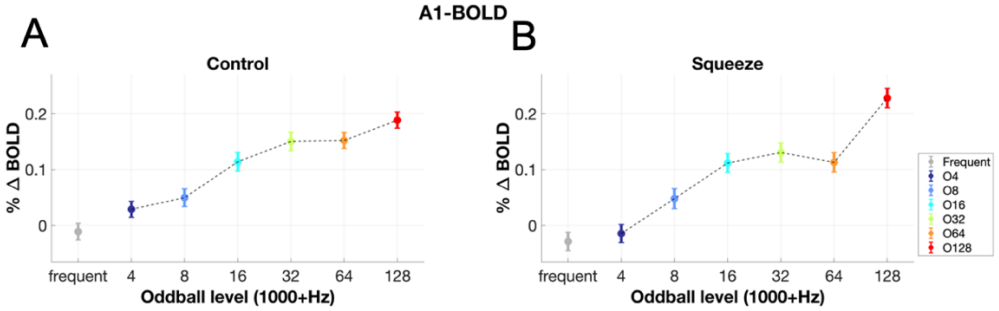

Figure S3. 3. Neurometric function is divided based on handgrip manipulation. (A) Pupillometric function for control (B) Pupillometric function for squeeze

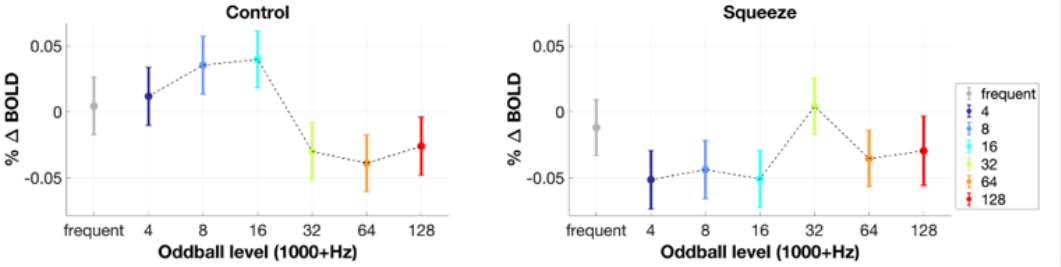

Figure S3. 4. Neurometric function is divided based on handgrip manipulation. (A) Pupillometric function for control (B) Pupillometric function for squeeze

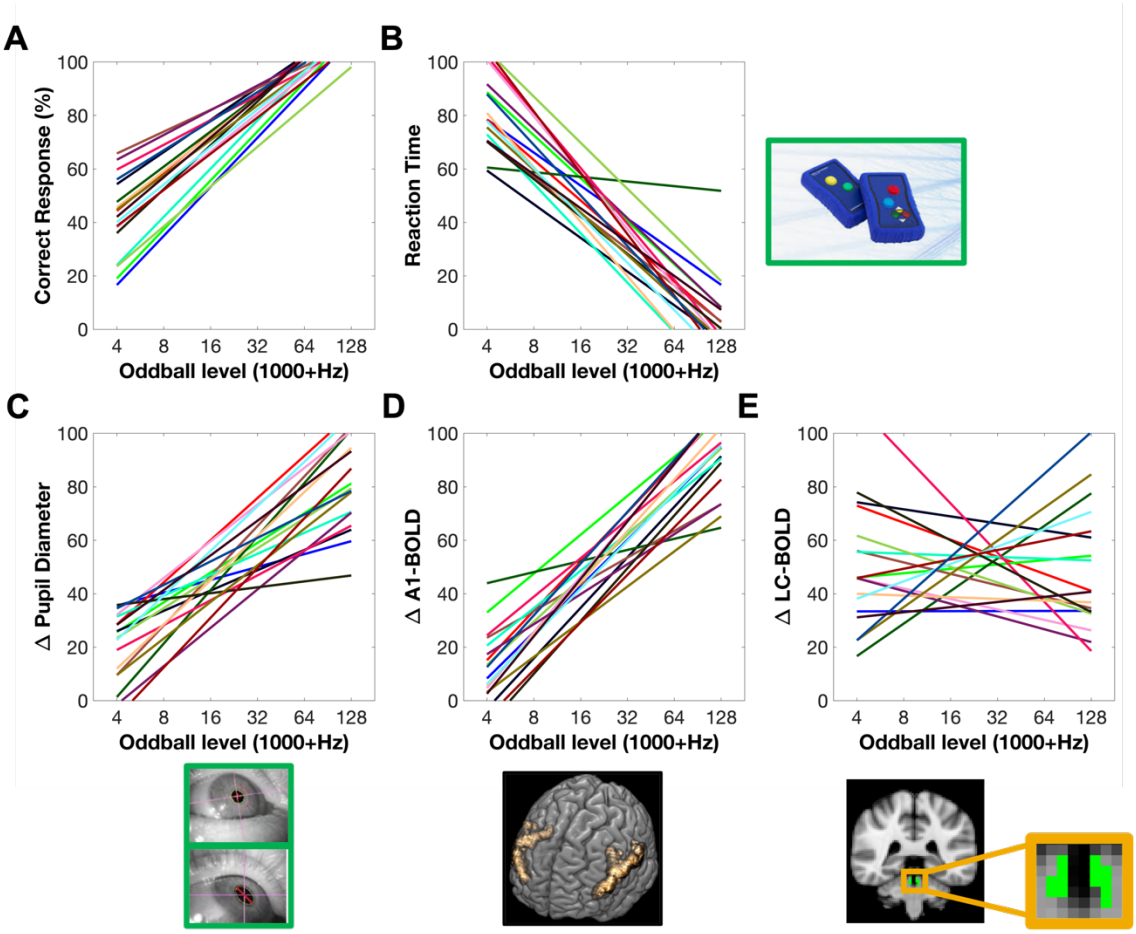

Figure S3. 5. Fitted lines across subjects after subject exclusion. (A) Accuracy (B) Normalized Reaction Time (C) Normalized pupil diameter change (D) Normalized A1-BOLD change and (E) Normalized LC-BOLD change. Each line shows a fit for a participant

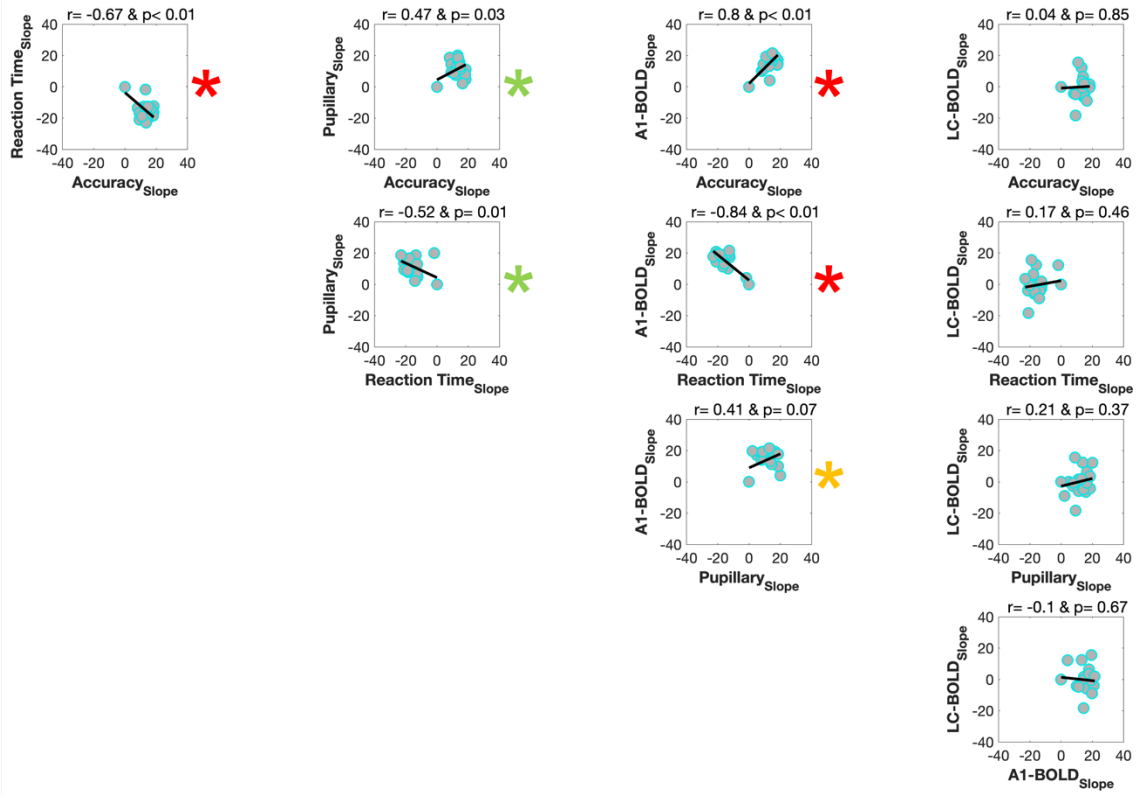

Figure S3. 6. Associations between the estimated slope of different measures (after subject exclusion). Behavioral performance (Accuracy and Reaction Time), pupillary response, A1-BOLD and LC-BOLD

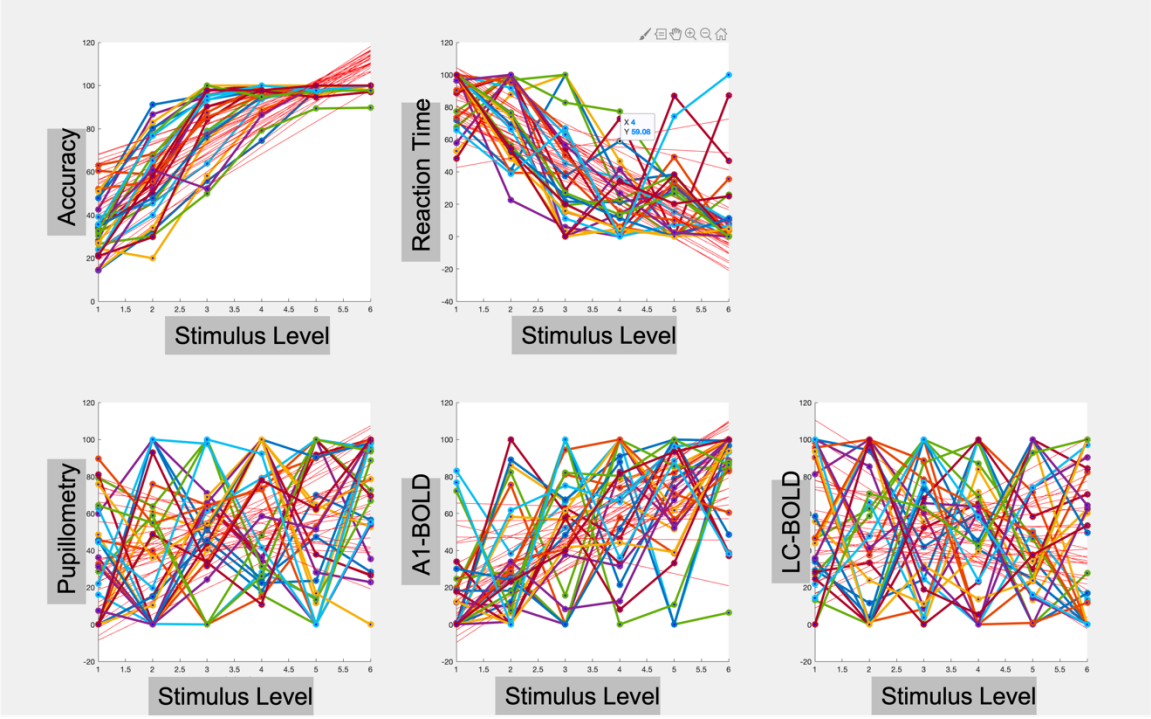

Figure S3. 7. All Stimulus-response functions overlayed with fitted lines.

Table S3. 1. Goodness of fit (R^2) for different measures. For Accuracy (ACC GOF), for Reaction Time (RT GOF), for Pupillometry (PUPIL GOF), for A1-BOLD (A1 GOF), and for LC-BOLD (LC GOF). Excluded subjects (positive slope for RT and negative slopes for pupillometry and A1-BOLD) are shown by orange shade.

	ACC GOF	RT GOF	PUPIL GOF	A1 GOF	LC GOF
Subject ID	R-squared	R-squared	R-squared	R-squared	R-squared
1	0.97	0.45	0.06	0.51	0.00
2	0.70	0.49	0.68	0.83	0.08
3	0.89	0.53	0.25	0.58	0.01
4	0.62	0.40	0.18	0.71	0.02
5	0.87	0.54	0.09	0.08	0.20
6	0.87	0.78	0.49	0.00	0.18
7	0.73	0.01	0.84	0.04	0.45
8	0.58	0.81	0.00	0.18	0.81
9	0.86	0.51	0.78	0.26	0.03
10	0.87	0.71	0.20	0.66	0.00
	0.00	0.00	0.00	0.00	0.00
12	0.61	0.61	0.34	0.57	0.68
13	0.93	0.72	0.36	0.55	0.09
14	0.91	0.96	0.50	0.97	0.04
15	0.79	0.19	0.02	0.24	0.48
16	0.72	0.83	0.58	0.27	0.06
17	0.85	0.79	0.25	0.46	0.87
18	0.72	0.72	0.61	0.72	0.00
19	0.73	0.44	0.01	0.91	0.26
20	0.72	0.70	0.06	0.00	0.37
21	0.72	0.72	0.06	0.78	0.04
22	0.77	0.64	0.47	0.43	0.42
23	0.89	0.78	0.63	0.66	0.13
24	0.82	0.89	0.02	0.54	0.45
25	0.64	0.75	0.32	0.74	0.01
26	0.89	0.93	0.87	0.68	0.03
	0.00	0.00	0.00	0.00	0.00
28	0.61	0.83	0.18	0.76	0.64
29	0.68	0.11	0.06	0.00	0.09
	0.00	0.00	0.00	0.00	0.00
31	0.72	0.46	0.40	0.96	0.01

General Discussion

Working Memory (WM) is a fundamental cognitive ability which correlates with a wide range of complex cognitive functions such as problem-solving, reasoning, learning, and planning of goal-directed behaviors (Miyake, & Shah, 1999; Swanson & Alloway, 2012). Due to the importance of WM, there has been a growing interest in understanding and enhancing WM by use of training interventions both in healthy individuals and populations with neurological disorders (Lawlor-Savage & Goghari, 2014). As a result, many companies started million-dollar businesses by creating off-the-shelf computerized programs and games. These companies claim broad generalization of training to untrained cognitive abilities, like Lumosity, CogniFit and Jungle Memory, to name a few. The study of the benefits of WM training have been pursued in lab settings of various research groups using rigorous scientific methodology and objective approaches as well (Deveau et al., 2015; Mohammed et al., 2017).

Numerous studies (Blacker, et al., 2017; Minear et al., 2016) have trained participants using a variety of WM tasks such as N-back, span tasks, immediate recall, etc. While most studies find improvements in the training task, it is controversial whether WM training gains transfer to similar WM tasks, and even more so to different tasks that may involve WM (Melby-Lervåg & Hulme, 2013; Au et al, 2015). For instance, some studies offered evidence of transfer from WM training to fluid intelligence and complex reasoning (Jaeggi et al., 2008; Klingberg et al., 2005), reading comprehension (Loosli, Buschkuehl, Perrig, & Jaeggi, 2012), arithmetic (Bergman-Nutley, & Klingberg, 2014), while others

reported no transfer effects to fluid intelligence or any other cognitive domains (Thompson et al., 2013; Estrada et al., 2015). To reconcile this dichotomous view, several meta-analyses interpreted these findings in support of the hypothesis that WM training is only beneficial to improve the trained task but has limited effect on other cognitive abilities (Sala, & Gobet, 2019), while others (Au et al., 2016) supported generalized efficacy of WM and attributed the discrepant results to incorrect analysis.

To address this controversy and further our understanding of WM and the factors that impact it, we have conducted a series of studies that were presented in three chapters of this thesis. In this process, we started with carefully characterizing the existing experimental paradigms in the field of WM training by focusing on behavioral measures. We then moved to gaining mechanistic understanding of ongoing brain activity by using brain EEG signals as important intermediaries of behavior. Finally, we advanced to establishing new experimental paradigms to set the stage for noninvasive estimation of LC activity. LC widely projects NE, and by doing so, modulates brain states and influences WM in many aspects ranging from information processing to retrieval and learning (Sara, 2009; Aston-Jones & Cohen, 2005).

More specifically, in chapter 1, we aimed to reconcile the ongoing debate on the efficacy of WM training and remind the reader that a more systematic review of the specific qualities of the training and transfer tasks is required rather than a dichotomous “effective” or “not effective” approach. We characterize the broad diversity of features (e.g., intervention length, stimulus modality, adaptivity) used in fifty-seven published studies

that used N-back training tasks and measured behavioral outcomes. The methodology used across studies is not consistent which is often ignored by existing meta-analyses. We demonstrate how these limitations deter cross-study comparison and prevent strong conclusions regarding efficacy of training from the published data. At this point however, we were not able to reveal anything about the underlying brain mechanisms. This brings us to the second chapter of this thesis.

In chapter 2, we aimed to understand the electrophysiological signatures of a popular WM task, N-Back task, by systematically comparing nine task structure and stimulus variations. The EEG's high temporal resolution compared to behavioral data provided a better understanding of the time course of the effect of stimulus encoding and response processing. Our results reveal significant differences in behavioral and electrophysiological signatures in response to both manipulations (task structure and stimulus type). Additionally, we observe differences beyond our experimental manipulations, such as the pre-processing method and the laboratory environment. We suggest that experimental factors such as stimulus type and task structure, but also analysis pipeline and laboratory differences, which are often overlooked, need to be accounted for when interpreting findings and making comparisons across studies.

After gaining mechanistic understanding using existing experimental paradigms, we combined our knowledge of chapters 1 and 2. This resulted in chapter 3 where we advanced to establish a new experimental paradigm to enable the study of the LC-NE system. As a first step, we verified our experimental paradigm by investigating stimulus

processing during a continuous performance auditory discrimination task. We collected a multi-measure dataset including behavioral, pupillometry and functional magnetic resonance imaging (fMRI) data. Our aim was to characterize LC-NE system activity by estimates of LC blood-oxygen-level- dependent (BOLD) signals, primary Auditory cortex (A1) BOLD signals, and pupillometry dilation responses and capture its moment-by-moment effect on stimulus processing and behavior. Here, we show that the paradigm enables extracting stimulus-response functions across all the three measures of behavior, pupillometry and fMRI. Thus, our paradigm is sensitive to capture the variability across different measures which can in turn be used to estimate LC-NE activity.

References

Anguera, J. A., Bernard, J. A., Jaeggi, S. M., Buschkuehl, M., Benson, B. L., Jennett, S., ... & Seidler, R. D. (2012). The effects of working memory resource depletion and training on sensorimotor adaptation. Behavioural brain research, 228(1), 107-115.

Arjona Valladares, A., Gómez, C. M., Rodríguez-Martínez, E. I., Barriga-Paulino, C. I., Gómez-González, J., & Diaz-Sánchez, J. A. (2020). Attention-deficit/hyperactivity disorder in children and adolescents: An event-related potential study of working memory. *European Journal of Neuroscience*, *52*(10), 4356-4369.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.*, *28*, 403-450.

Au, J., Sheehan, E., Tsai, N., Duncan, G. J., Buschkuehl, M., & Jaeggi, S. M. (2015). Improving fluid intelligence with training on working memory: A meta-analysis. Psychonomic bulletin & review, 22(2), 366-377.

Au, J., Katz, B., Buschkuehl, M., Bunarjo, K., Senger, T., Zabel, C., ... & Jonides, J. (2016). Enhancing working memory training with transcranial direct current stimulation. *Journal of cognitive neuroscience*, *28*(9), 1419-1432.

Baddeley, A. (2012). Working memory: theories, models, and controversies. Annual review of psychology, 63, 1-29.

Bergman-Nutley, S., & Klingberg, T. (2014). Effect of working memory training on working memory, arithmetic and following instructions. *Psychological research*, *78*(6), 869-877.

Blacker, K. J., Negoita, S., Ewen, J. B., & Courtney, S. M. (2017). N-back versus complex span working memory training. *Journal of cognitive enhancement*, *1*(4), 434-454.

Blasiman, R. N., & Was, C. A. (2018). Why is working memory performance unstable? A review of 21 factors. *Europe's journal of psychology*, *14*(1), 188.

Brouwer, A. M., Zander, T. O., Van Erp, J. B., Korteling, J. E., & Bronkhorst, A. W. (2015). Using neurophysiological signals that reflect cognitive or affective state: six recommendations to avoid common pitfalls. *Frontiers in neuroscience*, *9*, 136.

Christophel, T. B., Klink, P. C., Spitzer, B., Roelfsema, P. R., & Haynes, J. D. (2017). The distributed nature of working memory. *Trends in cognitive sciences*, *21*(2), 111-124.

Clewett, D. V., Huang, R., Velasco, R., Lee, T. H., & Mather, M. (2018). Locus coeruleus activity strengthens prioritized memories under arousal. *Journal of Neuroscience*, *38*(6), 1558-1574.

Cohen Hoffing, R., & Seitz, A. R. (2015). Pupillometry as a glimpse into the neurochemical basis of human memory encoding. *Journal of cognitive neuroscience*, *27*(4), 765-774.

D'Esposito, M., & Postle, B. R. (2015). The cognitive neuroscience of working memory. *Annual review of psychology*, *66*, 115-142.

Devilbiss, D. M., & Waterhouse, B. D. (2011). Phasic and tonic patterns of locus coeruleus output differentially modulate sensory network function in the awake rat. *Journal of neurophysiology*, *105*(1), 69-87.

Deveau, J., Jaeggi, S. M., Zordan, V., Phung, C., & Seitz, A. R. (2015). How to build better memory training games. *Frontiers in systems neuroscience*, *8*, 243.

Estrada, E., Ferrer, E., Abad, F. J., Román, F. J., & Colom, R. (2015). A general factor of intelligence fails to account for changes in tests' scores after cognitive practice: A longitudinal multi-group latent-variable study. *Intelligence*, *50*, 93-99.

Frydecka, D., Eissa, A. M., Hewedi, D. H., Ali, M., Drapała, J., Misiak, B., ... & Moustafa, A. A. (2014). Impairments of working memory in schizophrenia and bipolar disorder: the effect of history of psychotic symptoms and different aspects of cognitive task demands. *Frontiers in behavioral neuroscience*, *8*, 416.

Jaeggi, S. M., Buschkuehl, M., Jonides, J., & Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. Proceedings of the National Academy of Sciences, 105(19), 6829-6833.

Jaeggi, S. M., Studer-Luethi, B., Buschkuehl, M., Su, Y. F., Jonides, J., & Perrig, W. J. (2010). The relationship between n-back performance and matrix reasoning—implications for training and transfer. Intelligence, 38(6), 625-635.

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology*, *55*(4), 352.

Klingberg, T., Fernell, E., Olesen, P. J., Johnson, M., Gustafsson, P., Dahlström, K., ... & Westerberg, H. (2005). Computerized training of working memory in children with ADHD-a randomized, controlled trial. *Journal of the American Academy of child & adolescent psychiatry*, *44*(2), 177-186.

Lawlor-Savage, L., & Goghari, V. M. (2014). Working memory training in schizophrenia and healthy populations. *Behavioral Sciences*, *4*(3), 301-319.

Loosli, S. V., Buschkuehl, M., Perrig, W. J., & Jaeggi, S. M. (2012). Working memory training improves reading processes in typically developing children. *Child Neuropsychology*, *18*(1), 62-78.

Minear, M., Brasher, F., Guerrero, C. B., Brasher, M., Moore, A., & Sukeena, J. (2016). A simultaneous examination of two forms of working memory training: Evidence for near transfer only. *Memory & Cognition*, *44*(7), 1014-1037.

Melby-Lervåg, M., & Hulme, C. (2013). Is working memory training effective? A meta-analytic review. *Developmental psychology*, *49*(2), 270.

Melby-Lervåg, M., Redick, T. S., & Hulme, C. (2016). Working memory training does not improve performance on measures of intelligence or other measures of "far transfer" evidence from a meta-analytic review. *Perspectives on Psychological Science*, *11*(4), 512-534.

Miyake, A., & Shah, P. (Eds.). (1999). Models of working memory: Mechanisms of active maintenance and executive control. Cambridge University Press.

Mohammed, S., Flores, L., Deveau, J., Hoffing, R. C., Phung, C., Parlett, C. M., ... & Seitz, A. R. (2017). The benefits and challenges of implementing motivational features to boost cognitive training outcome. *Journal of Cognitive Enhancement*, *1*(4), 491-507.

Oliva, M. (2019). Pupil size and search performance in low and high perceptual load. *Cognitive, Affective, & Behavioral Neuroscience*, *19*(2), 366-376

Pergher, V., Shalchy, M. A., Pahor, A., Van Hulle, M. M., Jaeggi, S. M., & Seitz, A. R. (2019). Divergent research methods limit understanding of working memory training. Journal of Cognitive Enhancement, 1-21.

Rose, E. J., & Ebmeier, K. P. (2006). Pattern of impaired working memory during major depression. *Journal of affective disorders*, *90*(2-3), 149-161.

Sala, G., & Gobet, F. (2019). Cognitive training does not enhance general cognition. *Trends in cognitive sciences*, *23*(1), 9-20.

Sara, S. J., & Bouret, S. (2012). Orienting and reorienting: the locus coeruleus mediates cognition through arousal. *Neuron*, *76*(1), 130-141.

Shalchy, M. A., Pergher, V., Pahor, A., Van Hulle, M. M., & Seitz, A. R. (2020). N-Back Related ERPs Depend on Stimulus Type, Task Structure, Pre-processing, and Lab Factors. Frontiers in human neuroscience, 14.

Swanson, H. L., & Siegel, L. (2011). Learning disabilities as a working memory deficit. *Experimental Psychology*, *49*(1), 5-28.

Thompson, T. W., Waskom, M. L., Garel, K. L. A., Cardenas-Iniguez, C., Reynolds, G. O., Winter, R., ... & Gabrieli, J. D. (2013). Failure of working memory training to enhance cognition or intelligence. *PloS one*, *8*(5), e63614.

Vazey, E. M., Moorman, D. E., & Aston-Jones, G. (2018). Phasic locus coeruleus activity regulates cortical encoding of salience information. *Proceedings of the National Academy of Sciences*, *115*(40), E9439-E9448.

Wilhelm, O., Hildebrandt, A. H., & Oberauer, K. (2013). What is working memory capacity, and how can we measure it?. Frontiers in psychology, 4, 433.

Yaghoubi, K., Shalchy, M. A., Hussain, S., Chen, X., Benette, I., Mather, M., Hu, X., Seitz, A., & Peters, M. (2019). Computational fMRI Reveals Separable Representations Of Stimulus and Behavioral Choice In Auditory Cortex: A Tool for Studying the Locus Coeruleus Circuit. In *2019 Conference on Cognitive Computational Neuroscience*

Zokaei, N., & Husain, M. (2019). Working memory in Alzheimer's disease and Parkinson's disease. In Processes of Visuospatial Attention and Working Memory (pp. 325-344). Springer, Cham.