Individualized cognitive neuroscience needs 7T: Comparing numerosity maps at 3T and 7T MRI

Yuxuan Cai¹,², Shiri Hofstetter¹, Wietske van der Zwaag¹, Wietske Zuiderbaan¹, Serge O. Dumoulin¹,²,³,*

¹Spinnoza Centre for Neuroimaging, Amsterdam, Netherlands
²Experimental and Applied Psychology, VU University Amsterdam, Amsterdam, Netherlands
³Experimental Psychology, Helmholtz Institute, Utrecht University, Utrecht, Netherlands

**ABSTRACT**

The field of cognitive neuroscience is weighing evidence about whether to move from the current standard field strength of 3 Tesla (3T) to ultra-high field (UHF) of 7T and above. The present study contributes to the evidence by comparing a computational cognitive neuroscience paradigm at 3T and 7T. The goal was to evaluate the practical effects, i.e. model predictive power, of field strength on a numerosity task using accessible pre-processing and analysis tools. Previously, using 7T functional magnetic resonance imaging and biologically-inspired analyses, i.e. population receptive field modelling, we discovered topographical organization of numerosity-selective neural populations in human parietal cortex. Here we show that these topographic maps are also detectable at 3T. However, averaging of many more functional runs was required at 3T to reliably reconstruct numerosity maps. On average, one 7T run had about four times the model predictive power of one 3T run. We believe that this amount of scanning would have made the initial discovery of the numerosity maps on 3T highly infeasible in practice. Therefore, we suggest that the higher signal-to-noise ratio and signal sensitivity of UHF MRI is necessary to build mechanistic models of the organization and function of our cognitive abilities in individual participants.

1. Introduction

Cognitive neuroimaging studies typically require fast whole brain image acquisitions with high signal-to-noise ratio (SNR) and maximal sensitivity to small blood oxygenation level dependant (BOLD) signal changes for reliable detection. This is especially the case for computational neuroimaging where we go beyond the detection of activation to build computational models of neural function in individual participants (De Martino et al., 2018; Dumoulin and Knappen, 2018; Dumoulin and Wandell, 2008; Kay et al., 2008; Naselaris et al., 2011; Wandell, 1999; Wandell and Winawer, 2015). The use of magnetic resonance imaging (MRI) systems operating at field strengths greater than 3 Tesla (3T), i.e., ultra-high field (UHF) at 7T and above, is becoming popular in cognitive neuroscience since these systems provide greatly increased SNR and sensitivity to BOLD contrast.

One of the earliest discoveries using UHF in the field of cognitive neuroscience was the existence of topographic maps that represent dimensions of numerical cognition (Cai et al., 2021; Harvey et al., 2013; Harvey and Dumoulin, 2017). Following studies, extended this finding of cognitive topographic maps and uncovered maps representing object size (Harvey et al., 2015), time duration (Harvey et al., 2020; Protopapa et al., 2019) and haptic numerosity (Hofstetter et al., 2021). These discoveries suggested that topographic principles common in primary sensory and motor cortices may also be an organizational principle of cognitive functions in association cortex. However, all these studies used 7T functional MRI (fMRI), and anecdotal reports suggested failure to reconstruct these maps at lower field strengths.

Here, we ask whether these cognitive topographic maps can be reconstructed at 3T and will we use these maps and computational models to quantify the differences between 3T and 7T. We focus on visual topographic numerosity maps. Numerosity, the set size of a group of items, is critical to guide human and animals’ behaviour and decision (Burr and Ross, 2008; Carey, 2001; Nieder and Dehaene, 2009). Previously, using population receptive field (pRF) modelling (Dumoulin and Wandell, 2008) at 7T, we have demonstrated that neural population in fMRI recording sites (voxel) are selectively responding, i.e., tuned, to certain preferred numerosities and that this numerosity tuning can be captured with a logarithmic Gaussian model. Furthermore, different cortical locations have different preferred numerosities and these preferred numerosities increase systematically across the parietal cortex, i.e., form-
ing a numerosity topographic map (Cai et al., 2021; Harvey et al., 2013; Harvey and Dumoulin, 2017; Hofstetter et al., 2021).

In this study, we measure BOLD responses to a range of numerosities at 3T and 7T, respectively, and use pRF modelling to evaluate the responses of the numerosity-selective neural populations. We compare the variance explained by the numerosity model to the measured responses at the two field strengths as a function of the number scan runs. In this way, we quantify the extent to which 7T outperforms 3T in terms of the model predictive power. Though there is already an extensive literature on comparisons between field strengths (Duong et al., 2003; Geißler et al., 2013; Pohmann et al., 2016; van der Zwaag et al., 2009), this work directly compares the dependence of model predictive power on field strength in the field of computational neuroimaging.

2. Methods

2.1. Participants

We present data from three participants (one female, age range 22 – 45 years). All participants had normal or correct-to-normal visual acuity. All participants were well educated, with good mathematical abilities. Written informed consent was obtained before every scanning session. All experiments were approved by the ethic committee at University Medical Centre Utrecht.

2.2. Stimulus presentation

Visual stimuli were presented on a 69.84 × 39.29 cm LCD screen (Cambridge Research Systems) placed behind the 3T and the 7T MRI boxes which was viewed through a mirror mounted on the coil. The distance from the mirror to the display screen was 210 / 220 cm at the 3T / 7T scanner rooms, respectively. The stimuli were adjusted to have equal visual angle on the two display screens. The display resolution was 1920 × 1080 pixels.

The visual stimuli were generated in Matlab using PsychToolbox (Brainard, 1997; Kleiner et al., 2007; Pelli, 1997). A large red diagonal cross was displayed continuously across the entire screen (10.2° diameter), which served as an accurate fixation marker for participants. Stimuli consisting of various number of dots presented in the central 1.5° (diameter) of the visual field in a grey background. We used the “constant area” configuration from the original study (Harvey et al., 2013), which keeps the total surface area of all the presented dots combined constant of all the numerosities, ensuring equal luminance across all the dot arrays. The main numerosity stimuli of 1 to 7 dots were presented sequentially in ascending order, followed by a longer baseline period (13.5 s) containing 20 dots. Then the main stimuli were presented in descending order, followed by another baseline period (Fig. 1a). This sequence was repeated 4 times in each functional run. Before the first cycle initialized, there was a pre-scan period (12 s) presenting the baseline numerosity of 20 dots. Dots of all the numerosities were positioned randomly and homogeneously to avoid any links between numerosity and visual position and grouping effects (Fig. 1b). Numerosity stimuli were presented briefly (300 ms) as black dots to ensure participants did not have time to count. This was repeated every 750 ms, each time with a new random dot pattern presented, with 450 ms presentation of a uniform grey background between pattern presentations. Each pattern of the same numerosity was repeatedly presented six times, over 4500 ms, corresponding to 3 fMRI volume acquisitions (repetition time, i.e., TR), before the numerosity changed (Fig. 1c). In 10% of the stimuli presentations, dots were shown in white instead of black. Participants were instructed to press a button when they saw white dots to ensure they were paying attention to the stimuli during fMRI acquisition. No numerosity judgements were required during the experiment. Participants responded correctly on 90–100% of the white dots presentations in each run.

![Illustration of the experimental design and stimuli presentation.](image) Fig. 1. Illustration of the experimental design and stimuli presentation. a. Presented stimulus sequence in which numerosities consisting of 1 to 7 dots were shown in an ascending order followed by a baseline period containing 20 dots, then descended from 7 to 1 followed by another baseline period. b. Two examples of numerosity stimuli presented to the participant in the scanner. The dot array covered the central 1.5° (visual angle) diameter within a 10.2° diameter mean-luminance (grey) screen. A large, thin, red fixation cross passed diagonally through the centre of the display, and through the centre of the dot array. Participants were asked to fixate on the intersection of the cross and press a button when dots were shown in white. c. Schematic representation of stimulus presentation in one fMRI volume acquisition (TR). Numerosity stimulus was presented briefly (300 ms), followed by a 450 ms presentation of a uniform grey background before a new random positioned dot pattern presentation. Each pattern of the same numerosity was repeatedly presented six times, over 4500 ms, corresponding to 3 TRs, before the numerosity changed.

2.3. MRI acquisition

Scanning was carried out on two Philips Achieva scanners operating at 3T and 7T. Functional data were acquired using 32-channel receive head coils (Philips at 3T and Nova Medical at 7T). A multi-slice, single-shot gradient echo (GE) echo planar imaging (EPI) sequence was used at both scanners. Acquisition parameters for the EPI are listed in Table 1. At both systems, all functional runs had 248 volumes and each session had 8 runs. Three 3T and two 7T sessions were acquired for each participant on separate days. Anatomical images of each participant were collected at 7T using an MP2RAGE sequence (Marques et al., 2010) in separate sessions. The key MR parameters of the T1 were as follows: matrix size = 273 × 367 × 367, voxel size = 0.64 × 0.64 × 0.64 mm³, TR/TE = 5.5 s/2.2 ms, TI₁/TI₂ = 0.8/2.7 s, flip angle = 7°/5°. MP2RAGEs are relatively insensitive to the B₁-inhomogeneities present at 7T and yield good segmentation and co-registration results at high spatial resolution (Haast et al., 2018; Huntenburg et al., 2018).

2.4. Data pre-processing

Anatomical data pre-processing included skull stripping and resampling to a spatial resolution of 0.6 × 0.6 × 0.6 mm³. T₁ images were automatically segmented using cbs-tools (Bazin et al., 2014) and then manually edited to minimize auto-segmentation errors using ITK-SNAP (Yushkevich et al., 2006) (www.itksnap.org/). This provides a highly
accurate description of the cortical surface, an anatomical segmentation space used for analysis of cortical organization. The cortical surface was reconstructed at the grey-white matter border and rendered as a smoothed 3D surface. Pre-processing of the functional data was performed using AFNI (Cox, 1996; Cox and Hyde, 1997). The first 8 volumes of each run were discarded to account for signal equilibrium and participants’ adaptation to the immediate environment. Head movement and motion artefacts between and within the remaining volumes were measured and corrected for. All functional images collected at the same session were averaged to generate a common mean EPI image.

Pre-processed functional data were then analysed in mrVista, which is freely available at (https://github.com/vistalab/vistasoft). For each participant, at every session, the mean EPI image was aligned to the anatomy. Individual functional images were then imported and co-registered to the same anatomical space using the same transformation. To vary the signal strength, functional images were averaged with a variable number of runs, e.g., 8, 16 or 24 runs from the 3T and 8 or 16 runs from the 7T sessions. Subsequently, the averaged datasets were collapsed onto the nearest point on the cortical surface across depth, which generated a (folded) 2-dimensional grey matter surface. pRF modelling and subsequent statistical analyses were done at this space, except for the validation analyses using data points across cortical depth (i.e., uncollapsed data, see below). No spatial or temporal smoothing was applied to the functional data.

2.5. Numerosity pRF modelling

We applied pRF modelling to the data using a model that was developed to estimate numerosity tuning properties in human brains (Dumoulin and Wandell, 2008; Harvey et al., 2013). Specifically, a one-dimensional logarithmic model was adopted to predict neuronal responses at each stimulus time point of the numerosity presentation. The model describes tuning in logarithmic numerosity space using a Gaussian function characterized by two parameters: preferred numerosity (central position) and tuning width (standard deviation). A prediction of the neural response time course was produced by overlapping the stimulus at each time point with this tuning model. Then by convolving this prediction with a haemodynamic response function (HRF), a predicted time course was generated. For each voxel on the 2D cortical surface, the parameters were chosen from the prediction that fits the data most closely by minimizing the sum of squared errors between the predicted and observed fMRI time series. The model goodness-of-fit was described by the variance explained (R²). The neural responses of each voxel were described by the pRF model with a particular set of parameters. This modelling procedure was applied to the pre-processed functional data averaged with a variable number of runs, e.g., 8, 16 or 24 runs from the 3T or 8 and 16 runs from the 7T sessions. Thus, we reconstructed the numerosity maps at each field strength, for each participant, respectively.

2.6. Definition of region of interest

We defined region of interest (ROI) on the participants’ right hemispheres at the intraparietal sulcus (IPS), a key brain region for numerosity perception (Dehaene, 2001; Feigenson et al., 2004; Ritter et al., 2018; Nieder, 2016; Nieder et al., 2002), and was the first location where a topographic map of numerosity was found (Harvey et al., 2013). In this study, we refer to this ROI as NPC1 (numerosity map in parietal cortex 1) as defined in previous studies (Harvey and Dumoulin, 2017) and following naming conventions of newly discovered visual field maps in human cortex where homologues to non-human primates are unclear (Wandell et al., 2007). NPC1 lays in the right hemisphere, on the gyrus posterior to the superior postcentral sulcus, and its centre position was found at (22, −61, 60) in Montreal Neurological Institute (MNI) coordinates (Harvey et al., 2013; Harvey and Dumoulin, 2017).

2.7. Model-based analysis

To compare the model predictive power of the two field strengths, without relying on a predefined model for the functional responses, reference datasets were generated by averaging 8 functional runs from each field strength, respectively (Fig. 2a). The remaining individual functional runs were taken as independent test datasets (Fig. 2d). Every session was taken as the reference dataset in turn, including all 3T or 7T sessions. Hence, when taking one 7T session as the reference dataset, the independent test dataset included 8 functional runs from the other 7T session and 24 runs from the 3T sessions, etc. More comparisons between reference dataset and independent dataset were made by averaging runs across scanning sessions. The functional runs of the two 7T sessions were mixed based on odd or even order, resulting in 4 new reference datasets. Similarly, we created 3 new reference datasets consisting of eight 3T runs by recruiting every 2 runs from the total 24 3T functional runs. Thus, for each field strength, six different reference datasets and the remaining runs as test datasets were generated for further analysis.

We selected voxels from the reference dataset based on two criteria: (1) variance explained (R²) exceeded 30% and (2) the preferred tuning fell within the presented numerosity range (Fig. 2b). Model predicted time series of the selected voxels in NPC1 were extracted as the reference model (Fig. 2c). For each selected voxel, we extracted the time series of each individual run in the independent test datasets (Fig. 2d). These time series were shuffled (n = 100) and averaged with increasing number of runs included to produce a new time series of the voxel (Fig. 2e). By fitting the reference model with the averaged time series, we obtained the variance explained (R̂\text{model}) of the reference model as a function of increasing number of runs (Fig. 2f).

We iterated this procedure 6 times for each field strength while splitting the data into different pairs of reference and test datasets. We averaged the results of each field strength to compare the model predictive power between the two field strengths as a function of number of runs. We then performed a linear fit of how many 3T runs are required to have the same variance explained of one 7T run. This procedure was done for each participant individually. Finally, an overall factor between 3T and 7T in terms of number of runs was obtained by averaging the linear fits across participants.

2.8. Calculation of noise ceiling

To quantify the maximum explainable variance given the noise in the data, we computed the noise ceiling (NC) (Lage-Castellanos et al., 2019; Machens et al., 2004; Mante et al., 2005). Specifically, we employed the method described by Machens et al. (2004). Briefly, we calculated the noise ceiling as the fraction of variance in the residual noise (σᵣ²) over the variance in the response power (σᵢ²):

\[
NC = 1 - \frac{\sigmaᵣ²}{\sigmaᵢ²}
\]  

Table 1

| Field strength (B₀) | TR (ms) | Vox size (mm³) | TE (ms) | FA (°) | FOV (mm²) |
|---------------------|---------|----------------|---------|--------|-----------|
| 3T                  | 1500    | 1.98 × 1.98 × 2.00 | 28      | 80     | 46 × 190 × 190 |
| 7T                  | 1500    | 1.98 × 1.98 × 2.00 | 25      | 70     | 50 × 190 × 190 |
This is basically the maximal variance explained given the noise in the data. The response power is defined as the average variance over the session \( \langle s_t \rangle \) with \( t = 1 \ldots M \) time-points:

\[
\sigma_s^2 = \left\langle \frac{1}{M} \sum_{t=1}^{M} s_t^2 \right\rangle
\]

the variance in the residual noise is estimated as:

\[
\sigma_n^2 = \frac{n}{n-1} \left[ \left\langle \frac{1}{M} \sum_{t=1}^{M} s_t^2 \right\rangle - \frac{1}{M} \sum_{t=1}^{M} s_t^2 \right]^{-1}
\]

where \( n \) indicates the number of sessions, angular brackets denote averaging over sessions. Basically, adding up the variance of independent sessions will include the noise in each session, while computing the variance after averaging the sessions will remove the noise between sessions. Their difference is an estimate of the residual noise. The assumptions behind this estimation are minimal: the noise should have zero mean and a non-infinite variance, and should be independent between sessions.

As we used one session as a reference dataset in the analysis, we computed the noise ceiling across sessions each consisting of 8 runs either at 3T or 7T. Since the number of voxels selected for further analysis varied based on the reference sessions (see Fig. 2 of the analysis flowchart), the noise ceiling was calculated with the voxels selected based on different reference sessions in turn for each iteration. We averaged the noise ceiling of all the six iterations to have the noise ceiling of one session (i.e., 8 runs) at each field strength.

### 2.9. Calculation of tSNR

Temporal SNR (tSNR) is defined on a voxel-wise manner as the ratio of the mean across time divided by the standard deviation across time. To avoid bias by large response in active grey voxels, we calculated tSNR in white matter (WM) in addition to grey matter (GM). A whole-brain WM mask was defined from the segmented anatomy for each participant. The ROI for calculating tSNR in GM is confined to the numerosity map NPC1. We calculated tSNR as the average tSNR across voxels in the WM mask and GM ROI of each individual run at 3T and 7T, respectively. We reported the average tSNR across runs and participants at each field strength, respectively.

### 2.10. Comparing preferred numerosity and tuning width estimates at 3T and 7T

Pearson correlation analysis was performed between numerosity preference and tuning width estimates derived from the 3T and 7T data. This included the voxels that had variance explained above 30% in the maps constructed using all the acquired 3T (24 runs) or 7T (16 runs) data, and the preferred tuning fell within the presented range. Taking into account the functional resolution of the recording sites, the total number of data points used to calculate correlation’s probability was reduced by the factor by which functional voxels were up-sampled onto the 2D cortical surface. We quantified the similarity between the pRF estimates at 3T and 7T by dividing the subtraction of two estimates (e.g., \( X_{7T} \) and \( X_{3T} \)) by their mean, and converted to percentage: \( ((X_{7T} - X_{3T}) / (X_{7T} + X_{3T})/2) * 100\% \).

### 2.11. Validation analyses

In the model-based analysis we used the reference model and compared the variation explained to the test datasets at 3T and 7T. This analysis depends on the accuracy of the model. In order to perform a model-free analysis, we extracted the time series of the selected voxels (same criteria as in the model-based analysis) from the reference datasets as a reference time series. Analogous to the model-based analysis, the time series of each individual run in the test datasets (3T or 7T) were averaged with increasing number of runs to produce the averaged time series for each voxel. Applying with a Pearson correlation analysis, we obtained the correlation coefficients between the reference time series and the averaged time series from the test datasets as a function of increasing number of runs \( (r_{ref}) \). This procedure is illustrated in Fig. S1.

Furthermore, we validated our results by performing both the model-based and model-free analyses using all data points across the cortical
depth (i.e., un-collapsed data), and using all the data points within the ROI, without any threshold.

3. Results

3.1. BOLD responses of numerosity-selective neural populations at 3T and 7T

In Fig. 3a, we show two examples of representative time series and models. The representative recording site was selected randomly from the 3T data points which has the averaged variance explained amongst the selected voxels. We extracted the time series of this recording site from the datasets that averaging 24 and 16 runs at 3T and 7T, respectively. The 7T time series (blue points) exhibits larger response amplitude than the 3T time series (red points). The two model predictions explain 61% and 77% of the variance in these time series recorded at 3T (red line) and 7T (blue line), respectively. The higher percentage BOLD signal change and variance explained at 7T confirms the higher BOLD signal sensitivity and SNR at ultra-high field. The pRF models with a particular set of parameters that best fitted to each time series are shown in Fig. 3b. The peak response amplitude indicates the preferred numerosity and the full width half maximum (FWHM) reflects the tuning width of the pRF voxel. At this example recording site, the preferred numerosity was slightly smaller and the tuning width was larger when estimated at the 3T data (upper panel) than at the 7T data (lower panel).

3.2. Numerosity map is more reliably detected at 7T than at 3T

Fig. 4 presents the reconstructed numerosity maps of participant 1 using increasing number of runs acquired at 3T and 7T. Consistent with previous studies (Cai et al., 2021; Harvey et al., 2013; Harvey and Dumoulin, 2017), we found numerosity-selective neural populations in the parietal cortex that are topographically organized. We found a larger cortical extent above the variance explained threshold of 30% at the 7T (mean number of voxels ± SE: 904 ± 56) than the 3T data (695 ± 28). The 3T data were noisier and therefore could not be adequately captured by the model. At both field strengths, as the number of runs increased, the noise reduced and the topographic maps became more robust.

As Fig. 4 shows, the numerosity maps obtained at 3T were similar to those obtained at 7T. This was also found for the other participants (Fig. S2). We then quantified the similarity between the preferred numerosity and tuning width estimates at the two field strengths by a Pearson correlation analysis. The Pearson correlation analysis indicated that the numerosity preference estimates derived from the two field strengths were highly correlated (r > 70%) (Fig. S3a). However, this was not the case for tuning width (Fig. S3b). Overall, the preferred numerosity estimates at 7T were slightly higher than at 3T, while the tuning width estimates were broader when recorded at 3T (Fig. S3c). We speculate that the smaller tuning width at 7T maybe mediated by the larger sensitivity to smaller vessels (Duong et al., 2003; Yacoub et al., 2001), and the small differences in preferred tuning maybe influenced by the tuning width. However, given the sensitivity of tuning width estimates to algorithmic consideration, e.g., the HRF estimations (Dumoulin and Wandell, 2008; Lerma-Usubiaga et al., 2020) and data quality, we refrain from drawing too strong a conclusion.

3.3. One 7T run has four times the model predictive power of one 3T run

Fig. 5a shows the variance explained by the 7T reference model as a function of increasing number of runs of 3T and 7T data, together with the noise ceiling of the 7T data in one session (i.e. 8 runs, dashed lines). The variance explained increased as the number of runs increased at both field strengths. However, the increase in variance explained was faster at 7T than at 3T. In other words, more 3T runs were required to reach the same predictive power at 7T. Furthermore, the reference model always captured more variance of the 7T responses, thus the resulting variance explained was always higher than that at 3T. For example, averaging 24 3T runs ($R^2 = 53\%$) still could not reach the same variance explained of averaging 8 7T runs ($R^2 = 59\%$) for any of the participants.

The model predictive power is constrained by the noise present in the actual response. To quantify the maximum explainable variance (in one session) given the noise in the data, we computed the noise ceiling (see Methods). As shown in Fig. 5a, averaging 24 3T runs or averaging 8 7T runs always yielded a lower predictive power than the noise ceiling of one 7T session (8 runs).

Next, we then calculated how many 3T runs were required to achieve the same model predictive power as a function of the number of 7T runs (Fig. 5b). On average, one 7T run has 4 times the variance explained of one 3T run using the 7T reference model.

Similar results were obtained using the reference model derived from 3T reference datasets (Fig. S4). Averaging the same number of 3T runs (n = 8) could not reach the noise ceiling of one 3T session. However, the predictive power on the 3T and 7T data increased with increasing number of runs and ultimately outperformed the noise ceiling of 3T when more than 8 runs were included (Fig. S4a). When using the 3T reference model, the number of 3T runs to match the 7T data was smaller.
between grey (mean 72% (errors Fig. 38x519 to 398x737).

Comparison matter time data stimulus-driven in average, points the due and 4) tSNR 7T 6 runs one averaged session, (black).

Fig. 5. Quantification of field strength effects on pRF model predictive power as a function of number of runs, using the reference model derived from 7T reference datasets. a. The variance explained of the reference model as a function of increasing number of runs at 3T (red) and 7T (blue). Shaded areas indicate standard errors of the mean over iterations using different reference datasets (n = 6). The noise ceiling (dashed line) with 95% confidence intervals (grey bars) represents the maximum explainable variance (of one 7T session, i.e. 8 runs) given the noise in the data. b. Linear fits of the number of runs required at 3T to have equivalent model predictive power of one 7T run. Coloured-coded texts indicate the factor between 3T and 7T runs to achieve the same variance explained for each participant. On average, one 7T run has 4 times the model predictive power of one 3T run using the 7T reference model (black).

(3T)\textsuperscript{b}, likely due to the noisier data quality at 3T as indicated by the lower noise ceiling, and which likely also resulted in a noisier reference model. Overall, the 7T data had a higher noise ceiling (mean ± SD: 72%\pm 2.6%) than the 3T data (55% ± 3.7%). These results suggest that 3T data is noisier and the benefits in model predictive power is due to improved data quality, rather than model accuracy.

Last, we found similar results using model-free analyses of voxel-time series correlation (Fig. S5), model-based analyses using uncollapsed data points (Fig. S6) and using all data points within NPC1, i.e. no thresholding (Fig. S7).

3.4. Comparison of tSNR at 3T and 7T

Fig. 6 shows the tSNR maps of one example individual run at 3T (Fig. 6a) and 7T (Fig. 6b), and the averaged tSNR in white matter and grey matter across all runs and participants at the two field strengths (Fig. 6c). On average, one 7T run has higher tSNR in white matter (mean ± SE: 79 ± 4) than that at 3T (72 ± 2). However, at the cortical grey matter of the numerosity map (NPC1), on average, 3T has higher tSNR (74 ± 4) than 7T (66 ± 6) in an individual run. The difference between cortical GM and WM tSNR likely reflects the increased contributions of stimulus-driven BOLD signal fluctuations and physiological noise present in cortex at 7T.

4. Discussion

We recorded BOLD responses of numerosity-selective neural populations in human parietal cortex at 3T and 7T, respectively. We used identical numerosity stimuli and tasks and a similar functional MRI sequence which we optimized for each field strength. We applied identical pre-processing pipeline and analysed the data using biologically-inspired model-based analyses (Dumoulin and Wandell, 2008). Subsequently, we quantified the number of runs required to detect reliable numerosity maps at 3T, compared to 7T, for individual participants. Field strength effects on the functional data were examined using model predictive power. We were able to reconstruct the topographic numerosity maps in the intraparietal sulcus at 3T. However, the topographic maps derived from the 3T data were less reliable and required much more data than typically acquired in the field. The numerosity maps at both field strengths became more reliable with increasing number of runs, though the rate of increase was higher at 7T. On average across participants, one 7T run had about 4 times the model predictive power of one 3T run.

To acquire comparable data at the two different field strengths, we utilized a GE EPI sequence that we optimized to ensure a good SNR and signal strength at both field strengths. Specifically, a voxel volume...
size of $1.98 \times 1.98 \times 2.00 \, \text{mm}^3$ was used at both fields. At 7T, the TE was close to the tissue $T_2^*$ and hence optimum for BOLD contrast, while the TE at 3T (28 ms) was relatively short compared to the grey matter $T_2^*$ (Peters et al., 2007). Such a short TE is very widely used at 3T to allow acquisition of a higher number of slices in an achievable TR (Clare, 1997; Volz et al., 2019). At both field strengths the FA was set close to the Ernst angle, and as a result, it was slightly higher at 3T (80°) than at 7T (70°). Though obviously, these settings influence the results, we do not believe that they bias the results to favour one field strength over the other.

Previous studies show that imaging at UHF provides a leap forward in both higher SNR and BOLD signal sensitivity. In the context of fMRI, the static image SNR reflects MRI signal strength over the noise present in the image in the absence of signal. Pohmann et al. (2016) demonstrated that the image SNR showed a distinctly supralinear increase with field strength by a factor of $3.10 \pm 0.20$ from 3T to 7T, and $1.76 \pm 0.13$ from 7T to 9.4T over the entire cerebrum. However, fMRI signals include contributions from thermal noise and correlated interference due to head motion, scanner instability and non-neuronal physiological noise arising from cardiac and respiratory fluctuations. As the magnetic field strength increases, the relative contribution of non-neuronal physiological noise is also increased (Triantafyllou et al., 2005). If physiological noise contributions dominate over the thermal noise in the imaging voxel, the SNR is independent of signal strength (Krüger and Glover, 2001), resulting in a reduced ability to detect activation-induced signal changes. Although the physiological noise contribution is higher at UHF, the noise effect would be cancelled out when averaging multiple functional runs as the cardiac and respiratory signals are not task-locked. Thus, we believe that the data here presented after averaging over runs have only a small contribution from physiological noise sources.

Furthermore, magnetic field strength increase leads to significant increase in the BOLD contrast (Gati et al., 1997; Krasnow et al., 2003). The BOLD signal arises from the field inhomogeneity differences induced by the paramagnetic deoxyhaemoglobin in the capillaries and venous vessels and the surrounding tissue, which manifests as signal changes in the order of a few percent (Logothetis, 2002). This BOLD contrast scales approximately linearly with field strength. A quantitative analysis conducted by van der Zwaag et al. (2009) found that in brain tissue, the BOLD contrast (approximated by the relaxation rate change, $\Delta R_2^*$) increases linearly with field strength (0.98 $\pm$ 0.08 at 3T and 2.55 $\pm$ 0.22 at 7T). Yacoub et al. (2001) found a supralinear field strength dependence of BOLD contrast that increased by a factor of $2.13 \pm 0.23$ when going from 4T to 7T. The resulting increase in BOLD contrast is of great benefit for fMRI studies and can be exploited to reduce the number of functional runs required to demonstrate robust activation.

In the current study, the BOLD responses recorded at 7T benefit both from the increased SNR and augmented BOLD contrast, resulting in the factor of 4 times the number of 3T runs required to achieve equivalent model predictive power of one 7T run. The factor of the number of 3T runs to match the 7T data was smaller using 3T data as reference compared to using 7T data as reference. We suspect this is due to the noiser data quality of 3T data. Noisier data results in a pRF model that is less accurate, which in turn limits the amount of variance in the test data it can explain. The lower noise ceiling of 3T data than 7T data also indicate the noiser data quality at 3T. Overall, the 7T data is less noisy than the 3T data, yielding the higher noise ceiling. However, taking into account the noise in the 3T data, the model predictive power is comparable between different field strengths. In other words, relative to the noise ceiling, the pRF model is applicable and independent to field strengths. Furthermore, the similar results obtained from the model-free analysis suggest that the benefits of using less 7T runs to obtain robust numberosity maps is induced by data quality but not model accuracy.

Akin to previous studies we found 7T has higher tSNR than 3T in white matter. The higher tSNR in white matter at 7T is related to the higher SNR at UHF as tSNR will increase with increasing image SNR, until a field-strength dependant plateau value is reached (Krüger and Glover, 2001; Triantafyllou et al., 2005). The tSNR is lower in grey matter (i.e., NPC1) at 7T than 3T. This is likely caused by the larger (stimulus-driven) BOLD responses and higher physiological noise contributions at 7T. The different behaviour of tSNR in grey and white matter is more pronounced in this study because of the extra task-induced variance and the relatively short TE at 3T, e.g., compared to studies that fixed TE=$T_2^*$ in a resting state acquisition (Triantafyllou et al., 2005). The discrepancy that the grey matter tSNR at 7T is lower than at 3T while the noise ceiling is higher, is due to the fact that task-induced BOLD fluctuations are ‘signal’ when calculating the noise ceiling, while they contribute to ‘variance’ in the tSNR calculation.

This finding is in agreement with previous literature. To achieve a significant statistical power, many runs of a single participant and/or groups of participants are acquired at conventional field strength scanner (i.e. 3T). In such a way, noise present in fMRI time series is reduced when multiple runs are averaged together, which leads to a monotonic increase in statistical significance as increasing number of runs. Using a GO/NOGO task, Torrissi et al. (2018) compared GLM-based activation analyses and showed significant gains in statistical power at 7T. However, subjects were not selected at group level to match the same power at 3T. Gonzalez-Castillo et al. (2012) acquired 100 functional runs at each of the 3 participants at 3T and later performed a similar study at 7T (Gonzalez-Castillo et al., 2015), where much less 7T runs (~25 runs) were required to reach the same percent of grey matter voxels above a statistical threshold (activated) as 100 3T runs could achieve. This is a similar factor of 4 times improvement as we suggest here. At UHF scanner (i.e. 7T), higher SNR and tSNR will reduce the number of runs required from a single participant to detect activation with an expected statistical power (Murphy et al., 2007). We note that this is particularly relevant for computational neuroimaging, where signals of single voxels differ and are modelled separately. Furthermore, because the topographic map locations, size and orientations vary between participants
(Dumoulin et al., 2006; Harvey et al., 2013; Wandell et al., 2007), averaging of participants in a common space is often not feasible.

This study differs from other studies comparing field strengths dependence on BOLD signal in three aspects: first, we used a numerosity task that activated brain regions associated with high level cognition. This experimental design was the same as the paradigm initially used to uncover the topographic representation of numerosity at UHF (Harvey et al., 2013). Previous studies that compared the BOLD signal sensitivity between different field strengths mainly used simple tasks, e.g., flicker stimulation or finger tapping, to activate the primary sensory and/or motor cortex (Duong et al., 2003; Geißler et al., 2014, 2013; Schäfer et al., 2008). Only a few studies adopted high level cognitive tasks to compare lower field and UHF. For example, Jerde et al. (2008) compared the task-induced activation at 4T and 7T (Gourtzelidis et al., 2005) using a mental maze solving paradigm, and Geißler et al. (2014) compared the language network with a standard overt language fMRI paradigm between 3T and 7T. Although there have been many studies comparing field strengths performed by experts in physics and engineering (Geißler et al., 2013; Hutton et al., 2011; Li, 2013; Pohmann et al., 2016; Vaughan et al., 2001), additional empirical evidence using neurocognitive tasks may also aid the cognitive neuroscientist’s decision to execute fMRI experiments at UHF (De Martino et al., 2018; van der Zwaag et al., 2016). Second, rather than using conventional univariate analysis, such as GLM, we used a custom-built computational pRF modelling. Though the pRF model is conceptually similar to GLM by taking the best model fit as the predictor in the design matrix, there are several advantages of using pRF model to quantify the field strength dependant effect on model predictive power. (i) The pRF model is an explicit computational model and is expressed in terms of input-referred parameters (Dumoulin and Wandell, 2008; Wandell and Winawer, 2015) such as locations in the visual field rather than in terms of a statistic of the fMRI time series. Compared to a GLM, the pRF model characterizes the responses of neural populations that preferentially tuned to different stimuli (e.g., numerosities). The response differences to the presented numerosities could be converted into tuning functions, allowing for a comparison of the tuning parameters (i.e., preferred numerosity and tuning width), and model predictive performance (i.e., variance explained) of the fMRI signals. (ii) Our approach was motivated by the anecdotal suggestion that numerosity maps as discovered by the pRF modelling at 7T could not be reproduced at 3T. As such, the comparison between 3T and 7T became relevant for the pRF modelling. Though we show that we can reconstruct the numerosity maps at 3T, this requires much more data than typically acquired in the field. Last, we used model-based, model-free and other validation analyses, and these analyses showed similar results.

One of the limitations of the current study is that we only have three participants, but each participant was scanned for three sessions at 3T and two sessions at 7T. We prefer scanning multiple sessions on fewer participants than scanning more participants with fewer sessions for several reasons. First, the aim of the current study is to investigate whether we could detect numerosity maps at 3T and quantify how many functional runs are required at 3T to reach equivalent model predictive power at 7T. Thus, for each individual participant, it is necessary to have more than one 3T session so as to have enough signal strength to compare to a 7T session. Second, pRF model is commonly used to map functionally specialized brain regions on individual participant, for example, numerosity maps in the intraparietal sulcus. We ran the model in the native space of each participant’s cortical area thus it is not helpful to average these individual-specific cognitive maps across participants. Third, having more sessions on the same participant would help to reduce the confounds of between-subject variability for comparing different field strengths. To counter-balance the variability of the 3T data collection on separate days, we also collected two 7T sessions on two days. Each session was used as the reference dataset in turn to reduce session-specific variability (Vieussmann and Polimeni, 2021). Last, statistical power is a trade-off between number of trials per participant and number of participants (Baker et al., 2020). Studies with fewer participants but more trials can have the same statistical power as more participants with fewer trials. Thus, we opt for collecting more sessions on the same participant rather than having one session on multiple participants.

5. Conclusion

With the increasing popular application of computational model in neuroimaging, UHF MRI brings tremendous advantages in advancing our understanding of the brain function, such as increased sensitivity and greater spatial resolution. This study brings out another benefits of UHF MRI and demonstrates higher model predictive power at UHF. These results suggest that future cognitive neuroscience studies may benefit from UHF by collecting less data and preserving strong statistical power. Thus, UHF functional MRI paves the way for individualized cognitive neuroscience.

Originally, with all the control experiments involved, it took about 5 hours of scanning at 7T per individual participant to discover the numerosity maps (Harvey et al., 2013). Based on the results we report here, it would have required around 20 hours per participant to uncover the numerosity maps at 3T, which would have made the initial discovery of numerosity maps at 3T highly unfeasible in practice. To sum up, UHF benefits cognitive neuroscience with higher SNR and BOLD sensitivity, and thus reduces the number of runs (trials) required to achieve reliable activation compared to lower field strength.

Declaration of Competing Interest

The authors declare no competing interests.

Acknowledgements

This research was supported by an Ammodo KNAW Award (S. O. D.), an NWO-VICI grant 016.Vici.185.050 (S. O. D.), a KNAW research grant (S. O. D. and W. van der Z.) and a China Scholarship Council (CSC) scholarship [201706750008] (Y.C.).

Author contributions

S. O. D. conceived the study. W. van der Z. developed the MR sequences. Y. C. collected and analysed the data. W.Z. provided help with some of the analyses code. Y. C. wrote the paper. S. H., W. van der Z. and S. O. D. edited the paper. S. O. D. and S. H. provided supervision.

Data and code availability statement

The code generated during this study is available in the Vistasoft repository (https://github.com/vistalab/vistasoft).

The datasets supporting the current study have not yet been deposited in a public repository because of participants privacy concerns, but are available from the corresponding author on request.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2021.118184.

References

Baker, D.H., Villidaite, G., Lygo, F.A., Smith, A.K., Flack, T.R., Gouws, A.D., Andrews, T.J., 2020. Power contours: optimising sample size and precision in experimental psychology and human neuroscience. Psychol. Methods doi:10.1037/met0000337.

Bazin, P.-L., Weiss, M., Dineen, I., Schäfer, A., Trampel, R., Turner, R., 2014. A computational framework for ultra-high resolution cortical segmentation at 7Tesla. Neuroimage 93, 201–209. doi:10.1016/j.neuroimage.2013.03.077.

Brainard, D.H., 1997. The psychophysics toolbox. Spat. Vis. 10, 433–436.
Wandell, B.A., 1999. Computational neuroimaging of human visual cortex. Annu. Rev. Neurosci. 22, 145–173. doi:10.1146/annurev.neuro.22.1.145.
Wandell, B.A., Dumoulin, S.O., Brewer, A.A., 2007. Visual field maps in human cortex. Neuron 56, 366–383. doi:10.1016/j.neuron.2007.10.012.
Wandell, B.A., Winawer, J., 2015. Computational neuroimaging and population receptive fields. Trends Cogn. Sci. 19, 349–357. doi:10.1016/j.tics.2015.03.009.
Yacoub, E., Shmuel, A., Pfeuffer, J., Van De Moortele, P.F., Adriany, G., Andersen, P., Thomas Vaughan, J., Merkle, H., Ugurbil, K., Hu, X., 2001. Imaging brain function in humans at 7 Tesla. Magn. Reson. Med. 45, 588–594. doi:10.1002/mrm.1080.
Yushkevich, P.A., Piven, J., Hazlett, H.C., Smith, R.G., Ho, S., Gee, J.C., Gerig, G., 2006. User-guided 3D active contour segmentation of anatomical structures: significantly improved efficiency and reliability. Neuroimage 31, 1116–1128. doi:10.1016/j.neuroimage.2006.01.015.