Physiology & Behavior 158 (2016) 43–53

Contents lists available at ScienceDirect

Physiology & Behavior

journal homepage: www.elsevier.com/locate/phb

Within- and between-session replicability of cognitive brain processes: An MEG study with an N-back task

L. Ahonen [a,*], M. Huotilainen [a,b], E. Brattico [c,d]

[a] Brain Work Research Centre, Finnish Institute of Occupational Health, Finland
[b] BioMag Laboratory, HUS Medical Imaging Center, Helsinki University Central Hospital, Finland
[c] Center for Music in the Brain (MIB), Department of Clinical Medicine, Aarhus University, Denmark
[d] Cognitive Brain Research Unit, Institute of Behavioural Sciences, University of Helsinki, Finland

HIGHLIGHTS

- Repeated recordings showed over time stability in cognition associated features ERF
- Latency showed less fluctuation compared to amplitude in within-subject comparisons
- M170 latency and LPP correlated with task performance in a cognitively demanding task

ARTICLE INFO

Article history:
Received 26 June 2015
Received in revised form 17 December 2015
Accepted 3 February 2016
Available online 6 February 2016

Keywords:
MEG
ERF
Replicability
Test-retest
Cognition
GFP

ABSTRACT

In the vast majority of electrophysiological studies on cognition, participants are only measured once during a single experimental session. The dearth of studies on test-retest reliability in magnetoencephalography (MEG) within and across experimental sessions is a preventing factor for longitudinal designs, imaging genetics studies, and clinical applications. From the recorded signals, it is not straightforward to draw robust and steady indices of brain activity that could directly be used in exploring behavioral effects or genetic associations. To study the variations in markers associated with cognitive functions, we extracted three event-related field (ERF) features from time-locked global field power (GFP) epochs using MEG while participants were performing a numerical N-back task in four consecutive measurements conducted during two different days separated by two weeks.

We demonstrate that the latency of the M170, a neural correlate associated with cognitive functions such as working memory, was a stable parameter and did not show significant variations over time. In addition, the M170 peak amplitude and the mean amplitude of late positive component (LPP) also expressed moderate-to-strong reliability across multiple measures over time over many sensor spaces and between participants. The M170 amplitude varied more significantly between the measurements in some conditions but showed consistency over the participants over time. In addition we demonstrated significant correlation with the M170 and LPP parameters and cognitive load. The results are in line with the literature showing less within-subject fluctuation for the latency parameters and more consistency in between-subject comparisons for amplitude based features. The within-subject consistency was apparent also with longer delays between the measurements. We suggest that with a few limitations the ERF features show sufficient reliability and stability for longitudinal research designs and clinical applications for cognitive functions in single as well as cross-subject designs.

1. Introduction

Electromagnetic correlates of cognitive functions in the human brain form an intriguing topic, which has been studied abundantly for decades [1]. In electrophysiology, cognitive processes are traditionally studied with long-latency neural responses of the event-related potential (ERP), or field (ERF) that are extracted from the continuous electroencephalograph (EEG) or magnetoencephalograph (MEG), respectively, by signal averaging. Typically, the ERPs or ERFs, with associations to cognition, peak several hundreds of milliseconds after the onset of an event and originate in associative cortical areas. The use of MEG as a method can be preferable in some situations since it reveals activity with high spatial and temporal precision to provide information on the overall stability in neural activation during comprehensive

[*] Corresponding author.
E-mail address: lauri.ahonen@ttl.fi (L. Ahonen).

http://dx.doi.org/10.1016/j.physbeh.2016.02.006
0031-9384/© 2016 Elsevier Inc. All rights reserved.

cognitive tasks [2, 3]. Unraveling the neural basis of human information processing is intriguing and potentially beneficial task since the longer-latency ERP components have shown some promise as tools in clinical applications [4, 5]. The importance of electromagnetic measures relies on their extreme accuracy in time, which enables a deep understanding of the temporal succession of neural events. In contrast, the temporal approximation of metabolic measures, such as those provided by functional magnetic resonance imaging (fMRI), typically average over several seconds, skipping over the fast, local neural events and evidencing only the dominant, systemic ones [6].

Variations in recorded brain responses result partially from noise and partially from true and persistent inter-individual differences, e.g., endophenotypes. Endophenotypes are specific traits that are meaningfully associated with a disorder of interest, a type of behaviour or exposure to a specific environment, and are an interesting source of information for brain research to tackle [7]. Since the N-back has shown promise in linking genetic traits to cognitive performance [8], we chose to use the task in our test-retest study on ERF reliability since it is cognitively a much more demanding task than the ones used previously in studies of ERP/ERF replicability [9]. Some properties in evoked brain activations are successfully linked to genome and gene expression [10, 11]. However, due to the interaction and variations in environmental factors, internal conditions in participants' physiological state, as well as task dependent variables, the results of electrophysiological measurements on higher cognitive activations are difficult to interpret [12, 13]. The lack of studies concentrating on test-retest reliability and replicability of electrophysiological correlates of working memory is a serious concern and partly preventing eletrophysiological research on the topic.

Using PubMed searches with keywords ‘replicability’ and ‘test-retest’, and restricting the results to studies with MEG, we found 16 studies of which none considered cognitive task-related activation. In addition, one reliability study by [14] on graph metrics stability was found outside the PubMed search. The study reports greater stability in connections between cortical areas in cognitively demanding situations compared to the resting state. Within EEG research, test-retest studies on different features of evoked potentials has a long history reaching back three decades. A large number of studies on the test-retest reliability in EEG inspect the mismatch negativity (MMN) [15, 16, 17] reporting fair stability in early ERP components, both at individual and at interindividual level. And many of the studies focus also on latter components and error-related negativity in EEG [18, 19, 20, 21]. These studies have found stability in P3 component latency over weeks, however reporting earlier components as more stable over longer period of time. The studies regarding error-related features report fair stability in interindividual tests but suggest high number of trials. MEG studies with reliability as their main research question concentrate mainly on early sensory responses, e.g., on the auditory N1 response [22], and on somatosensory evoked fields (SEF) [23]. These studies suggest equal stability for both EEG and MEG signals. We found only three studies focusing on the replicability of evoked responses during a demanding cognitive task [14, 24, 25]. Huffmeijer et al. [24] recommend more trials to be used for latter components in ERPs while the early sensory components can be studied with fewer trials. While [25] reports stronger replicability to test-retest amplitudes compared to split-half amplitudes of various ERP components.

Here, we adopted a basic visually presented N-back paradigm as a cognitive task. N-back is a classic working memory test and has been used in electrophysiological studies as a cognitive task for several reasons [1]. Performing an N-back task requires monitoring, updating, and manipulating the information flow on-line and is assumed to occupy numerous key processes within working memory and other executive functions [26]. N-back is abundantly used and reviewed in the field of neuroimaging and imaging genetics, mainly in fMRI [27, 28], and has also been used in a replicability study of fMRI responses [29]. Thus, it is well suited for studying stability of neural activations. Recently, the task has also received publicity within the field of cognitive training, advocating its use as a cognitive performance measure [30].

We aimed to explore the source of variability between participants and to study the stability of repeated measures within participants. In this repeated-measures cognitive MEG paradigm, we investigated the effect of daily variations within healthy participants performing cognitively demanding tasks against the instrument derived and random noise sources. To explore the traces of individual ERFs we computed the global field power (GFP) for the MEG data. In MEG, GFP reduces the dimensions of the multisensory electrophysiological data and yet serves as an excellent quantifier for neural activity. GFP is a global and well-established quantifier of the overall neuronal field strength. It is based on spatial standard deviation, and quantifies the amount of activity of all neuronal sources at a given time instant to its largest possible extent. Hence, it serves as an excellent summation to study traces of event related fields (ERF) [31, 32]. It is also a measure with very few presumptions. Unlike many techniques such as source modelling, GFP does not require a priori assumptions on the studied brain responses, allowing a more direct and easily replicable estimate of total brain activity. Thus, GFP is a good quantifier of MEG activity also when large amounts of recordings need to be analyzed in automated paradigms such as in, e.g., imaging genetics.

In particular, we focused on the ERF component termed M170, peaking at around 150–200 ms from event onset, and reflecting attention [33] and cognitive processes such as face recognition [34], and complex lexical decisions [35, 36]. The loci of M170 neural generators converge to left or right fusiform gyrus, depending on the task [37]. We also examined the long-latency ERF component labeled late positive potential (LPP). This somewhat controversial ERF feature is elicited during evaluative classification of various stimuli [38, 39]. For extracting the LPP, we measured the difference between the target and non-target stimuli in a post-response time window. This modulation of ERF strength begins approximately 300–400 ms after stimulus onset and lasts several hundreds of milliseconds [40]. Its neural generators have been identified in lateral to frontal regions for cognitive tasks [41]. Despite its controversial status, LPP seems to, e.g., consistently reflect the awareness of an error [42, 43].

2. Methods

2.1. Protocol, participants, and questionnaires

Seven healthy right-handed participants (2 males, mean(sd) age 26(5.8) years) were recorded in four replicated measurements. Subjects were recruited via mailing lists and compensated for the time used (equivalent to ca. 24€). The study consisted of four separate sessions for each participant. The measurements were conducted during two days separated by a period of approximately two weeks so that each measurement day included two repeated measurements. Each measurement consisted of an N-back task and two other cognitive tasks that lasted altogether for approximately an hour. Thus each session consisted of 2 h for the tasks, and an additional hour for preparations and questionnaires about vigilance, performance, and mood (KSS (Karolinska Sleepiness Scale), NASA-TLX (NASA Task Load Index), and POMS (Profile of Mood States) [44, 45, 46] respectively). Here we analyze the test-retest reliability in all of the concluding 28 N-back blocks (7 subjects, 4 blocks each), resulting in over 500 min of recorded MEG data.

We aimed at inducing some natural variation in the mental state of the participants during the N-back measurements. For this reason, the two session days differed in the type of pause the participants had between the two measurement blocks (see Fig. 1): one break was made pleasant and the other one unpleasant. Other parameters such as caffeine consumption and the time of the day were controlled. A common workload score was evaluated from NASA-TLX questionnaires. Mood

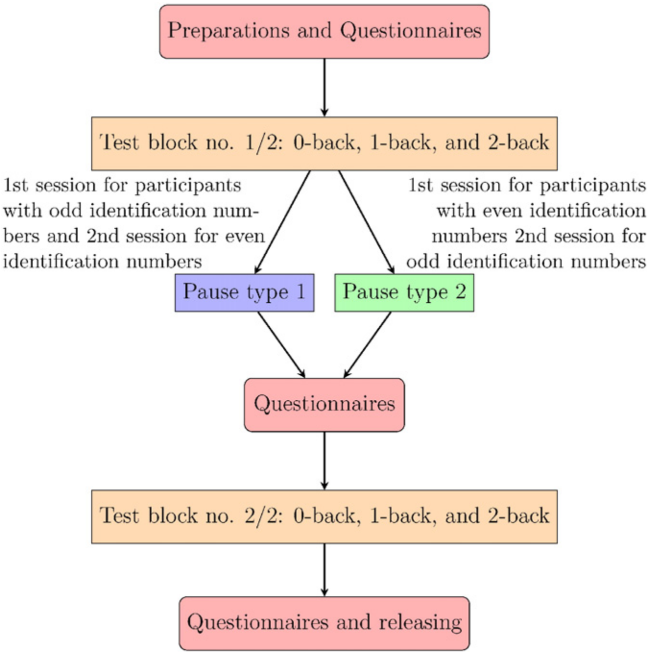

Preparations and Questionnaires
Test block no. 1/2: 0-back, 1-back, and 2-back
1st session for participants with odd identification numbers and 2nd session for even identification numbers
1st session for participants with even identification numbers 2nd session for odd identification numbers
Pause type 1
Pause type 2
Questionnaires
Test block no. 2/2: 0-back, 1-back, and 2-back
Questionnaires and releasing

Fig. 1. Schema of the session.

and vigilance were evaluated by questionnaires (POMS and KSS, respectively). We analyzed the POMS by computing total mood disturbance scores. The sessions were made different to explore the stability in electrophysiological activity under varying environmental and internal states. The pause types were controlled to differ in solely enjoyability: during the pleasant pause (type 1), the participants listened to their favourite music and savored a pleasant snack; during the unpleasant pause (type 2), participants were exposed to a recording of street and construction yard noise and consumed an unpleasant snack. The sound environments were controlled for mean loudness and the snacks for calories. The study was granted ethical approval by the review board of the Hospital District of Helsinki and Uusimaa, Finland. The experiment was carefully explained to the participants, and written consent was obtained before attending the first session. The study protocol followed the Declaration of Helsinki.

2.2. N-back task

We used a basic N-back task with numerical stimuli. The task was presented with Presentation software (Neurobevaioral Systems, Inc., Version 14.9). The stimuli were bright white numbers on a black background. They were presented at the center of the participants visual field occupying ca. 1.7 degree vertical visual angle on a screen in the measurement chamber. The participants monitored a stimulus train of 180 consecutive trials for each memory load level (0-back, 1-back, and 2-back, see below). Prior to presenting the stimulus, a fixation cross in the middle of the screen was presented. The number stimulus was visible for 1500 ms and thereafter the fixation cross for 500 ms, stimulus onset asynchrony thus being 2000 ms. A button press response between 300–1950 ms after stimulus onset was accepted for the analysis. A tenth of the stimuli was distracted with a noise distractor beginning 300–900 ms after the onset of the stimulus presentation. These stimuli were used to distract the participant from the N-back task. Those events when a distractor sound was present are analyzed elsewhere and were discarded from the analysis of performance and all brain measures.

The paradigm had three levels of memory load; in the 0-back condition, participants were looking for a predetermined number, whereas in the 1-back and 2-back the task was to determine whether the stimulus matched the previous stimulus, or the one before that, respectively. The stimulus trains were predetermined pseudo-random lists. One third of the stimuli corresponded to the task, resulting in 60 match and 120 non-match trials in each of the load levels. The participants took part in two sessions and conducted the task twice per a session, resulting in 1944 (0.9 $*$ 180 trials, 3 levels, 2 blocks, 2 sessions) non-distracted trials for each participant.

2.3. Response design

The participants used their thumbs to respond with an in-house device connected to the measurement system using optic fiber technology. In the N-back paradigm, a forced-choice response between match and non-match was applied. The right thumb was used for matching stimuli and the left for non-matching stimuli. The response was indicated by lifting the corresponding thumb from the hand-held device.

Responses were categorized according to the task load and the stimulus type (match or non-match). Only the correct responses were included in the further analysis. We used the median response time as a behavioral metric. The median was chosen despite a predefined time window for response, since in a task with varying requirements, the median gives the most stable results [47].

2.4. MEG recordings

MEG recordings were carried out in the BioMag laboratory of the Helsinki University Central Hospital with a 306-channel Elekta Neuromag Vector View MEG device placed in a three-layer magnetic shielded room (Euroshield, Eura, Finland). The Elekta Neuromag Vector View is comprised of 204 orthogonal planar gradiometers and 102 magnetometers in a head-shaped helmet. During the recordings, the participants were sitting in a comfortable position and their heads were covered by the MEG sensor array. In addition to the MEG channels, EEG (64 channels) and electro-oculography (EOG), stimulus triggers, and digital timing signals for synchronization were recorded simultaneously into the data file while the participants were performing the N-back task. These signals were used for artefact detection, time synchronization, and noise control. The position of the participant's head with respect to the sensor helmet was determined with help of four head-position-indicator (HPI) coils. Participant's head was positioned similarly in the beginning of each measurement block. Data from all MEG channels were band-pass filtered with 0.1–170 Hz filter, sampled at 500 Hz and stored locally.

2.5. Data analysis

The MEG data was analyzed using Martinos MNE [48], Brainstorm [49], MATLAB (8.3, MathWorks), and R language and environment for statistical computing [50]. Preprocessing was conducted with MNE and Brainstorm. For the statistical analysis, the data was exported to R environment.

First, the Martinos MNE software was used to filter with a 1–20 Hz band-pass filter, which is a typical filter for cognitive MEG studies. The effects of filtering on signal to noise ratio (SNR) are assessed in Appendix A. Thereafter, eye blink artefacts were attenuated in Brainstorm with signal space projection (SSP) by visually inspecting and removing the corresponding SSP component. The data were then epoched according to stimulus and response triggers. The epochs started 150 ms before and continued 1000 ms after the onset of the stimulus. Pre-stimulus interval was used for determining the baseline. In addition, epochs with signal amplitudes (peak-to-peak) exceeding 3000 fT or fT/m were discarded.

Global field power (GFP) for each preprocessed trial epoch was subsequently computed in MATLAB, as defined by [51, 52]. GFP was examined in a space including all MEG sensors as well as in three separate sub-spaces, named right lateral frontal (RLF), left lateral frontal (LLF), and occipital, denoting partial selections of included sensors on the frontal hemispheres and the occipital lobe. GFP of all epochs were subsequently exported to R statistical software. In R, the GFP time epochs were baseline corrected and averaged for ERF feature extraction. Instead of one average we computed sample of bootstrapped averages to illustrate also the uncertainty of individual ERF average. This way we were able to observe the individual differences over different measurements. The ERF components extracted for test-retest analysis were an early peak, the M170, and a late modulation in signal (peaking at 600–900 ms).

The M170 peaks were determined using local polynomial regression fitting (loess) in an automated algorithm. The method reduces the noise-derived variation in the signals [53] and allows an automatic peak detection. The fitting was applied to a signal average of each task load (0-back, 1-back, and 2-back) and response (match and non-match) combination separately. The parameters for the fitting algorithm were adjusted to result in an R^2 fit of 0.9 for every signal. The peaks for M170 amplitude and latency were determined as being the next peak after 100 ms and before 250 ms post-stimulus by a simple algorithm searching for locally highest values on the slopes of fitted signal. The LPP was defined as a signal amplitude average between 600 and 900 ms post-stimulus. Three key features, the M170 peak amplitudes and latencies and the LPPs mean amplitude, were subjected to the statistical analysis.

2.6. Statistical analysis

The behavioral results were analyzed according to the response times in the three task loads and the four measurement blocks for learning effects using general linear models (GLMs) and ANOVA. The three extracted ERF features (M170 amplitude, latency, and LPP mean amplitude) were analyzed using GLMs to examine differences between participants ('subject', 7 levels) and the measurements ('block' & 'session', 4 levels) within participants.

The measurements were compared in a pair-wise manner to explore the main effects (e.g., for learning) between measurements within the sessions ('block'), between latter measurements of each sessions, i.e., after different pause types and the measurements between the session days. These three independent variables were used to test the variation differences in all the extracted ERF features. Participant and task load were used as the parameters in our statistical models to test the interactions.

For between participant consistency, intraclass correlation coefficients (ICC) (see [54] for details) were computed for each task load (0-, 1-, and 2-back), response (match, non-match), and measurement ('block' & 'session'). ICC was calculated as defined by [55]. When using reliability analysis such as ICC instead of simple correlation coefficient the difference in mean of the ERF features between participants is utilized in the analysis. It constitutes more rigorous analysis of test-retest reliability than the zero order correlation coefficient.

To confirm the task dependence of the extracted electrophysiological features, we analyzed the effect of response time (RT) on the ERF features by computing the regression for each task load-response-subject mean against median response times in the N-back task. We also examined the M170 and LPP differences between the slow performance and fast performance participants (RT limit 500 ms in 2-back condition) within each task load to further verify that the used features are task related.

3. Results

3.1. Questionnaires

The questionnaire data did not indicate a significant effect of environmental conditions on mood or vigilance. The KSS data showed that vigilance was stable within participants within the visits (χ^2-test for within session ratings). The subjective stress induced by the tasks (NASA-TLX) did not show trends within sessions nor did it express correlation with the pause type (χ^2-test between sessions). Similarly, according to ANOVA the total mood disturbance was stable across both sessions and did not show variation in questionnaires filled after different types of pauses. ANOVA showed more variation in the mood across the two counter-balancing groups ($F = 4.01, p = 0.06$) than across the pause type. The workload scores did not vary significantly between session either. In sum, changes in the environmental factors, i.e., the pause type did not affect performance, mood state, or alertness.

3.2. Behavioral

We found that for lower task load levels over 90% of the responses were correct and were thus qualified for further analysis. ANOVA ($F = 4.153, p < 0.02$) demonstrated significant differences in accuracy between load levels but multiple comparisons (Tukey's range test) revealed that accuracy varied significantly only between 2-back and 0-back conditions. When also the effect of participant is taken into account the difference between 2-back and 1-back conditions became significant. The behavioral data showed significant differences in response times between the three task load levels, i.e., in 0-back versus 1-back and 2-back task loads (ANOVA, $F = 28.99, p < 0.001$). This was true also for each task load pair as revealed by multiple comparisons test. Repeated measures ANOVA showed no difference between the performance of the first test session and the second session, $F = 1.12, ns.$ (see Table 1). This indicates no significant improvement in performance and thus no learning effect. There was also no statistically significant difference between the latter test blocks of the sessions and the different type of pause (ANOVA, $F = 0.525, ns.$). The ERF features showed similar disordinal variations in pairwise comparisons.

3.3. MEG data

After artefact rejection, 94.6% of the correct trials were included to the subsequent analyses. We adjusted the fitting parameters of the algorithm reported in Section 2.5 to obtain 0.9 or higher values for multiple R^2 tests for the fitting curves. The residuals of the fitted polynomials were normally distributed. Some examples of the ERFs with the estimated M170 peaks and smoothing curves are shown in Fig. 2, which also illustrates the variation in the GFP averages within-subject for one task load, in single block, and one stimulus type.

3.3.1. ANOVA and ICC results

Our analyses with ANOVA resulted (Tables 2, 3, 4) significant differences in all parameters for some of the sensor selections when investigating the pairwise main effects. The measurements were divided into pairs according to different pause types, sessions, and measurement blocks within a session.

Overall, occipital sensor selection demonstrated more significant variations across the measurements than the other sensor selections. For all the analyzed features, least variation was found in M170 latency

Table 1
Mean of the response time medians (sd) in measurements over all participants. In milliseconds.

Task load	1st session		2nd session	
	1st block	2nd block	1st block	2nd block
0-Back	426(86)	430(91)	431(122)	430(97)
1-Back	486(98)	451(95)	503(95)	487(112)
2-Back	594(156)	531(134)	640(209)	564(180)
Diff. 2-, 0-back	173(99)	111(83)	216(132)	138(122)

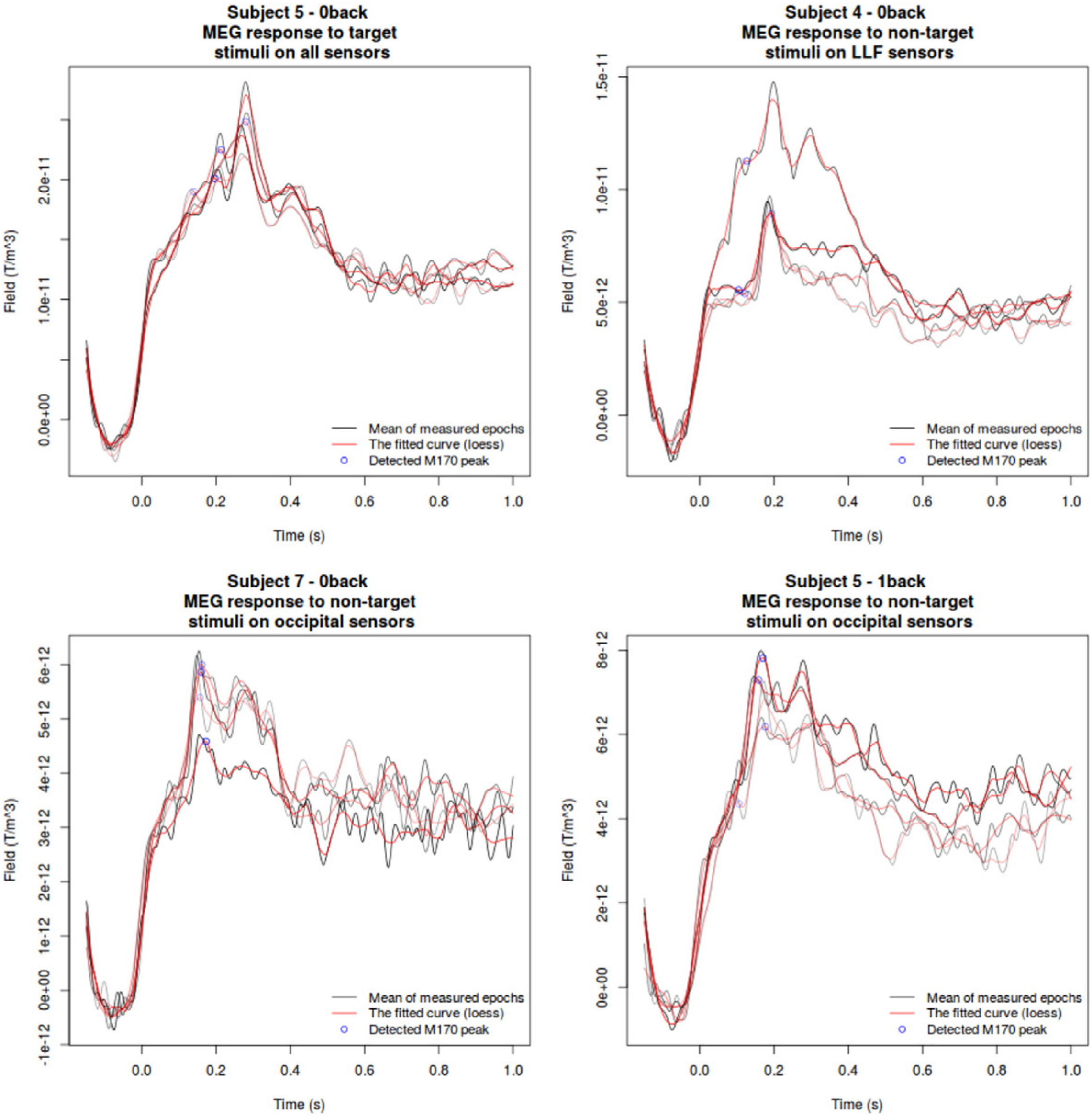


Fig. 2. Examples of fitting on averaged GFP epochs by subject, cognitive load, and response (target/non-target). Each signal illustrate the mean for a single block, the local polynomial regression fitting (loess), and the M170 peak defined by the automatic algorithm. Individual block means are overlaid in the figures.

and most in M170 amplitude. The ICC analysis on the other hand showed significant reliability for the M170 amplitude (1-back and 2-back $ICCs>0.44, ps<0.001$) and excellent reliability for LPP (1-back and 2-back $ICCs>0.83, ps<0.001$) in the conditions with higher cognitive load when all the sensors were included into the analysis. The full ICC results can be found in Appendix A.

Table 2
GLM main effect of session number.

Sensor block	Statistic	M170 ampl.	M170 latency	LPP ampl.
All sensors	Df	1	1	1
	F-value	0.0214	0.0003	5.3006
	p-Value	*ns.*	*ns.*	0.02282
LLF sensors	Df	1	1	1
	F-value	2.0281	09,638	0.0530
	p-value	*ns.*	*ns.*	*ns.*
RLF sensors	Df	1	1	1
	F-value	5.9904	0.2269	0.8976
	p-Value	0.01565	*ns.*	*ns.*
Occipital sensors	Df	1	1	1
	F-value	2.8876	2.9762	0.1284
	p-Value	0.09153	0.08675	*ns.*

Table 3
GLM main effect of pause type.

Sensor block	Statistic	M170 ampl.	M170 latency	LPP ampl.
All sensors	Df	1	1	1
	F-value	0.1031	2.4706	0.0047
	p-Value	*ns.*	*ns.*	*ns.*
LLF sensors	Df	1	1	1
	F-value	1.9158	0.1064	4.7062
	p-Value	*ns.*	*ns.*	0.03178
RLF sensors	Df	1	1	1
	F-value	0.0029	0.0057	0.479
	p-Value	*ns.*	*ns.*	*ns.*
Occipital sensors	Df	1	1	1
	F-value	5.0772	3.9286	0.0258
	p-Value	0.02583	0.04947	*ns.*

Table 4
GLM main effect between blocks (before and after pause).

Sensor block	Statistic	M170 ampl.	M170 latency	LPP ampl.
All sensors	Df	1	1	1
	F-value	8.6184	0.0465	0.2552
	p-Value	0.003904	*ns.*	*ns.*
LLF sensors	Df	1	1	1
	F-value	0.3480	0.941	0.0816
	p-Value	*ns.*	*ns.*	*ns.*
RLF sensors	Df	1	1	1
	F-value	2.0149	1.1284	1.1904
	p-Value	*ns.*	*ns.*	*ns.*
Occipital sensors	Df	1	1	1
	F-value	3.1796	0.3294	0.3911
	p-Value	0.07678	*ns.*	*ns.*

Table 6
GLM interaction effects of subject and session type.

Sensor block	Statistic	M170 ampl.	M170 latency	LPP ampl.
All sensors	Df	6	6	6
	F-value	3.6499	1.3226	2.7212
	p-Value	00,039	*ns.*	0.0225
LLF sensors	Df	6	6	6
	F-value	2.2002	1.1337	3.5082
	p-Value	0.0580	*ns.*	0.0052
RLF sensors	Df	6	6	6
	F-value	4.7178	0.5957	2.5700
	p-Value	0.0005	*ns.*	0.0297
Occipital sensors	Df	6	6	6
	F-value	8.8782	3.2789	1.9644
	p-Value	0.0001	0.0080	0.0880

According to the ANOVA, the within-session variation (between blocks) showed no significant effects for M170 latency or the LPP. M170 amplitude instead varied significantly even within-session ($F=8.61, p=0.004$). The counter-balanced between day variation (Table 3) demonstrates significant main effects for all the features in some sensor selections (M170 amplitude in occipital sensors $F=5.08, p=0.03$, M170 latency in occipital sensors $F=3.93, p=0.05$, and mean LPP amplitude in LLF sensors $F=4.71, p=0.03$).

Tables 5, 6, and 7 show that the interaction effect of the subject and the block pairs correspond to the main effect evaluation. Again the LPP and M170 amplitude varied more drastically than the M170 latency. The M170 latency showed an interaction effect only in the occipital sensor selection. The interaction effects between block pairs and subjects were, however, disordinal (Fig. 3. Whereas, within session effects for M170 latency and mean LPP were purely subject derived. Due to the disordinal interaction, the effects showed in Tables 5, 6, and 7 cannot be referred as main effects for either measurement block or subject.

The within subject variation was much smaller than between-subject variation (6.5 vs. 11.7% for the M170 latency, 0.3 vs. 4% for the M170 amplitude, and 0.5 vs. 2% for LPP) as expected.

The within subject consistency was thereby more prominent in amplitude based measures (within variation 0.25 of total variation) than in latency (within variations 0.5 of total variation).

The ICC analysis expressed increasing trend for correlation with higher cognitive load. This was evident across the sensor selections and ERF features. The ICC was most correlated for 1-back task in LPP in the all sensor space ($ICC=0.87, p<0.001$). In the conditions 1-back and 2-back, the ICC ranked ERF parameters as follows, LPP was most correlated ($ICC_{\min}=0.42, p<0.001$), M170 amplitude second ($ICC_{\min}=0.18, p=0.02$), and M170 latency least ($ICC_{\min}<0$, ns.).

3.4. Correlation between MEG and behavioral results

To confirm the dependence of the behavioral data on the ERF features we examined if differences between participants were due to differences in strategy, cognitive abilities, or internal state, and if they appear in the ERF features. The data provides evidence that ranking the participants according to performance is reflected in all of the ERF features. In Fig. 4, the regression (latency change is >10 ms per response second ($p>0.95$)) suggests that the ERF features are affected by the response times. This is evident when all the sensors are included in the computation of GFP, as well as if we only analyze RLF sensors.

Moreover, we divided the participants into two groups, namely fast and slow performers, to see if the speed in given responses to the task is shown in the ERF features. We wished to test whether the differences in the parameters according to response times are due to biological or strategic distinctions across the participants. We found significant differences between participants with median response time above 500 ms and participants whose median response times were faster than 500 ms in the 2-back task. The faster responders had a significantly higher mean LPP amplitude (*t*-test, $t=3.0, df=26, p=0.005$ and $t=2.8, df=27, p<0.009$, respectively) for the 2-back and the 1-back task loads, but not for the 0-back task load (shown in Fig. 5).

4. Discussion

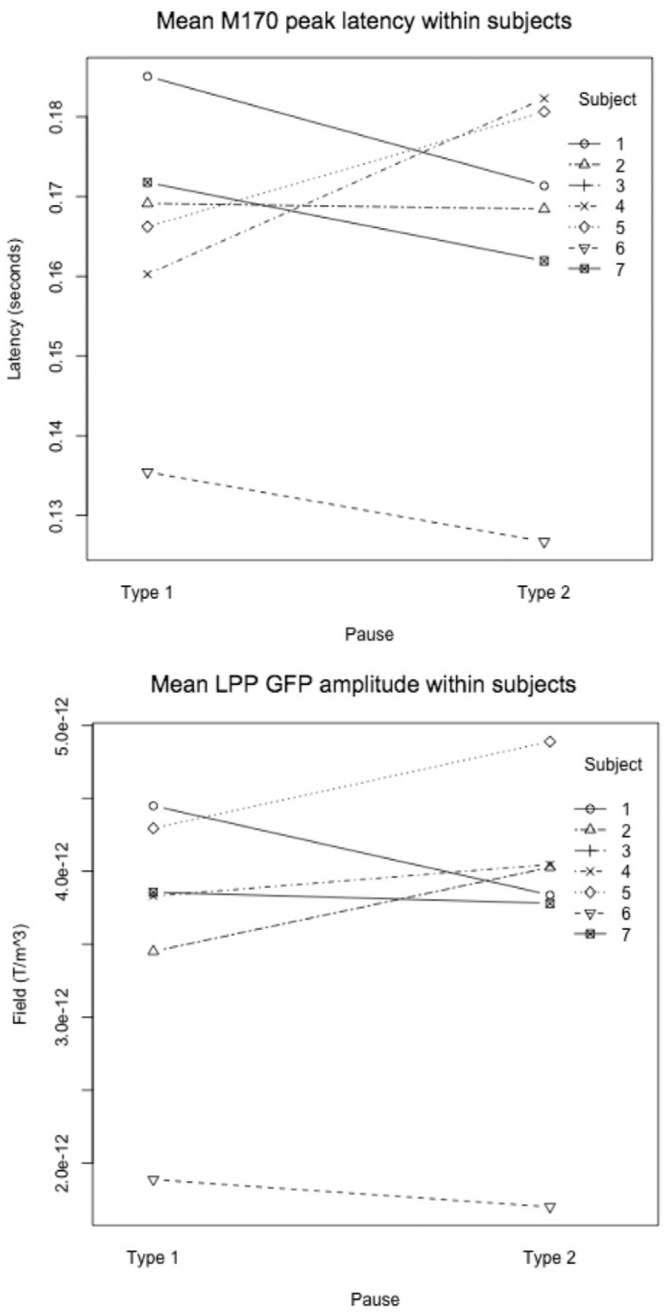
The primary objective of this study was to examine the replicability and test-retest reliability of evoked field components with associations to cognitive functions such as working memory in MEG. GFP was used as a measure for overall event-related activation in

Table 5
GLM Interaction effects of subject and session number.

Sensor block	Statistic	M170 ampl.	M170 latency	LPP ampl.
All sensors	Df	6	6	6
	F-value	3.6678	1.8225	1.6374
	p-Value	0.0038	*ns.*	*ns.*
LLF sensors	Df	6	6	6
	F-value	2.1771	0.9626	4.5237
	p-Value	0.0604	*ns.*	0.0007
RLF sensors	Df	6	6	6
	F-value	3.4148	0.5522	2.4821
	p-Value	0.0062	*ns.*	0.0349
Occipital sensors	Df	6	6	6
	F-value	9.4304	3.4809	1.9432
	p-Value	0.0001	0.0055	0.0914

Table 7
GLM interaction effects of subject and block number.

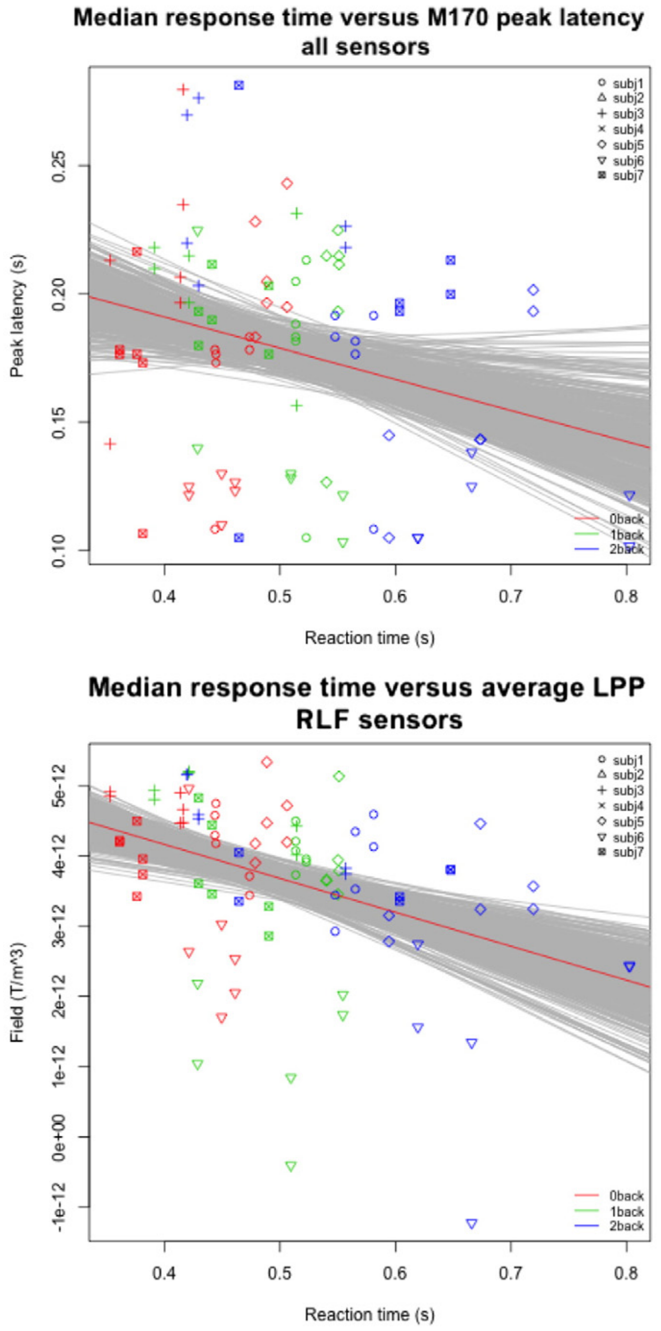
Sensor block	Statistic	M170 ampl.	M170 latency	LPP ampl.
All sensors	Df	6	6	6
	F-value	1.8973	1.0697	1.1090
	p-Value	0.0990	*ns.*	*ns.*
LLF sensors	Df	6	6	6
	F-value	4.1128	1.7571	1.4041
	p-Value	0.0017	*ns.*	*ns.*
RLF sensors	Df	6	6	6
	F-value	0.4805	0.5141	1.2651
	p-Value	*ns.*	*ns.*	*ns.*
Occipital sensors	Df	6	6	6
	F-value	2.0722	0.6249	0.1459
	p-Value	0.0728	*ns.*	*ns.*



Mean M170 peak latency within subjects

Mean LPP GFP amplitude within subjects

Fig. 3. Interaction plot between sessions with different pause types and within participant changes in M170 peak latency and mean LPP amplitude in the right frontal sensor block.



Median response time versus M170 peak latency all sensors

Median response time versus average LPP RLF sensors

Fig. 4. Above: Effect of response time on M170 latency over all sensors. Below: Effect of response time on mean LPP amplitude in RLF sensor block. Red lines depict the regression of the measured data and gray lines are regressions for bootstrapped datasets used for computing the confidence intervals.

the cerebral cortices, and we found reliable ERF features showing little test-retest variation across multiple measurements. We also found that with higher cognitive load the ERF features express more intraclass correlation between participants. The variation in M170 latency was least significant in the frontal regions as the test-retest correlation in LPP was most prominent in the occipital area.

In most of the research paradigms only one recording is available from each individual. This yields a great challenge to the estimation of the contributing sources of intra-individual variability. To study the dependencies of the intra-individual differences to other factors, such as genetic or behavioral measures, it is of high importance to study the reliability of brain responses in a paradigm with multiple measures per participant, preferably in different moods and across days [56].

Importantly, our results are in line with earlier studies on stability of electrophysiological features [24, 14, 57], by demonstrating that intra-individual stability is high compared to inter-individual variation. Furthermore, our results also showed greater variability in intra-individual ERF parameters with longer delay between the measurements. Interestingly this is especially salient for the task load dependent feature of the M170 latency.

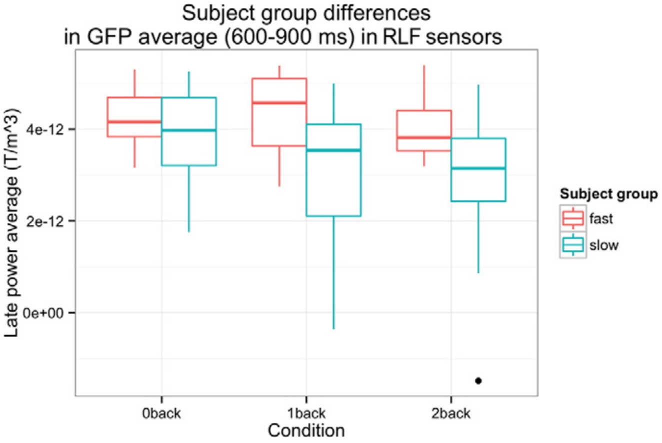


Fig. 5. The effect of group to mean LPP power. Participants divided into groups according to the median response time in 2-back task (fast < 500 ms < slow).

In addition, we demonstrated that the selection of the time-locked ERF metrics needs to be done carefully. For instance the major fluctuations in occipital area between measurements with M170 peak latency might appear due to problems in finding proximate peaks in ERFs. In the partial sensor selections the method for isolating the peak to represent M170 might be prone to alternations of head position between measurements and equipment derived noise. However whole-head GFP is virtually independent of head positioning and hence more stable parameter than, e.g., dipole models. Despite the variations in some of the sensor selection results, latency is a rather stable feature and does not vary significantly between measurements. Some of the EEG studies on reliability suggest that for, e.g., N170 type ERPs show adequate to excellent test/reteset reliability for amplitude [24]. We found that M170 amplitude correlates over subjects and over the measurements although, it is fragile to variation between individual measurements. This finding advocates that M170 is quite a precise parameter over our test population. In addition the LPP correlated greatly between the participants and had less measure to measure variation than the M170 amplitude.

The stability differences between the extracted ERF features suggest temporal structures in noise distributions. The interesting signals in MEG are several orders of magnitude smaller than environmental noise [3], and environmental noise is not completely stable; it changes over time [58]. These changes might cause erroneous interpretations of MEG signals even on averaged ERFs. In our results, fluctuation is shown as disordinal variation in the interaction effects in between-day measures and between participants. Random variation trends suggests that this originates from the state of the measurement equipment rather than a change in the participants electrophysiological signals. This can also be assumed by looking into stability differences between areas. The areas related to working memory and task execution, as shown by, e.g., [59], display more stability between measurements. The occipital fluctuations in pair-wise measurement comparisons may derive from weaker peak detection since the overall stability is better for the peak independent feature, i.e., LPP. In addition, the constant mood and vigilance states within different environmental factors and over different visits may have contributed to the signal stability. Literature suggests that emotional content may affect to the ERP components widely [60]. Stress also alters the electrophysiology in cognitively demanding tasks in variety of ERP components [61].

The right frontal sub-division of sensors (RLF), which showed the least variation in pair-wise comparisons, is proximate to the cortical areas reported as highly relevant for the N-back task [27]. This may have affected to the ICC results here since the target and non-target stimuli might elicit differing responses on this area [62]. These differences may vary between participants. In addition, our behavioral results combined with MEG outcome features imply that the RLF sensors incorporate the most variation due to response time and current task load. This is supported by findings that the right-hemispheric frontal lobe is highly involved in task-related processing [63]. Earlier studies have also shown that the observed lateralization is related to the use of verbal (alphanumerical) stimuli [27]. In general, the partial sensor selection results suggest that task relevant ERF latencies extracted from smaller areas show less variability than signal amplitudes and whole sensor array statistics.

The arguments stated above imply that task-related MEG/ERF analysis are to be selected with caution when planning a multi-session study. It also seems sensible to use a task in the experimental paradigm to assure that participants retain a similar time-dependent cognitive state in both sessions. This is also shown in earlier studies [14, 57]. Moreover, the engagement of the participants should also be controlled, e.g., by gamification.

Prior research on regional connectivity suggests that improved performance, i.e., learning, might relate to the emergence of more reliable brain network configurations [14]. The performance differences reflected in the ERF features might also be due to effort-related vigilance. However, further study is needed to examine the cognition and person-related factors behind the group difference in the N-back task. We found a persistent decrease among slower responders in LPP, i.e., for the participants whose behavioral response was more proximate to the analyzed time segment in the ERF. This suggests that the effect is not directly response derived, but rather a delayed positivity or negativity related to performance.

Using more advanced analysis methods would reveal additional prospects for signal replicability. Comparing replicability between raw sensor signals and source space models would provide valuable information on the effect of advanced analysis techniques on equipment derived noise. In the future, specific ERF source space parameters should be compared to, e.g., GFP in order to evaluate the signal stability during different steps in the analysis. Also, our sample size is modest and including, e.g., patient groups would reveal more about intra-subject stability of ERF components.

5. Conclusion

We demonstrated the stability of task-dependent brain responses in a cognitively demanding MEG experiment. We propose the features of task related ERFs as appropriate measures of cognitive brain functions in longitudinal designs, such as cross-over studies or imaging genetic studies. The findings are in line and well inserted in the already existing literature on test-retest reliability in EEG [24, 25]. The literature suggests that ERP latencies show reliability in components such as N170. Therefore these features qualify as an adequate measures for endophenotype models, and personality trait research. The late ERF components appear to be most affected to the attention specific parameters of the task, and while the signal to noise ratio (SNR) remains limited, i.e., the number of averaged epochs is high, it is demonstrated to be potential measure for cognitive activity as well [24]. The cognitive load should be controlled in the paradigms. Our results reveal the potential in clinical applications and applications utilizing brain derived variables for automated electrophysiological analysis and computationally extracted parameters, even in sensor signals. However, the generalization of the results must be investigated with higher participant count and with clinical populations in further studies.

Acknowledgements

The authors declare that there is no conflict of interest regarding the publication of this paper. The study was supported by the SalWe Research Program for Mind and Body (Tekes - the Finnish Funding Agency for Technology and Innovation grant 1104/10) and Danish National Research Foundation (project number 117). It was conducted in accordance with the Declaration of Helsinki (1964). We would like to thank the MEG team at McConnell Brain Imaging Centre, in McGill University, Montreal for help during the preparation of the study and with methodological problems, and Finnish Institute of Occupational Health personnel for guidance and help in latter phases of the project.

Appendix A. Filtering

Here's the ICC analysis results in a table and some extra data on sensor selection and filtering results.

The sensor numbers for different sensor selections used in the article,

Due to the arbitrary (literature based) cutoff frequencies in our data analysis, we assessed the effect of filtering on replicability. We compared the data of the original sampling filter (0.1 Hz high-pass) to that of used (1 Hz high-pass) filter, and evaluated the signal to noise ratio (SNR).

The high pass filtering of data reduced the noise significantly. For example, for mean LPP amplitude the noise levels drop to one half in occipital sensors and below one tenth in the sensors on the left prefrontal cortex. Fig. A.6 illustrates the reduction of the noise in a few examples for single participant in single stimulus type in one task load in one of the sensor blocks.

The channel numbers for the used channel sub-spaces:

- right lateral frontal (RLF): [76:81,109:126,138:150]
- left lateral frontal (LLF): [1:6,9:24,31:48]
- occipital: [187:188,193:198,214:220,235:246]

The full ICC results are found in Table A.8.

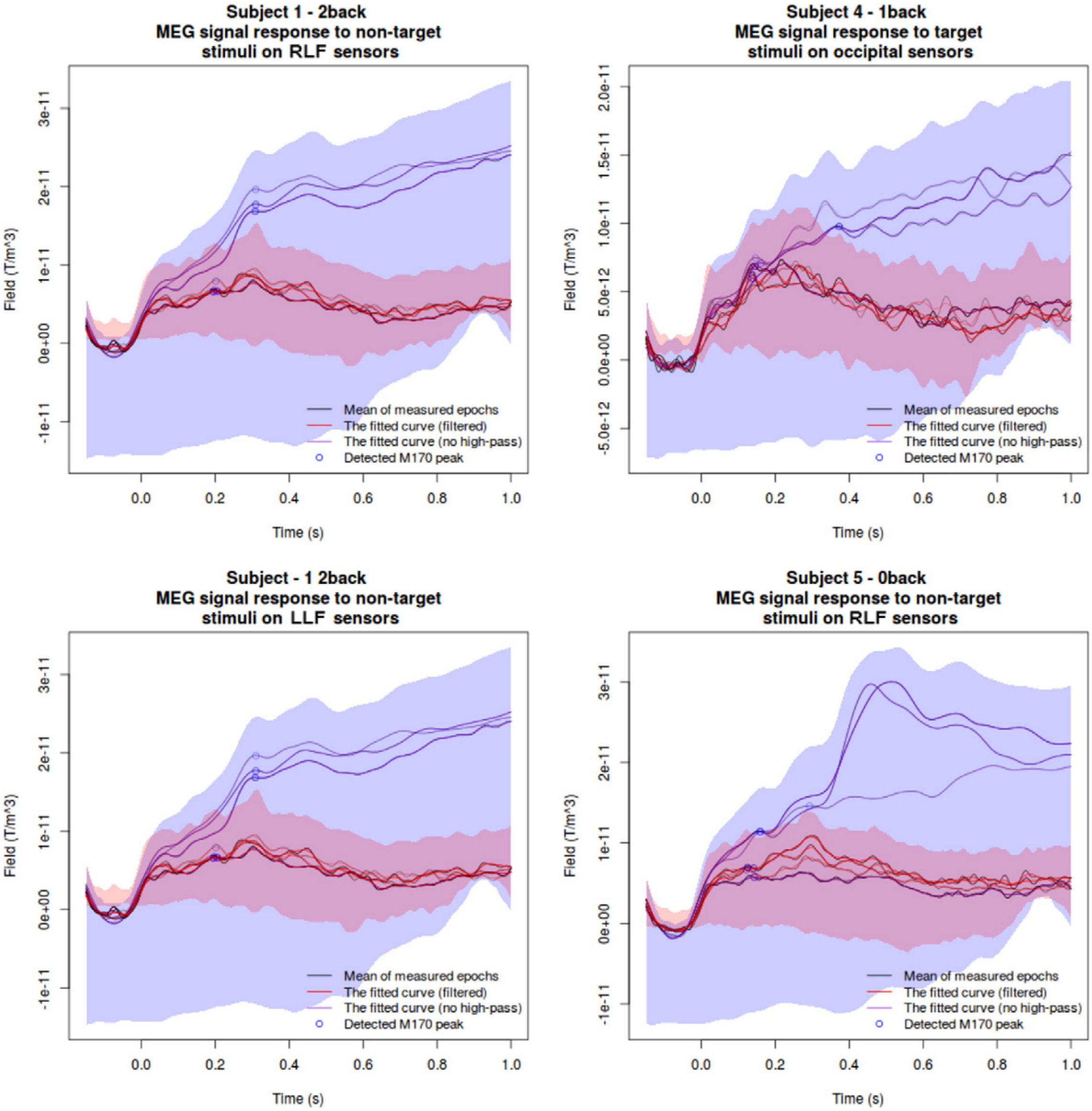


Fig. A.6. Examples of the effect of high-pass filtering on different participants, task difficulties, and responses. The blue shaded area represents the deviation in the data with original 0.1 Hz high-pass sampling filter, and the red shaded section the deviation in the 1 Hz high-pass filtered signals.

Table A.8
Intraclass correlation coefficient analysis for all sensor blocks and task loads within each ERF feature.

Sensor Block		M170 ampl.	*F*-value (*p*-value)	M170 lat.	*F*-value (*p*-value)	LPP ampl.	*F*-value (*p*-value)
All sensors	0-Back	0.133	2.2377 ($p=0.055$)	0.152	2.463 ($p=0.037$)	−0.039	0.695 ($p=0.236$)
	1-Back	0.578	10.738 ($p<0.001$)	0.296	4.352 ($p=0.002$)	0.869*	53.879 ($p<0.001$)
	2-Back	0.445	7.667 ($p<0.001$)	0.319	4.493 ($p=0.001$)	0.833*	39.461 ($p<0.001$)
LLF sensors	0-Back	0.084	1.719 ($p=0.13$)	−0.089	0.366 ($p=0.89$)	−0.131	0.067 ($p=0.99$)
	1-Back	0.175	2.786 ($p=0.02$)	−0.055	0.540 ($p=0.74$)	0.799*	31.670 ($p<0.001$)
	2-Back	0.487	9.013 ($p<0.001$)	0.107	2.011 ($p=0.08$)	0.835*	39.128 ($p<0.001$)
RLF sensors	0-Back	0.010	1.082 ($p=0.38$)	−0.059	0.508 ($p=0.79$)	0.008	1.062 ($p=0.40$)
	1-Back	0.241	3.429 ($p=0.01$)	0.045	1.337 ($p=0.27$)	0.774*	27.039 ($p<0.001$)
	2-Back	0.178	2.850 ($p=0.02$)	0.220	3.268 ($p=0.009$)	0.417	6.550 ($p<0.001$)
Occipital sensors	0-Back	0.469	8.183 ($p<0.001$)	0.403	6.323 ($p<0.001$)	0.058	1.500 ($p=0.19$)
	1-Back	0.783*	31.751 ($p<0.001$)	0.425	6.362 ($p<0.001$)	0.866*	58.555 ($p<0.001$)
	2-Back	0.661	13.214 ($p<0.001$)	0.335	5.950 ($p<0.001$)	0.768*	24.828 ($p<0.001$)

* Higher than 0.7 correlation.

References

[1] M. Gazzaniga, R. Ivry, G. Mangun, Cognitive Neuroscience: The Biology of the Mind, Norton, 2009 (ISBN 9780393927955, http://books.google.ca/books?id=9uB_PwAACAAJ).

[2] R. Hari, R. Salmelin, Magnetoencephalography: From SQUIDs to Neuroscience, Neuroimage 20th Anniversary Special Edition, Neuroimage, 61 (2)2012 386–396, http://dx.doi.org/10.1016/j.neuroimage.2011.11.074.

[3] M. Hämäläinen, R. Hari, R.J. Ilmoniemi, J. Knuutila, O.V. Lounasmaa, Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain, Rev. Mod. Phys. 65 (2) (1993) 413–497, http://dx.doi.org/10.1103/RevModPhys.65.413.

[4] V. Sakkalis, Applied strategies towards EEG/MEG biomarker identification in clinical and cognitive research, Biomark. Med 5 (1) (2011) 93–105, http://dx.doi.org/10.2217/bmm.10.121.

[5] S. Giaquinto, Evoked potentials in rehabilitation. A review, Funct. Neurol. 19 (4) (2004) 219–225.

[6] E.L. Hall, S.E. Robson, P.G. Morris, M.J. Brookes, The relationship between MEG and fMRI, Neuroimage (2014), http://dx.doi.org/10.1016/j.neuroimage.2013.11.005.

[7] I.I. Gottesman, T.D. Gould, The endophenotype concept in psychiatry: etymology and strategic intentions, Am. J. Psychiatry 160 (4) (2003) 636–645, http://dx.doi.org/10.1176/appi.ajp.160.4.636.

[8] G.A.M. Blokland, K.L. McMahon, J. Hoffman, G. Zhu, M. Meredith, N.G. Martin, P.M. Thompson, G.I. de Zubicaray, M.J. Wright, Quantifying the heritability of task-related brain activation and performance during the N-back working memory task: a twin fMRI study, Biol. Psychol. 79 (1) (2008) 70–79, http://dx.doi.org/10.1016/j.biopsycho.2008.03.006.

[9] R. Ntnen, R.J. Ilmoniemi, K. Alho, Magnetoencephalography in studies of human cognitive brain function, Trends Neurosci. 17 (9) (1994) 389–395, http://dx.doi.org/10.1016/0166-2236(94)90048-5 (http://www.sciencedirect.com/science/article/pii/0166223694900485, ISSN issn0166-2236).

[10] H. Renvall, E. Salmela, M. Vihla, M. Illman, E. Leinonen, J. Kere, R. Salmelin, Genome-wide linkage analysis of human auditory cortical activation suggests distinct loci on chromosomes 2, 3, and 8, J. Neurosci. 32 (42) (2012) 14511–14518, http://dx.doi.org/10.1523/JNEUROSCI.1483-12.2012.

[11] Y. Agam, M. Vangel, J.L. Roffman, P.J. Gallagher, J. Chaponis, S. Haddad, D.C. Goff, J.L. Greenberg, S. Wilhelm, J.W. Smoller, D.S. Manoach, Dissociable genetic contributions to error processing: a multimodal neuroimaging study, PLoS One 9 (7) (2014) e101784, http://dx.doi.org/10.1371/journal.pone.0101784.

[12] R. Hari, S. Levänen, T. Raij, Timing of human cortical functions during cognition: role of MEG, Trends Cogn. Sci. 4 (12) (2000) 455–462.

[13] A. Sorrentino, L. Parkkonen, M. Piana, A.M. Massone, L. Narici, S. Carozzo, M. Riani, W.G. Sannita, Modulation of brain and behavioural responses to cognitive visual stimuli with varying signal-to-noise ratios, Clin. Neurophysiol. 117 (5) (2006) 1098-105, http://dx.doi.org/10.1016/j.clinph.2006.01.011.

[14] L. Deuker, E.T. Bullmore, M. Smith, S. Christensen, P.J. Nathan, B. Rockstroh, D.S. Bassett, Reproducibility of graph metrics of human brain functional networks, NeuroImage 47 (4) (2009) 1460–1468, http://dx.doi.org/10.1016/j.neuroimage.2009.05.035.

[15] E. Pekkonen, T. Rinne, R. Näätänen, Variability and replicability of the mismatch negativity, Electroencephalogr. Clin. Neurophysiol. 96 (6) (1995) 546–554.

[16] A.H. Lang, O. Eerola, P. Korpilahti, I. Holopainen, S. Salo, O. Aaltonen, Practical issues in the clinical application of mismatch negativity, Ear Hear. 16 (1) (1995) 118–130.

[17] C. Escera, E. Yago, M.D. Polo, C. Grau, The individual replicability of mismatch negativity at short and long inter-stimulus intervals, Clin. Neurophysiol. 111 (3) (2000) 546–551.

[18] D.A. Sklare, G.E. Lynn, Latency of the P3 event-related potential: normative aspects and within-subject variability, Electroencephalogr. Clin. Neurophysiol. 59 (5) (1984) 420–424.

[19] K.B. Walhovd, A.M. Fjell, One-year test-retest reliability of auditory ERPs in young and old adults, Int. J. Psychophysiol. 46 (1) (2002) 29–40.

[20] D.M. Olvet, G. Hajcak, Reliability of error-related brain activity, Brain Res. 1284 (2009) 89–99, http://dx.doi.org/10.1016/j.brainres.2009.05.079.

[21] M.J. Larson, S.A. Baldwin, D.A. Good, J.E. Fair, Temporal stability of the error-related negativity (ERN) and post-error positivity (Pe): the role of number of trials, Psychophysiology 47 (6) (2010) 1167–1171, http://dx.doi.org/10.1111/j.1469-8986.2010.01022.x.

[22] J. Virtanen, J. Ahveninen, R.J. Ilmoniemi, R. Näätänen, E. Pekkonen, Replicability of MEG and EEG measures of the auditory N1/N1m-response, Electroencephalogr. Clin. Neurophysiol. 108 (3) (1998) 291–298.

[23] M. Schaefer, N. Nöennig, A. Karl, H.-J. Heinze, M. Rotte, Reproducibility and stability of neuromagnetic source imaging in primary somatosensory cortex, Brain Topogr. 17 (1) (2004) 47–53.

[24] R. Huffmeijer, M.J. Bakermans-Kranenburg, L.R.A. Alink, M.H. van Ijzendoorn, Reliability of event-related potentials: the influence of number of trials and electrodes, Physiol. Behav. 130 (2014) 13–22, http://dx.doi.org/10.1016/j.physbeh.2014.03.008.

[25] S.M. Cassidy, I.H. Robertson, R.G. O'Connell, Retest reliability of event-related potentials: evidence from a variety of paradigms, Psychophysiology 49 (5) (2012) 659–664, http://dx.doi.org/10.1111/j.1469-8986.2011.01349.x.

[26] A. Baddeley, Working Memory, Oxford Psychology SeriesClarendon Press, 1987 (ISBN 9780198521334, URL http://books.google.fi/books?id=ZKWbdv__vRMC).

[27] A.M. Owen, K.M. McMillan, A.R. Laird, E. Bullmore, N-back working memory paradigm: a meta-analysis of normative functional neuroimaging studies, Hum. Brain Mapp. 25 (1) (2005) 46–59, http://dx.doi.org/10.1002/hbm.20131.

[28] A. Meyer-Lindenberg, D.R. Weinberger, Intermediate phenotypes and genetic mechanisms of psychiatric disorders, Nat. Rev. Neurosci. 7 (10) (2006) 818–827, http://dx.doi.org/10.1038/nrn1993.

[29] M.M. Plichta, A.J. Schwarz, O. Grimm, K. Morgen, D. Mier, L. Haddad, A.B.M. Gerdes, C. Sauer, H. Tost, C. Esslinger, P. Colman, F. Wilson, P. Kirsch, A. Meyer-Lindenberg, Test-retest reliability of evoked BOLD signals from a cognitive-emotive fMRI test battery, NeuroImage 60 (3) (2012) 1746–1758, http://dx.doi.org/10.1016/j.neuroimage.2012.01.129.

[30] S.M. Jaeggi, M. Buschkuehl, J. Jonides, W.J. Perrig, Improving fluid intelligence with training on working memory, Proc. Natl. Acad. Sci. U. S. A. 105 (19) (2008) 6829–6833, http://dx.doi.org/10.1073/pnas.0801268105.

[31] H.L. Hamburger, M.A. vd Burgt, Global field power measurement versus classical method in the determination of the latency of evoked potential components, Brain Topogr. 3 (3) (1991) 391–396.

[32] W. Skrandies, Global field power and topographic similarity, Brain Topogr. 3 (1) (1990) 137–141.

[33] L. Anllo-Vento, S.J. Luck, S.A. Hillyard, Spatio-temporal dynamics of attention to color: evidence from human electrophysiology, Hum. Brain Mapp. 6 (4) (1998) 216–238.

[34] J. Liu, A. Harris, N. Kanwisher, Stages of processing in face perception: an MEG study, Nat. Neurosci. 5 (9) (2002) 910–916, http://dx.doi.org/10.1038/nn909.

[35] A. Tarkiainen, P. Helenius, P.C. Hansen, P.L. Cornelissen, R. Salmelin, Dynamics of letter string perception in the human occipitotemporal cortex, Brain 122 (Pt 11) (1999) 2119–2132.

[36] C.-H. Hsu, C.-Y. Lee, A. Marantz, Effects of visual complexity and sublexical information in the occipitotemporal cortex in the reading of Chinese phonograms: a single-trial analysis with MEG, Brain Lang. 117 (1) (2011) 1–11, http://dx.doi.org/10.1016/j.bandl.2010.10.002.

[37] E. Zweig, L. Pylkkänen, A visual M170 effect of morphological complexity, Lang. Cogn. Process. 24 (3) (2009) 412–439, http://dx.doi.org/10.1080/01690960802180420 (URL http://www.tandfonline.com/doi/abs/10.1080/01690960802180420).

[38] J.T. Cacioppo, S.L. Crites, W.L. Gardner, G.G. Bernston, Bioelectrical echoes from evaluative categorizations: I. A late positive brain potential that varies as a function of trait negativity and extremity, J. Pers. Soc. Psychol. 67 (1) (1994) 115–125.

[39] S.L. Crites, J.T. Cacioppo, W.L. Gardner, G.G. Berntson, Bioelectrical echoes from evaluative categorization: II. A late positive brain potential that varies as a function of attitude registration rather than attitude report, J. Pers. Soc. Psychol. 68 (6) (1995) 997–1013.

[40] G. Hajcak, J.P. Dunning, D. Foti, Motivated and controlled attention to emotion: time-course of the late positive potential, Clin. Neurophysiol. 120 (3) (2009)

505–510, http://dx.doi.org/10.1016/j.clinph.2008.11.028 (URL http://www.sciencedirect.com/science/article/pii/S1388245708012728, ISSN 1388-2457).

[41] K.J. Yoder, J. Decety, Spatiotemporal neural dynamics of moral judgment: a high-density {ERP} study, Neuropsychologia 60 (0) (2014) 39–45, http://dx.doi.org/10.1016/j.neuropsychologia.2014.05.022 (URL http://www.sciencedirect.com/science/article/pii/S0028393214001705, ISSN issn0028-3932).

[42] T.W. Picton, D.T. Stuss, The component structure of the human event-related potentials, Prog. Brain Res. 54 (1980) 17–48, http://dx.doi.org/10.1016/S0079-6123(08)61604-0.

[43] D.S. Ruchkin, R. Johnson Jr., D. Mahaffey, S. Sutton, Toward a functional categorization of slow waves, Psychophysiology 25 (3) (1988) 339–353.

[44] S.G. Hart, L.E. Staveland, Development of NASA-TLX (Task Load Index): results of empirical and theoretical research, in: P.A. Hancock, N. Meshkati (Eds.),Human Mental Workload, Vol. 52 of Advances in Psychology, North-Holland 1988, pp. 139–183, http://dx.doi.org/10.1016/S0166-4115(08)62386-9 (http://www.sciencedirect.com/science/article/pii/S0166411508623869).

[45] M. Gillberg, G. Kecklund, T. Akerstedt, Relations between performance and subjective ratings of sleepiness during a night awake, Sleep 17 (3) (1994) 236–241.

[46] D.M. McNair, M. Lorr, L.F. Droppleman, Profile of Mood States, University of California, 1971.

[47] R. Ratcliff, Methods for dealing with reaction time outliers, Psychol. Bull. 114 (3) (1993) 510–532.

[48] A. Gramfort, M. Luessi, E. Larson, D.A. Engemann, D. Strohmeier, C. Brodbeck, L. Parkkonen, M.S. Hämäläinen, MNE software for processing MEG and EEG data, NeuroImage 86 (2014) 446–460, http://dx.doi.org/10.1016/j.neuroimage.2013.10.027.

[49] F. Tadel, S. Baillet, J.C. Mosher, D. Pantazis, R.M. Leahy, Brainstorm: a user-friendly application for MEG/EEG analysis, Comput. Intell. Neurosci. 2011 (2011) 879716, http://dx.doi.org/10.1155/2011/879716.

[50] R Core Team, R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Vienna, Austria, 2014 (http://www.R-project.org/).

[51] D. Lehmann, W. Skrandies, Reference-free identification of components of checkerboard-evoked multichannel potential fields, Electroencephalogr. Clin. Neurophysiol. 48 (6) (1980) 609–621.

[52] D. Lehmann, W. Skrandies, Spatial analysis of evoked potentials in man–a review, Prog. Neurobiol. 23 (3) (1984) 227–250.

[53] H. Akaike, A new look at the statistical model identification, IEEE Trans. Autom. Control 19 (6) (1974) 716–723, http://dx.doi.org/10.1109/TAC.1974.1100705 (ISSN issn0018-9286).

[54] G.G. Koch, Intraclass Correlation Coefficient, John Wiley & Sons, Inc., 2004http://dx.doi.org/10.1002/0471667196.ess1275.pub2 (http://dx.doi.org/10.1002/0471667196.ess1275.pub2, ISBN 9780471667193).

[55] R.H. Finn, A note on estimating the reliability of categorical data, Educ. Psychol. Meas. 30 (1) (1970) 70–76, http://dx.doi.org/10.1177/001316447003000106.

[56] D. Nutt, S. Wilson, A. Lingford-Hughes, J. Myers, A. Papadopoulos, S. Muthukumaraswamy, Differences between magnetoencephalographic (MEG) spectral profiles of drugs acting on {GABA} at synaptic and extrasynaptic sites: a study in healthy volunteers, Neuropharmacology 88 (0) (2015) 155–163, http://dx.doi.org/10.1016/j.neuropharm.2014.08.017 (http://www.sciencedirect.com/science/article/pii/S0028390814003001, note{GABAergic} Signaling in Health and Disease, ISSN issn0028-3908).

[57] M. Näpflin, M. Wildi, J. Sarnthein, Test-retest reliability of EEG spectra during a working memory task, NeuroImage 43 (4) (2008) 687–693, http://dx.doi.org/10.1016/j.neuroimage.2008.08.028.

[58] S. Taulu, M. Kajola, J. Simola, Suppression of interference and artifacts by the signal space separation method, Brain Topogr. 16 (4) (2004) 269–275.

[59] I. Burunat, V. Alluri, P. Toiviainen, J. Numminen, E. Brattico, Dynamics of brain activity underlying working memory for music in a naturalistic condition, Cortex 57 (0) (2014) 254–269, http://dx.doi.org/10.1016/j.cortex.2014.04.012 (http://www.sciencedirect.com/science/article/pii/S0010945214001270, ISSN issn0010-9452).

[60] A. Zinchenko, P. Kanske, C. Obermeier, E. Schrger, S.A. Kotz, Emotion and goal-directed behavior: ERP evidence on cognitive and emotional conflict, Soc. Cogn. Affect. Neurosci. 10 (11) (2015) 1577–1587, http://dx.doi.org/10.1093/scan/nsv050 (http://scan.oxfordjournals.org/content/10/11/1577.abstract).

[61] A.J. Shackman, J.S. Maxwell, B.W. McMenamin, L.L. Greischar, R.J. Davidson, Stress potentiates early and attenuates late stages of visual processing, J. Neurosci. 31 (3) (2011) 1156–1161, http://dx.doi.org/10.1523/JNEUROSCI.3384-10.2011 (http://www.jneurosci.org/content/31/3/1156.abstract).

[62] D. Choi, Y. Egashira, J. Takakura, M. Motoi, T. Nishimura, S. Watanuki, Gender difference in N170 elicited under oddball task, J. Physiol. Anthropol. 34 (1) (2015) 7, http://dx.doi.org/10.1186/s40101-015-0045-7 (http://www.jphysiolanthropol.com/content/34/1/7, ISSN issn1880-6805).

[63] M.F. Glabus, B. Horwitz, J.L. Holt, P.D. Kohn, B.K. Gerton, J.H. Callicott, A. Meyer-Lindenberg, K.F. Berman, Interindividual differences in functional interactions among prefrontal, parietal and parahippocampal regions during working memory, Cereb. Cortex 13 (12) (2003) 1352–1361.