Wüthrich Florian (Orcid ID: 0000-0001-6007-8668)
Lefebvre Stephanie (Orcid ID: 0000-0003-4833-2197)
Nadesalingam Niluja (Orcid ID: 0000-0002-5221-5148)
Bernard Jessica (Orcid ID: 0000-0001-7697-3675)

# Test-retest reliability of a finger-tapping fMRI task in a healthy population

## Task-based fMRI test-retest reliability

Florian Wüthrich[1-3], Stephanie Lefebvre[1,2], Niluja Nadesalingam[1-3], Jessica A. Bernard[4], Vijay A. Mittal[5-9], Stewart A. Shankman[5,6], Sebastian Walther[1,2]

[1] Translational Research Center, University Hospital of Psychiatry and Psychotherapy, University of Bern, Switzerland
[2]Translational Imaging Center (TIC), Swiss Institute for Translational and Entrepreneurial Medicine, Bern, Switzerland
[3] Graduate School for Health Sciences, Universityx of Bern, Switzerland
[4]Texas A&M University, Department of Psychological and Brain Sciences, Texas A&M Institute for Neuroscience, Texas A&M University, College Station, TX, USA
[5]Northwestern University, Department of Psychiatry and Behavioral Sciences, Chicago, IL, USA.
[6]Northwestern University, Department of Psychology, Evanston, IL, USA
[7]Northwestern University, Institute for Innovations in Developmental Sciences, Evanston/Chicago, IL, USA
[8]Northwestern University, Institute for Policy Research, Evanston, IL, USA
[9]Northwestern University, Medical Social Sciences, Chicago, IL, USA

Correspondence: Florian Wüthrich, MD
Translational Research Center
University Hospital of Psychiatry and Psychotherapy
Bolligenstrasse 111
CH 3000 Bern 60, Switzerland
Phone: +41 31 932 84 19
E-Mail: florian.wuethrich@upd.unibe.ch

## Acknowledgements

We would like to thank all study participants for taking part in this study.

## Data Availability Statement

The data used in this study are available from SW upon reasonable request.

## Ethics Statement

Written informed consent was obtained from all participants. The study protocol adhered to the Declaration of Helsinki and was approved by the local ethics committee. The OCoPS-P trial has been registered on ClinicalTrials.gov with identifier: NCT03921450.

Funding
This study was supported by the Swiss National Science Foundation (#182469 to SW) and the National Institute of Mental Health (#R01 MH118741 to SAS, VAM, and SW).

Conflict of interest
SW has received honoraria from Janssen, Lundbeck, Mepha, Neurolite, and Sunovion. All other authors report no conflicts of interest.

Abstract
Measuring brain activity during fMRI tasks is one of the main tools to identify brain biomarkers of disease or neural substrates associated with specific symptoms. However, identifying correct biomarkers relies on reliable measures. Recently, poor reliability was reported for task-based fMRI measures. The present study aimed to demonstrate the reliability of a finger-tapping fMRI-task across two sessions in healthy participants.

Thirty-one right-handed healthy participants aged 18-60 took part in two MRI sessions three weeks apart during which we acquired finger-tapping task-fMRI. We examined the overlap of activations between sessions using Dice-Similarity-Coefficients, assessing their location and extent. Then, we compared amplitudes calculating Intraclass-Correlation-Coefficients (ICC) in three sets of Regions-of-Interest (ROIs) in the motor-network: Literature-based ROIs (10mm-radius spheres centered on peaks of an activation-likelihood-estimation), anatomical ROIs (regions as defined in an atlas), and ROIs based on conjunction analyses (super-threshold voxels in both sessions).

Finger-tapping consistently activated expected regions, e.g. left primary sensorimotor cortices, premotor area, and right cerebellum. We found good-to-excellent overlap of activations for most contrasts (Dice-coefficients .54-.82). Across time, ICCs showed large variability in all ROI-sets (.04-.91). However, ICCs in most ROIs indicated fair-to-good reliability (mean=.52). The least specific contrast consistently yielded the best reliability.

Overall, the finger-tapping task showed good spatial overlap and fair reliability of amplitudes on group-level. While caution is warranted interpreting correlations of activations with other variables, identification of activated regions in response to a task and their between-group comparisons are still valid and important modes of analysis in neuroimaging to find population tendencies and differences.

Keywords: Dice similarity coefficient; DSC; intraclass correlation coefficient; ICC; motor task;

Abbreviations: MR or MRI: magnetic resonance imaging; fMRI: functional MRI; BOLD: blood oxygen level dependent signal; ROI: region of interest; MP2RAGE: magnetization-prepared 2 rapid gradient echoes sequence; mbep2d: multiband accelerated echo planar imaging sequence; TR: repetition time; TE: echo time; FOV: field of view; FWHM: full width at half maximum; CI: confiddence interval; FWE: family-wise error rate; ICC: intraclass correlation coefficient; DSC: Dice similarity coefficient; ALE: activation likelihood estimation; TIF: sound

paced thumb-index finger tapping; TIFfast: unpaced thumb-index finger tapping; TAF: sound paced thumb-alternating finger opposition; TAFfast: unpaced thumb-alternating finger opposition; M1: primary motor cortex S1: primary sensory cortex; SMA: supplementary motor area; COVID-19: coronavirus disease 2019.

## Introduction

Measuring the neural substrates associated with a motor or cognitive task using functional MRI has been extensively used in neurosciences in the past two decades (Sadraee, Paulus, & Ekhtiari, 2021) and has increasingly been utilized in clinical applications, e.g. for preoperative mapping for brain surgery (Jalilianhasanpour et al., 2021; Manan, Franz, & Yahya, 2021) or neurofeedback therapies (Dudek & Dodell-Feder, 2021; Thibault, MacPherson, Lifshitz, Roth, & Raz, 2018). The majority of studies evaluating task-based brain activation do so by contrasting the blood oxygenation level-dependent (BOLD) signals during active and control conditions as described by Ogawa in 1990 (Ogawa, Lee, Kay, & Tank, 1990). This approach allows for a wide variety of task designs and examination of various neural processes. To be able to draw conclusions and base further research on previous results, the reliability of the measure is utterly critical. However, there are different forms of reliability and which of its forms is critical depends on the measure and construct under investigation: Internal consistency reliability is crucial for tasks measuring rapidly changing states (e.g. emotions), interrater (i.e. inter-scanner) reliability is essential for multicentric studies, the discovery of traits and prognostic or predictive markers relies on test-retest reliability.

Reliability of functional imaging has been measured using several different metrics. One of the first and most crude measures to assess reliability is comparing the number of activated voxels. However, this only allows estimating whether the amount of activation is comparable, but does not contain information on spatial distribution of this activation. Consequently, this method has fallen out of use for evaluation of fMRI reliability (Cohen & DuBois, 1999). An

alternative approach assesses the spatial distribution of activation by measuring the spatial overlap of brain activation can be performed using the Dice or Jaccard coefficients (Dice, 1945; Jaccard, 1912). This form of reliability is especially crucial for studies aiming to identify brain regions that are involved in a specific task. Finally, intraclass correlation coefficients (ICC) can additionally assess the amplitudes or weights of activations within voxels or regions of interest over time (Shrout & Fleiss, 1979). Reliable amplitudes are crucial when correlational or regression analyses are planned.

Several publications have reported low reliability of task-based fMRI and a recent meta-analysis by Elliott et al. reported low test-retest reliability across various tasks, especially on the single-subject level (Bennett & Miller, 2013; Elliott et al., 2020). Poor subject-level reliability sets limits to the minimal observable effect sizes for correlational analyses. However, examination and comparison of activated regions among groups is still the most frequently used form of analysis and utilizes spatial group-level reliability of superthreshold clusters of voxels. Moreover, a large proportion of the literature on task fMRI reliability is based on data from older scanners. Considering the tendency to higher field strengths, shorter TR, acquisition acceleration, and optimized processing pipelines, more studies assessing reliability in modern settings are needed.

While the average test-retest reliability in the meta-analysis of Elliott et al. was poor, there was a large range across the included studies, suggesting that the reliability of task-based fMRI might vary on the specific task and its implementation. Interestingly, four of the 10 studies with the highest reliability used motor tasks: Friedman et al. (2008) examined paced alternating button pressing with audiovisual cues, Rath et al. (2016) investigated fist-clenching, Estevez et al. (2014) studied robot-assisted elbow motion, and Kimberley et al.

(2008) used a drawing task. Motor tasks may be ideal to examine test-retest reliability, as the targeted brain regions are well characterized. One of the most classic motor tasks is finger-tapping. Body movement may induce head movement in the scanner, which should be minimized. Finger-tapping tasks allow minimization of the coupling of body and head movement, as the hands can usually move freely even in the confined space of a scanner and with elbows fixed for stabilization. Frequent implementations of this task are the use of a button box or free moving thumb-finger opposition. Both, button pressing and thumb-finger opposition fMRI test-retest reliability have been examined in the past (Ibinson et al., 2022; Lee et al., 2010; Marshall et al., 2004; Yoo, Wei, Dickey, Guttmann, & Panych, 2005). However, button pressing might differ from the more naturalistic thumb-finger opposition, especially in fast, unpaced paradigms and there are surprisingly few reliability studies of these tasks, considering their frequent use. Moreover, the studies examining thumb-finger opposition were either conducted with longer repetition time (TR ≥ 3s) and in lower field strength (i.e. 1.5T) than the current standard, examined only one variation of finger-tapping (either externally or self-paced), or restricted the analyses to either spatial overlap or region-of-interest activation amplitude comparison. In this study, we aimed to assess spatial and amplitude test-retest reliability of an fMRI task investigating fine motor behavior on a group-level testing multiple versions of finger tapping in two separate sessions three weeks apart. We expected relatively consistent activation of a motor network most pronounced in left primary motor and sensory cortex (M1, S1), premotor cortex, supplementary motor area (SMA), parietal regions, basal ganglia, and right cerebellum (Witt, Laird, & Meyerand, 2008).

## Materials & Methods

### Participants

We recruited 42 right-handed healthy participants from the general population in Switzerland as a control group for a larger project (OCoPS-P, ClinicalTrials.gov identifier: NCT03921450). Participants were recruited via advertisements and word-of-mouth. Inclusion criteria were right-handedness as confirmed by the Edinburgh Handedness Inventory (Oldfield, 1971), age between 18 and 60, ability and willingness to participate in the study. Exclusion criteria were substance abuse other than nicotine, history of psychiatric disorders or medical conditions impairing movements, epilepsy, history of head trauma with loss of consciousness, and contraindications for MR scans, i.e. metal objects in the body or pregnancy. Written informed consent was obtained from all participants. The study protocol adhered to the Declaration of Helsinki (World Medical Association, 2013) and was approved by the local ethics committee (KEK-BE 2018-02164). Out of 42 participants, 31 data sets were included in the analyses. Reasons for exclusion were withdrawal of consent (n=2), cancellation of the second session due to the COVID-19 pandemic (n=3), technical/language issues (n=2), insufficient task performance with at least one of the task conditions never performed correctly (n=2), and excessive motion in the scanner (n=2). Demographic characteristics and task performance in fast conditions are provided in table 1.

### Image acquisition

Participants underwent two imaging sessions that were scheduled three weeks apart at the same hour of the day. At both sessions, we acquired structural and functional neuroimaging data at the Translational Imaging Center Bern of sitem-insel Bern on a 3T Magnetom Prisma scanner (Siemens Healthcare, Erlangen, Germany). First, we acquired structural T1-weighted images (MP2RAGE, 176 slices, FOV 240 x 256 mm, voxelsize 1x1x1mm, TR=5000ms, TE=2.98ms, flip angles=4°/5°) and then task-based fMRI (mbep2d, 660 volumes, covering 11

minutes, 72 slices, FOV 230x230mm, voxelsize 2.5x2.5x2.5mm, TR=1000ms, TE=37ms, flip angle=30°).

### fMRI task

The task was in a block design with 5 repetitions of 4 movement conditions, with a fixed duration of 17 seconds for each block. Active conditions were separated by two different control conditions of random length between 12-17 seconds. The order of active and control conditions remained consistent across all repetitions and sessions. Subjects performed all tasks with the right hand. Participants were instructed verbally before the scans and written cues were displayed via a projector at the beginning of each condition.

The four active conditions consisted of (i) sound-paced Thumb-Index Finger tapping (TIF) at .5 Hz; (ii) unpaced, as fast as possible Thumb-Index Finger tapping (TIFfast); (iii) paced Thumb-Alternating Finger opposition (TAF) at .5 Hz; and (iv) unpaced, as fast as possible Thumb-Alternating Finger opposition (TAFfast). The rest conditions following paced active conditions were combined with the pacing sound and the instruction to listen but not move. Runs were separated by a short break with a length between 6-12 seconds. When no instructions were displayed, a fixation cross was presented in all conditions. See figure 1 for a schematic depiction of the task design. Stimuli were presented and onsets of conditions logged using E-Prime (Version 2.0.10 Psychology Software Tools, Pittsburgh, PA, USA). Sounds were delivered via MR-safe headphones. We videotaped participants' hands during the task and verified the correct execution of each condition. Additionally, to evaluate the reliability of motor performance, we counted number of taps/oppositions for the fast movement conditions.

## Preprocessing and first-level analysis

Preprocessing was performed in SMP12 (Revision 7771, Welcome Trust, London, U.K., https://www.fil.ion.ucl.ac.uk/spm/) and MATLAB (R2020b, MathWorks, Natick, USA) and was identical for both sessions. The MP2RAGE sequence acquires images at two inversion times and calculates a unified resulting image with higher cerebral tissue contrast but increased extracerebral noise that may interfere with segmentation and co-registration (Marques et al., 2010). Therefore, we masked the unified image with the thresholded second inversion image to suppress the background. Then, we applied segmentation and normalization to MNI space within CAT12 (Christian Gaser 2018, http://www.neuro.uni-jena.de/cat/), and smoothing with a 5 mm FWHM kernel to structural images. Functional images were realigned, co-registered to the corresponding structural image, normalized using the DARTEL (Ashburner, 2007) approach and smoothed using a 5mm FWHM kernel. Subjects with mean framewise displacement >.5 mm or displacement >2.5 mm in one of the three translations or >2.5° in one of the three rotations were excluded from the analysis.

We built subject-wise first-level models in SPM12 with one regressor for each of the conditions (4 movement conditions and two rest conditions), as well as regressors for each of the three translations and three rotations from realignment as covariates. We then contrasted beta-values of each of the four movement conditions with the corresponding rest condition (TIF – Listen; TIFfast – Rest; TAF – Listen; TAFfast – Rest), as well as all tapping – all resting conditions. The resulting beta-difference maps were the input of the group-level ICC-analyses, while we used the resulting t-maps for group-level overlap-analyses.

## Statistical analyses

To assess reliability of task performance we calculated the average number of taps per second for the unpaced (fast) conditions and conducted paired t-tests and intraclass correlation

coefficient ($ICC_{3,k}$) analyses between both sessions for these performance metrics in R (version 4.0.3, The R Foundation for Statistical Computing).

We applied several strategies to evaluate imaging reliability. First, we explored differences in activation amplitude between sessions. Second, we examined the overlap of significant activations across the sessions to evaluate consistency of their spatial distribution. Finally, we calculated ICCs to investigate consistency of activation amplitudes in three sets of ROIs. We modelled paired t-tests between sessions for each of the four imaging contrasts in SPM to evaluate whether there were significant differences in activations between sessions. To evaluate spatial similarity of activations, we calculated the Dice similarity coefficients (DSC) for each contrast. The DSC is a simple measure for the overlap of clusters and is defined as:

$$DSC = \frac{2 * |X \cap Y|}{|X| + |Y|}$$

where X and Y are the extent of each session's activations at a given threshold (Dice, 1945; Rombouts et al., 1997). Since the DSC is highly dependent on the chosen threshold (Duncan, Pattamadilok, Knierim, & Devlin, 2009; Fernandez et al., 2003), we performed these analyses with three different thresholds: First p = .05 to capture and compare as much activation as possible, then the two standard thresholds of SPM p = .001 and family-wise error corrected $p_{FWE}$ = .05 (~p = 4.7239e-07). We did not apply any cluster forming threshold. DSC provides information on spatial reliability of activations and is distinct from t-tests: DSCs compare extent and localization of significant activations between sessions, depicting spatial similarity of these activations. T-tests compare amplitudes of all (de)activations, including nonsignificant ones, depicting amplitude differences. It is important to note that incongruences in DSC-analyses do not necessarily relate to significant differences in t-tests.

Additionally, to assess reliability of amplitudes, we extracted contrast values in three different sets of regions of interest (ROIs) and calculated intraclass correlation coefficients $ICC_{(3,k)}$ between sessions. $ICC_{(3,k)}$ (hereafter ICC) is:

$$ICC_{(3,k)} = \frac{BMS - EMS}{BMS}$$

where BMS is between subject mean square and EMS is error mean square (Shrout & Fleiss, 1979). Therefore, the ICC depicts the proportion of true variance in the total variance. ICC-analysis has become a standard for several types of reliability analyses. While DSC allows examination of spatial distribution of two categories (activated, not activated), ICC allows examination of amplitude reliability in a region of interest (ROI).

ROIs can be the primary unit of analysis and are often defined a priori based on previous literature or anatomical regions. Another frequent use of ROIs is to define them based on significant clusters from a whole-brain analysis to examine correlations with a variable of interest. To account for these different modes of ROI construction, we examined three sets of ROIs: First, a set of spheres with 10mm radius centered on peaks reported in an activation likelihood estimation by Hardwick et al. (2013) (table S1, figure S1). Due to the proximity of bilateral M1 and S1 peaks, they share 40% of their volumes. To ensure consistent ROI creation, we did not modify these ROIs. Second, a set consisting of anatomical ROIs exported from the AAL-atlas (Tzourio-Mazoyer et al., 2002) (table S2, figure S2), and finally a functionally defined set for which we conducted a conjunction analysis of activations of both sessions for each tapping condition and defined significant clusters at a threshold of $p_{FWE} < .05$ as ROIs (table S3, figures S3-S7). Note that the three sets differ in shape, size and location of the ROIs despite similar naming (figure S8).

To assess the influence of the broad range of age on the observed reliability, we split the sample in half, doing a median split at 31 years and repeated all test-retest reliability analyses in both age groups. To our knowledge, there is no consensus regarding the interpretation of DSC. Therefore, we will apply the guidelines of Cicchetti (1994) to both, DSC and ICC values. Coefficients below .4 will be considered poor, between .4 and .59 fair, between .6 and .74 good, and >.75 excellent.

## Results

### Tapping performance and reliability

Tapping performance for the two unpaced (fast) conditions are provided in table 1. Participants tapped slightly faster in thumb-index finger tapping than in thumb-alternating finger opposition (Δ=.76, 95%-CI .40 - 1.11, $p < .001$). Paired t-tests of performance showed no significant improvement over time (TIFfast: mean difference = .14 Taps/s, $p = .12$; TAFfast: mean difference = .05 Taps/s, $p = .58$). Intraclass correlation coefficients indicated excellent reliability of performance in both conditions (TIFfast: ICC = .94, 95%-CI .90 - .97; TAFfast: ICC = .92, 95%-CI .86 - .96).

In the paced conditions, the number of trials excluded due to incorrect tapping were comparable for TIF (2.6%) and TAF (3.2%) ($X^2 = .1$, $p = .75$). For TAF, more trials were excluded at follow-up (5.8%) than at baseline (.01%) ($X^2 = 4.9$, $p = .027$). Similarly but statistically only at trend level, more TIF trials were excluded at follow-up (4.5%) than at baseline (.01%) ($X^2 = 3.13$, $p = .08$).

### Activations

All contrasts showed the expected activations in the motor network in response to right-hand finger tapping: Left primary motor and sensory cortices, premotor and supplementary motor areas, and bilateral cerebellum. Additionally, all contrasts but TIF showed activations

in left parietal and bilateral frontal cortices, as well as in subcortical structures, such as putamen or thalamus (see table S3 for clusters in conjunction analyses of each contrast).

### Imaging reliability

Paired t-tests of task fMRI activations showed no significant differences between sessions for any of the five contrasts at $p_{FWE} < .05$. However, at $p < .001$ and $p < .05$ we found clusters with higher activation at follow-up in bilateral precuneus for TIFfast, TAFfast, and all tapping vs. all rest. Additionally, three clusters in bilateral operculum and left cerebellum showed higher activation at baseline than at follow-up for TIF at $p < .05$.

DSC analyses yielded comparable coefficients across all thresholds and contrasts (figures 2-6). The overlap between the two sessions was good to excellent in all cases, except for TIF at the two lower thresholds (.001 and FWE-.05, Figure 2), which were in the fair range. The contrast all tapping vs. all rest yielded the highest DSC values for all three thresholds (Figure 6). The DSC values are provided in table 2a.

ICCs in literature based, 10 mm spherical ROIs showed a large range and variability (.04 - .91). The ROIs with at least fair ICC for all contrasts were bilateral primary motor and sensory cortices, SMA, and left putamen (table S1). There were no ROIs with poor ICC for all contrasts, but left cerebellum and bilateral thalamus had no good or excellent ICC in any of the contrasts.

The atlas-based ROIs showed a relatively large range of ICCs (.08 - .81). ROIs with at least fair reliability in all contrasts were bilateral primary sensory cortex and cerebellum, left putamen, and right SMA. Bilateral primary motor cortices showed poor ICCs only for the paced thumb-alternating finger contrast (table S2). Again, no ROI had poor reliability in all contrasts, but bilateral thalamus, left SPL, and right putamen had no good or excellent ICC in any of the conditions.

Finally, ICC in ROIs based on conjunction analyses showed a similar variability with most values in fair, good and excellent ranges. Range of ROI size was immense with 7 to 4451 voxels, as was range of ICC with .13 - .74. The all-tapping vs. all-rest contrast was the only one with at least fair ICCs in all ROIs (table S3).

The average ICCs were in the fair range for paced and unpaced TIF and TAF contrasts for all three ROI-sets, except TIFfast (good) and TAF (poor) in the atlas-based ROIs, while the all-tapping vs. all-rest contrast yielded average ICCs in the good range in all three ROI-sets (table 2b).

### Age groups

Characteristics and tapping performance of age groups are provided in supplementary table S4. They did not differ in sex ($X^2$ = .8; p = .37), education (t = .97; p = .34), or tapping performance (all p > .55). Both groups tapped slightly faster in thumb-index finger tapping than in thumb-alternating finger opposition (Δyoung = .82; p < .001, Δold = .69; p < .001), but paired t-tests of performance showed no significant difference between the sessions in the younger or older half of the sample (all p > .18). Intraclass correlation coefficients indicated excellent reliability of performance in both conditions in both groups (all ICC ≥ .92).

For the two more liberal thresholds (p < .05 and p < .001), overlap was in the fair-to-good range in both groups for all conditions but TIF. The younger group had poor overlap for this contrast at all thresholds. The older half of our sample showed numerically higher overlap in all cases but the most liberal threshold (p < .05) in the all tapping vs. all rest condition. We noticed a sharp drop in overlaps between p < .001 and $p_{FWE}$ < .05 in all conditions for both age groups. Comparison of all DSC between the younger and older half of the sample using a

Mann-Whitney-U-test showed no significant difference. Since there was a sharp drop of coefficients at $p_{FWE} < .05$, we also compared the DSC for the two more liberal thresholds between the age groups and again, found no significant difference. Dice coefficients per age group and condition are provided in supplementary table S5 and supplementary figure S8.

Separate ICC analyses in the age groups showed averages of ICCs in the fair-to-good range for most conditions using the literature-based or the conjunction-based ROIs, regardless of age group (supplementary table S6). The anatomical, atlas-based ROIs had averages of ICCs in the poor range for three conditions in the younger half of the sample. Again, in most direct comparisons, the older half of the sample showed numerically higher reliability than the younger half. Additionally, we compared the ICCs per age group, condition and ROI category using two-sample t-tests. In 9 out of 18 comparisons, ICCs were significantly higher in the older half of the sample, while the younger half had higher ICCs in only one comparison. There was no significant difference in the remaining 8 comparisons (see supplementary table S6).

## Discussion

In the present study, we evaluated the test-retest reliability of fMRI derived brain activations for four simple motor tasks in a right-handed healthy population. We found good reliability regarding spatial distribution and satisfactory reliability for amplitudes of activations on group level.

Regarding task performance, participants showed no significant improvement across the two sessions in both unpaced movement conditions. Therefore, we may assume that there was no relevant training effect. This is in line with literature; although within and between session training effects have been shown for an intersession interval of 24h, no training effect was

observed at an interval of two weeks (Nguemeni, Stiehl, Hiew, & Zeller, 2021; Sardroodian, Madeleine, Mora-Jensen, & Hansen, 2016). Reliability of tapping performance of both unpaced conditions was excellent with ICCs of >.9, demonstrating behavioral robustness of the motor tasks themselves. The increased number of trials with incorrect paced tapping might hint at a reduction in motivation or attention at follow-up compared with baseline.

As expected, we detected activations in left primary motor and sensory cortex, left premotor cortex, left SMA and right cerebellum for all tasks. Again, this is in line with literature (see (Witt et al., 2008) for an ALE meta-analysis). In the relatively more demanding conditions, additional brain regions were recruited, i.e. alternating finger opposition evoked activity in more regions than index finger tapping and unpaced, fast tapping recruited more regions than paced, slower tapping. Furthermore, clusters of activated voxels tended to be larger in conditions that are more demanding, reflecting the increased need of neural resources for these task conditions (Goble et al., 2010; Van Impe et al., 2013). The higher signal at follow-up in precuneus during the fast conditions might actually represent a weaker deactivation of the default mode network that is associated with mind wandering (Buckner, Andrews-Hanna, & Schacter, 2008; Fox, Spreng, Ellamil, Andrews-Hanna, & Christoff, 2015), possibly reflecting a reduction of focus at follow-up.

We found good to excellent spatial activation overlap in all five contrasts with little variance across all three tested statistical thresholds for activation maps ($p<.05$; $p<.001$; $p_{FWE}<.05$), as demonstrated by the Dice similarity coefficients. This demonstrates reliable spatial identification of activated voxels in response to finger tapping at the most commonly applied thresholds. The all-tapping vs. all-rest contrast had the largest overlap at all three thresholds, but the differences were relatively small. This indicates that fMRI can reliably identify the

brain regions activated in response to these tasks at group level. Good test-retest spatial overlap has been reported for several task designs, but the range of reported overlaps was immense even among finger tapping tasks (Bennett & Miller, 2010; Gountouna et al., 2010; Ibinson et al., 2022; Yetkin, McAuliffe, Cox, & Haughton, 1996). The larger overlap of the fast, unpaced compared with the paced contrasts might reflect the behavioral performance: Most of the volume of non-overlap for the paced contrasts consisted of activations at baseline that were missing at follow-up, possibly paralleling the higher number of correct runs at baseline for these contrasts. Conversely, the non-overlap for the fast contrasts included more activation only during follow-up. We found no significant difference in tapping performance in the fast conditions, but there was a subtle numerical increase of tapping speed at follow-up. Moreover, the larger volume of activation in the more demanding fast conditions might represent recruitment of a higher proportion of available neural resources, resulting in a smaller volume for potential non-overlap.

Similarly, ROI-based analyses of ICC showed a large span of reliability of activation amplitude across ROIs. Interestingly, we found ICCs >.4 in most ROIs independent of the mode of ROI selection. This was unexpected, as the sets of ROIs differed in shape, size and location of the ROIs with potentially little overlap between them (see supplementary figure S8). However, average ICC per contrast was only in the fair range in conjunction and literature based ROIs. In the anatomically defined ROIs, TAF had poor average ICC, whereas TIFfast was in the good range. Notably, the all tapping vs. all rest contrast showed the least amount of variability with all but one ROIs having at least fair reliability, and average ICCs being in the good range for all three methods of ROI definition. These results indicate that it is possible to associate amplitudes of activations with other variables. Similar to the reports on spatial overlap, reported studies on activation amplitudes using test-retest ICC show large variability (Aron,

Gluck, & Poldrack, 2006; Bennett & Miller, 2010; Friedman et al., 2008). Generally, motor tasks tend to yield higher reliability than cognitive tasks (Bennett & Miller, 2010; Fliessbach et al., 2010). However, Havel et al. (2006) reported hand movements to have higher reliability than movement in other anatomical regions, pointing to differences even between motor tasks. Regarding the variability within the ROI analyses, S1 and M1 bilaterally seem to have superior ICCs across most contrasts, while there is no detectable pattern in the other regions. We suggest that M1 and S1 are consistently recruited during the finger tapping tasks and therefore achieve higher ICCs than areas with different functional specialization. It remains to be established whether studies in much larger samples would detect interpretable patterns of ICC distribution.

In both, the spatial overlap and ICC analyses, the all-tapping vs. all-rest contrast had the highest reliability. However, this is also the least specific contrast. In the present case, the higher number of trials and the longer acquisition period included in the more general contrast may have led to statistically more robust but less specific responses (Gordon et al., 2017). Moreover, outlier responses to specific tasks loose impact through averaging across several subjects and trials. Extending this notion in the opposite direction, this may explain the low subject-wise reliability that recent studies reported (Elliott et al., 2020), as the total number of trials in a single subject is usually substantially lower than the number of trials in a whole group of subjects. Friedman et al. (2008) previously demonstrated this relationship of number of trials and reliability on the group-level in a finger tapping task. It is important to note that we aimed to examine group-level reliability in the present study. This is reflected by the design of our task that allows for a maximum of five trials per tapping condition. Moreover, high group-level reliability does not necessitate high subject-level reliability and vice-versa (Frohner, Teckentrup, Smolka, & Kroemer, 2019; Gordon et al., 2017). However,

examination of single subject test-retest reliability may inform the interpretation of group-level reliability. For example, high group reliability with low subject reliability would argue for a population tendency of a state that is unstable in the individual, while low group but high individual reliability could reflect heterogeneity in stable individual traits.

Various factors other than acquisition duration and number of trials have been reported to increase reliability of task-based fMRI: shorter between-sessions interval, block-design had higher reliability than event-related, cortical activations had higher reliability than subcortical ones, healthy populations generated more robust results than patients, but evidence on the effect of these factors is conflicting (Bennett & Miller, 2010, 2013; Elliott et al., 2020). Moreover, we found numerically higher reliability in the older half of the study sample compared with the younger half for both, overlap and amplitudes of activations. However, this did not pertain to the motor behavior itself. Larger between-subject variability may have increased ICCs in the older half and may have had a smoothing effect in the overlap analyses.

For research contexts, reliability should not be examined in isolation, as larger effect sizes as well as larger sample size can enhance the detection of effects even with less reliable measures. Moreover, there are sources of uncertainty beside the reliability of the BOLD signal in fMRI: Even when evaluating the same set of images, there is substantial variability depending on the choice of toolbox for the analyses, the preprocessing, models, and even operating systems, computers and versions of the toolboxes (Bowring, Maumet, & Nichols, 2019; Carp, 2012; Pauli et al., 2016).

Some limitations require consideration for this study. First, our sample size is limited. However, it is in the range of typical fMRI studies. The limited sample size prevented more in-depth investigation of possible age effects and our median split resulted in age groups with

vastly different age ranges, as the younger half spanned 12 years, while the older half spanned 24 years. Second, we examined reliability between only two sessions. Reliability between multiple sessions might differ from the one observed here. Third, we had a limited number of trials per condition, as the task consisted of five runs and some trials were excluded due to incorrect tapping. Moreover, since our sessions took place three weeks apart, the female participants in the present sample were probably in different phases of their menstrual cycle in the two sessions. Effects of the menstrual cycle on brain networks, including the somatomotor network, and motor behavior have been demonstrated (Bayer & Hausmann, 2012; Pellegrini, Zoghi, & Jaberzadeh, 2018; Pritschet et al., 2020). Finally, there is an unknown amount of true variability that is unrelated to the measure. The true neural response to the task may vary due to subject and session specific variables, such as participants being tired or varying concentration and motivation. In fact, the increased number of incorrect paced trials and the higher signal in precuneus during unpaced trials at follow-up are suggestive of differences in focus between the sessions.

## Conclusion

In sum, the presented tapping tasks can reliably identify brain regions that are activated in response to the task. Test-retest reliability was good in spatial and fair in amplitude domain on group level. Subject and group level reliability are distinct properties of a task and task design should reflect the level of intended analyses (i.e. subject vs. group). Although the reliability of the amplitudes was often only in the fair range and caution is warranted when examining correlations of activations with other variables, identification of activated regions in response to a task and their comparisons between groups are still a valid and important mode of analysis in neuroimaging to find population tendencies and differences.

## Data Availability Statement

The data used in this study are available from SW upon reasonable request.

## Ethics Statement

Written informed consent was obtained from all participants. The study protocol adhered to the Declaration of Helsinki and was approved by the local ethics committee.


## Funding

This study was supported by the Swiss National Science Foundation (#182469 to SW) and the National Institute of Mental Health (#R01 MH118741 to SAS, VAM, and SW).


## Conflict of interest

SW has received honoraria from Janssen, Lundbeck, Mepha, Neurolite, and Sunovion. All other authors report no conflicts of interest.


Aron, A. R., Gluck, M. A., & Poldrack, R. A. (2006). Long-term test-retest reliability of functional MRI in a classification learning task. *Neuroimage, **29***(3), 1000-1006. doi:10.1016/j.neuroimage.2005.08.010

Ashburner, J. (2007). A fast diffeomorphic image registration algorithm. *Neuroimage, **38***(1), 95-113. doi:10.1016/j.neuroimage.2007.07.007

Bayer, U., & Hausmann, M. (2012). Menstrual cycle-related changes of functional cerebral asymmetries in fine motor coordination. *Brain Cogn, **79***(1), 34-38. doi:10.1016/j.bandc.2012.02.003

Bennett, C. M., & Miller, M. B. (2010). How reliable are the results from functional magnetic resonance imaging? *Ann N Y Acad Sci, **1191***, 133-155. doi:10.1111/j.1749-6632.2010.05446.x

Bennett, C. M., & Miller, M. B. (2013). fMRI reliability: influences of task and experimental design. *Cogn Affect Behav Neurosci, **13***(4), 690-702. doi:10.3758/s13415-013-0195-1

Bowring, A., Maumet, C., & Nichols, T. E. (2019). Exploring the impact of analysis software on task fMRI results. *Hum Brain Mapp, **40***(11), 3362-3384. doi:10.1002/hbm.24603

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: anatomy, function, and relevance to disease. *Ann N Y Acad Sci, **1124***, 1-38. doi:10.1196/annals.1440.011

Carp, J. (2012). On the plurality of (methodological) worlds: estimating the analytic flexibility of FMRI experiments. *Front Neurosci, **6***, 149. doi:10.3389/fnins.2012.00149

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, **6***(4), 284-290. doi:10.1037/1040-3590.6.4.284

Cohen, M. S., & DuBois, R. M. (1999). Stability, repeatability, and the expression of signal magnitude in functional magnetic resonance imaging. *J Magn Reson Imaging, **10***(1), 33-40. doi:10.1002/(sici)1522-2586(199907)10:1<33::aid-jmri5>3.0.co;2-n

Dice, L. R. (1945). Measures of the Amount of Ecologic Association Between Species. *Ecology, **26***(3), 297-302. doi:10.2307/1932409

Dudek, E., & Dodell-Feder, D. (2021). The efficacy of real-time functional magnetic resonance imaging neurofeedback for psychiatric illness: A meta-analysis of brain and behavioral outcomes. *Neurosci Biobehav Rev, **121***, 291-306. doi:10.1016/j.neubiorev.2020.12.020

Duncan, K. J., Pattamadilok, C., Knierim, I., & Devlin, J. T. (2009). Consistency and variability in functional localisers. *Neuroimage, **46***(4), 1018-1026. doi:10.1016/j.neuroimage.2009.03.014

Elliott, M., Knodt, A., Ireland, D., Morris, M., Poulton, R., Ramrakha, S., . . . Hariri, A. (2020). What is the Test-Retest Reliability of Common Task-fMRI Measures? New Empirical Evidence and a Meta-Analysis. *Psychological Science, **31***(7), 792-806. doi:10.1177/0956797620916786

Estevez, N., Yu, N., Brugger, M., Villiger, M., Hepp-Reymond, M. C., Riener, R., & Kollias, S. (2014). A reliability study on brain activation during active and passive arm movements supported by an MRI-compatible robot. *Brain Topogr, **27***(6), 731-746. doi:10.1007/s10548-014-0355-9

Fernandez, G., Specht, K., Weis, S., Tendolkar, I., Reuber, M., Fell, J., . . . Elger, C. E. (2003). Intrasubject reproducibility of presurgical language lateralization and mapping using fMRI. *Neurology, **60***(6), 969-975. doi:10.1212/01.wnl.0000049934.34209.2e

Fliessbach, K., Rohe, T., Linder, N. S., Trautner, P., Elger, C. E., & Weber, B. (2010). Retest reliability of reward-related BOLD signals. *Neuroimage, **50***(3), 1168-1176. doi:10.1016/j.neuroimage.2010.01.036

Fox, K. C., Spreng, R. N., Ellamil, M., Andrews-Hanna, J. R., & Christoff, K. (2015). The wandering brain: meta-analysis of functional neuroimaging studies of mind-wandering and related spontaneous thought processes. *Neuroimage, **111***, 611-621. doi:10.1016/j.neuroimage.2015.02.039

Friedman, L., Stern, H., Brown, G. G., Mathalon, D. H., Turner, J., Glover, G. H., . . . Potkin, S. G. (2008). Test-retest and between-site reliability in a multicenter fMRI study. *Hum Brain Mapp, **29***(8), 958-972. doi:10.1002/hbm.20440

Frohner, J. H., Teckentrup, V., Smolka, M. N., & Kroemer, N. B. (2019). Addressing the reliability fallacy in fMRI: Similar group effects may arise from unreliable individual effects. *Neuroimage, **195***, 174-189. doi:10.1016/j.neuroimage.2019.03.053

Goble, D. J., Coxon, J. P., Van Impe, A., De Vos, J., Wenderoth, N., & Swinnen, S. P. (2010). The neural control of bimanual movements in the elderly: Brain regions exhibiting age-related increases in activity, frequency-induced neural modulation, and task-specific compensatory recruitment. *Hum Brain Mapp, **31***(8), 1281-1295. doi:10.1002/hbm.20943

Gordon, E. M., Laumann, T. O., Gilmore, A. W., Newbold, D. J., Greene, D. J., Berg, J. J., . . . Dosenbach, N. U. F. (2017). Precision Functional Mapping of Individual Human Brains. *Neuron, **95***(4), 791-807 e797. doi:10.1016/j.neuron.2017.07.011

Gountouna, V. E., Job, D. E., McIntosh, A. M., Moorhead, T. W., Lymer, G. K., Whalley, H. C., . . . Lawrie, S. M. (2010). Functional Magnetic Resonance Imaging (fMRI) reproducibility and variance components across visits and scanning sites with a finger tapping task. *Neuroimage, **49***(1), 552-560. doi:10.1016/j.neuroimage.2009.07.026

Hardwick, R. M., Rottschy, C., Miall, R. C., & Eickhoff, S. B. (2013). A quantitative meta-analysis and review of motor learning in the human brain. *Neuroimage, **67***, 283-297. doi:10.1016/j.neuroimage.2012.11.020

Havel, P., Braun, B., Rau, S., Tonn, J. C., Fesl, G., Bruckmann, H., & Ilmberger, J. (2006). Reproducibility of activation in four motor paradigms. An fMRI study. *J Neurol,* ***253***(4), 471-476. doi:10.1007/s00415-005-0028-4

Ibinson, J. W., Gillman, A. G., Schmidthorst, V., Li, C., Napadow, V., Loggia, M. L., & Wasan, A. D. (2022). Comparison of test-retest reliability of BOLD and pCASL fMRI in a two-center study. *BMC Med Imaging,* ***22***(1), 62. doi:10.1186/s12880-022-00791-9

Jaccard, P. (1912). THE DISTRIBUTION OF THE FLORA IN THE ALPINE ZONE.1. *New Phytologist,* ***11***(2), 37-50. doi:https://doi.org/10.1111/j.1469-8137.1912.tb05611.x

Jalilianhasanpour, R., Beheshtian, E., Ryan, D., Luna, L. P., Agarwal, S., Pillai, J. J., . . . Gujar, S. K. (2021). Role of Functional Magnetic Resonance Imaging in the Presurgical Mapping of Brain Tumors. *Radiol Clin North Am,* ***59***(3), 377-393. doi:10.1016/j.rcl.2021.02.001

Kimberley, T. J., Birkholz, D. D., Hancock, R. A., VonBank, S. M., & Werth, T. N. (2008). Reliability of fMRI during a continuous motor task: assessment of analysis techniques. *J Neuroimaging,* ***18***(1), 18-27. doi:10.1111/j.1552-6569.2007.00163.x

Lee, J. N., Hsu, E. W., Rashkin, E., Thatcher, J. W., Kreitschitz, S., Gale, P., . . . Marchand, W. R. (2010). Reliability of fMRI motor tasks in structures of the corticostriatal circuitry: implications for future studies and circuit function. *Neuroimage,* ***49***(2), 1282-1288. doi:10.1016/j.neuroimage.2009.09.072

Manan, H. A., Franz, E. A., & Yahya, N. (2021). The utilisation of resting-state fMRI as a pre-operative mapping tool in patients with brain tumours in comparison to task-based fMRI and intraoperative mapping: A systematic review. *Eur J Cancer Care (Engl),* ***30***(4), e13428. doi:10.1111/ecc.13428

Marques, J. P., Kober, T., Krueger, G., van der Zwaag, W., Van de Moortele, P. F., & Gruetter, R. (2010). MP2RAGE, a self bias-field corrected sequence for improved segmentation and T1-mapping at high field. *Neuroimage,* ***49***(2), 1271-1281. doi:10.1016/j.neuroimage.2009.10.002

Marshall, I., Simonotto, E., Deary, I. J., Maclullich, A., Ebmeier, K. P., Rose, E. J., . . . Chappell, F. M. (2004). Repeatability of motor and working-memory tasks in healthy older volunteers: assessment at functional MR imaging. *Radiology,* ***233***(3), 868-877. doi:10.1148/radiol.2333031782

Nguemeni, C., Stiehl, A., Hiew, S., & Zeller, D. (2021). No Impact of Cerebellar Anodal Transcranial Direct Current Stimulation at Three Different Timings on Motor Learning in a Sequential Finger-Tapping Task. *Front Hum Neurosci,* ***15***, 631517. doi:10.3389/fnhum.2021.631517

Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *Proc Natl Acad Sci U S A,* ***87***(24), 9868-9872. doi:10.1073/pnas.87.24.9868

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

Pauli, R., Bowring, A., Reynolds, R., Chen, G., Nichols, T. E., & Maumet, C. (2016). Exploring fMRI Results Space: 31 Variants of an fMRI Analysis in AFNI, FSL, and SPM. *Front Neuroinform,* ***10***, 24. doi:10.3389/fninf.2016.00024

Pellegrini, M., Zoghi, M., & Jaberzadeh, S. (2018). Biological and anatomical factors influencing interindividual variability to noninvasive brain stimulation of the primary motor cortex: a systematic review and meta-analysis. *Rev Neurosci,* ***29***(2), 199-222. doi:10.1515/revneuro-2017-0048

Pritschet, L., Santander, T., Taylor, C. M., Layher, E., Yu, S., Miller, M. B., . . . Jacobs, E. G. (2020). Functional reorganization of brain networks across the human menstrual cycle. *Neuroimage,* ***220***, 117091. doi:10.1016/j.neuroimage.2020.117091

Rath, J., Wurnig, M., Fischmeister, F., Klinger, N., Hollinger, I., Geissler, A., . . . Beisteiner, R. (2016). Between- and within-site variability of fMRI localizations. *Hum Brain Mapp,* ***37***(6), 2151-2160. doi:10.1002/hbm.23162

Rombouts, S. A., Barkhof, F., Hoogenraad, F. G., Sprenger, M., Valk, J., & Scheltens, P. (1997). Test-retest analysis with functional MR of the activated area in the human visual cortex. *AJNR Am J Neuroradiol,* ***18***(7), 1317-1322. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/9282862

Sadraee, A., Paulus, M., & Ekhtiari, H. (2021). fMRI as an outcome measure in clinical trials: A systematic review in clinicaltrials.gov. *Brain Behav,* ***11***(5), e02089. doi:10.1002/brb3.2089

Sardroodian, M., Madeleine, P., Mora-Jensen, M. H., & Hansen, E. A. (2016). Characteristics of Finger Tapping Are Not Affected by Heavy Strength Training. *J Mot Behav,* ***48***(3), 256-263. doi:10.1080/00222895.2015.1089832

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass Correlations - Uses in Assessing Rater Reliability. *Psychological Bulletin,* ***86***(2), 420-428.

Thibault, R. T., MacPherson, A., Lifshitz, M., Roth, R. R., & Raz, A. (2018). Neurofeedback with fMRI: A critical systematic review. *Neuroimage,* ***172***, 786-807. doi:10.1016/j.neuroimage.2017.12.071

Tzourio-Mazoyer, N., Landeau, B., Papathanassiou, D., Crivello, F., Etard, O., Delcroix, N., . . . Joliot, M. (2002). Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage,* ***15***(1), 273-289. doi:10.1006/nimg.2001.0978

Van Impe, A., Bruijn, S. M., Coxon, J. P., Wenderoth, N., Sunaert, S., Duysens, J., & Swinnen, S. P. (2013). Age-related neural correlates of cognitive task performance under increased postural load. *Age (Dordr),* ***35***(6), 2111-2124. doi:10.1007/s11357-012-9499-2

Witt, S. T., Laird, A. R., & Meyerand, M. E. (2008). Functional neuroimaging correlates of finger-tapping task variations: an ALE meta-analysis. *Neuroimage,* ***42***(1), 343-356. doi:10.1016/j.neuroimage.2008.04.025

World Medical Association. (2013). World Medical Association Declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA,* ***310***(20), 2191-2194. doi:10.1001/jama.2013.281053

Yetkin, F. Z., McAuliffe, T. L., Cox, R., & Haughton, V. M. (1996). Test-retest precision of functional MR in sensory and motor task activation. *AJNR Am J Neuroradiol,* ***17***(1), 95-98. Retrieved from https://www.ncbi.nlm.nih.gov/pubmed/8770256

Yoo, S. S., Wei, X., Dickey, C. C., Guttmann, C. R., & Panych, L. P. (2005). Long-term reproducibility analysis of fMRI using hand motor task. *Int J Neurosci,* ***115***(1), 55-77. doi:10.1080/00207450490512650

| Table 1 – Sample characteristics and task performance | | |
|---|---|---|
| | Baseline (Mean ± SD) | Follow-Up (Mean ± SD) |
| Age (years) | 35.7 ± 12.2 | |
| Sex (n, % female) | 16 (51.6) | |
| Education (years) | 16.2 ± 3.3 | |
| TIFfast performance (Taps/second) | 3.71 ± 1.15 | 3.86 ± .95 |
| TAFfast performance (Taps/second) | 3.00 ± .91 | 3.05 ± .86 |
| TIF: Paced thumb-index finger tapping; TAF: paced thumb alternating finger opposition; TIF/TAFfast: unpaced condition with movement as fast as possible | | |

| Table 2a – Dice Similarity Coefficients (DSC) | | | | | |
|---|---|---|---|---|---|
| Contrast | TIF | TIFfast | TAF | TAFfast | AllC |
| Threshold p=.05 | .624 | .764 | .652 | .750 | .781 |
| Threshold p=.001 | .569 | .747 | .642 | .778 | .819 |
| Threshold $p_{FWE}$=.05 | .543 | .605 | .683 | .706 | .710 |
| TIF: Paced thumb-index finger tapping; TAF: paced thumb alternating finger opposition; TIF/TAFfast: unpaced condition with movement as fast as possible; FWE: family-wise error rate. | | | | | |

| Table 2b – Average Intraclass Correlation Coefficients (ICC) per contrast and ROI-set | | | | | | |
|---|---|---|---|---|---|---|
| | TIF | TIFfast | TAF | TAFfast | AllC | Average across contrasts |
| Literature | .48 ± .15 | .54 ± .21 | .45 ± .18 | .54 ± .19 | .61 ± .14 | .52 ± .06 |
| Atlas | .53 ± .08 | .62 ± .10 | .33 ± .17 | .52 ± .14 | .63 ± .11 | .53 ± .12 |
| Conjunction | .48 ± .18 | .54 ± .18 | .44 ± .14 | .49 ± .16 | .60 ± .10 | .51 ± .06 |
| Mean ± SD of ICCs per tapping contrast and ROI set. TIF: Paced thumb-index finger tapping; TAF: paced thumb alternating finger opposition; TIF/TAFfast: unpaced condition with movement as fast as possible. | | | | | | |

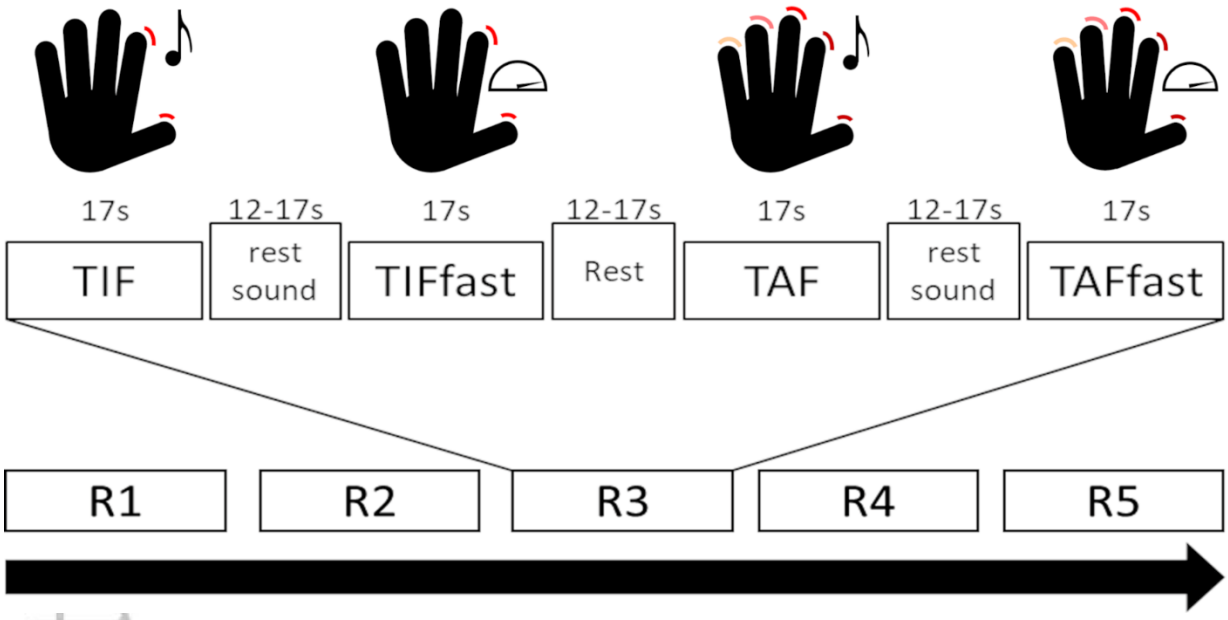


Figure 1: Schematic depiction of the examined finger-tapping task. TIF: Paced thumb-index finger tapping; TAF: paced thumb alternating finger opposition; TIFfast/TAFfast: unpaced condition with movement as fast as possible; R1 – R5: runs (repetitions). Mean difference of active conditions and their respective rest condition was used for group analyses.

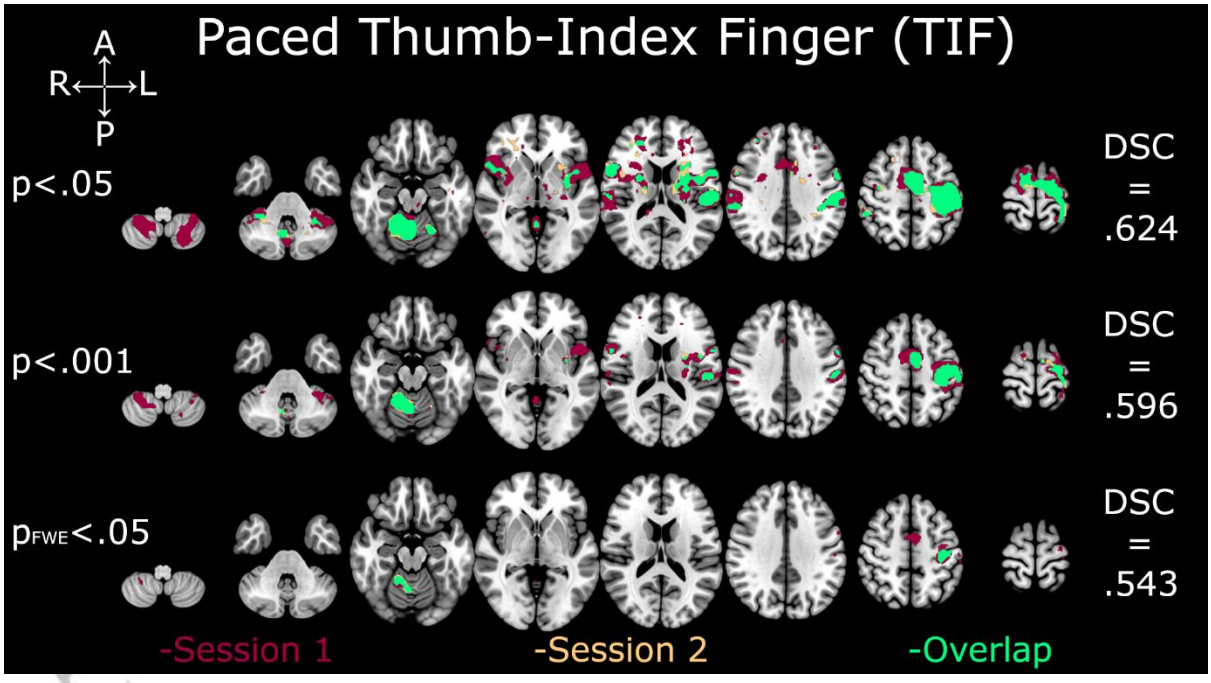


Figure 2: Depiction of Dice Similarity Coefficients for paced Thumb-Index Finger tapping (TIF). Single sessions and overlapping activation of TIF at the three examined thresholds. DSC: Dice Similarity Coefficient; FWE: Family-Wise Error.

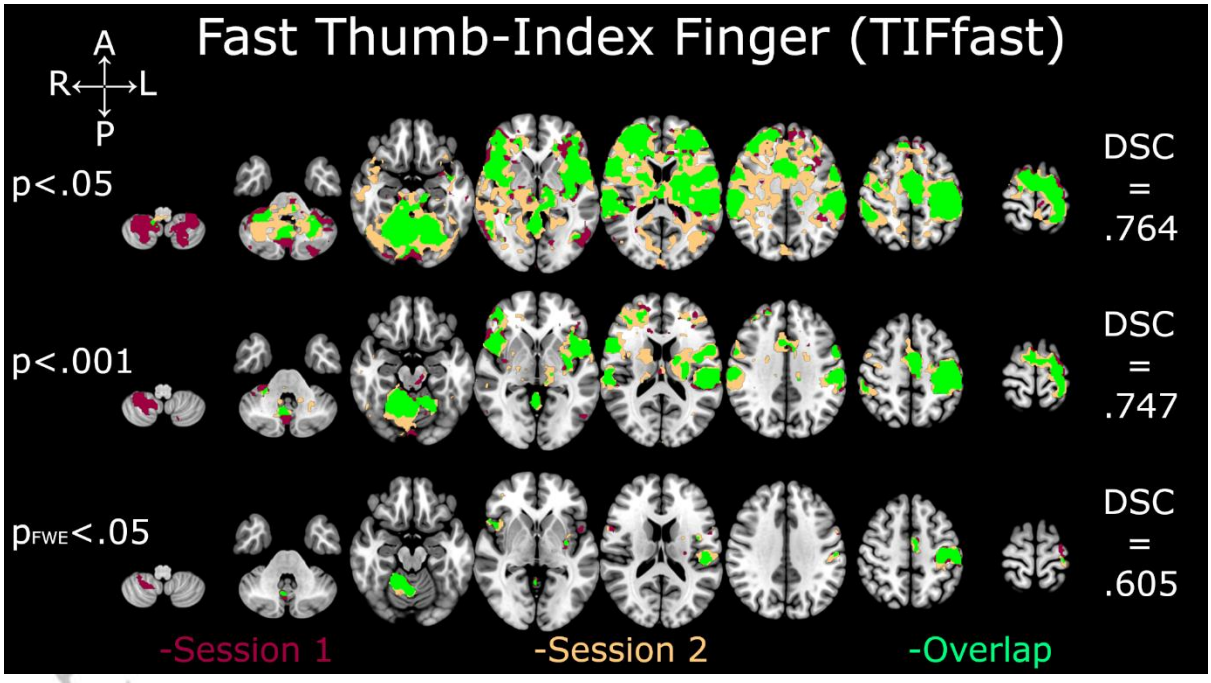


Figure 3: Depiction of Dice Similarity Coefficients for unpaced Thumb-Index Finger tapping (TIFfast). Single sessions and overlapping activation of TIFfast at the three examined thresholds. DSC: Dice Similarity Coefficient; FWE: Family-Wise Error.

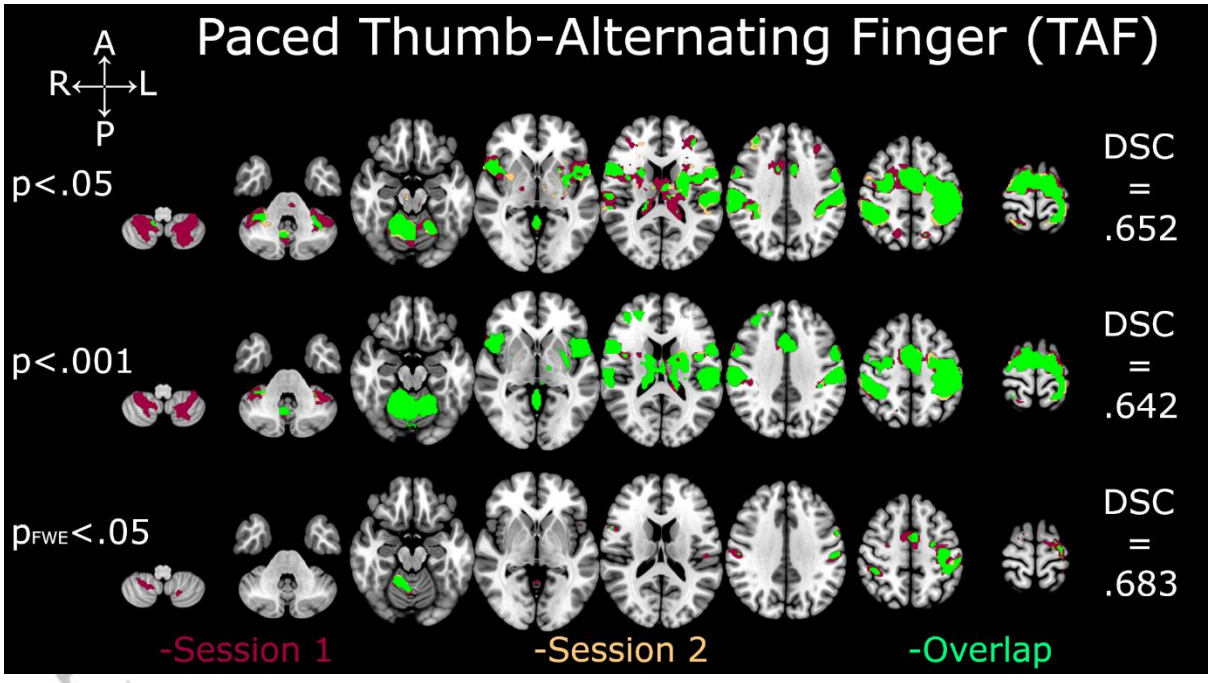


Figure 4: Depiction of Dice Similarity Coefficients for paced Thumb-Alternating Finger opposition (TAF). Single sessions and overlapping activation of TAF at the three examined thresholds. DSC: Dice Similarity Coefficient; FWE: Family-Wise Error.

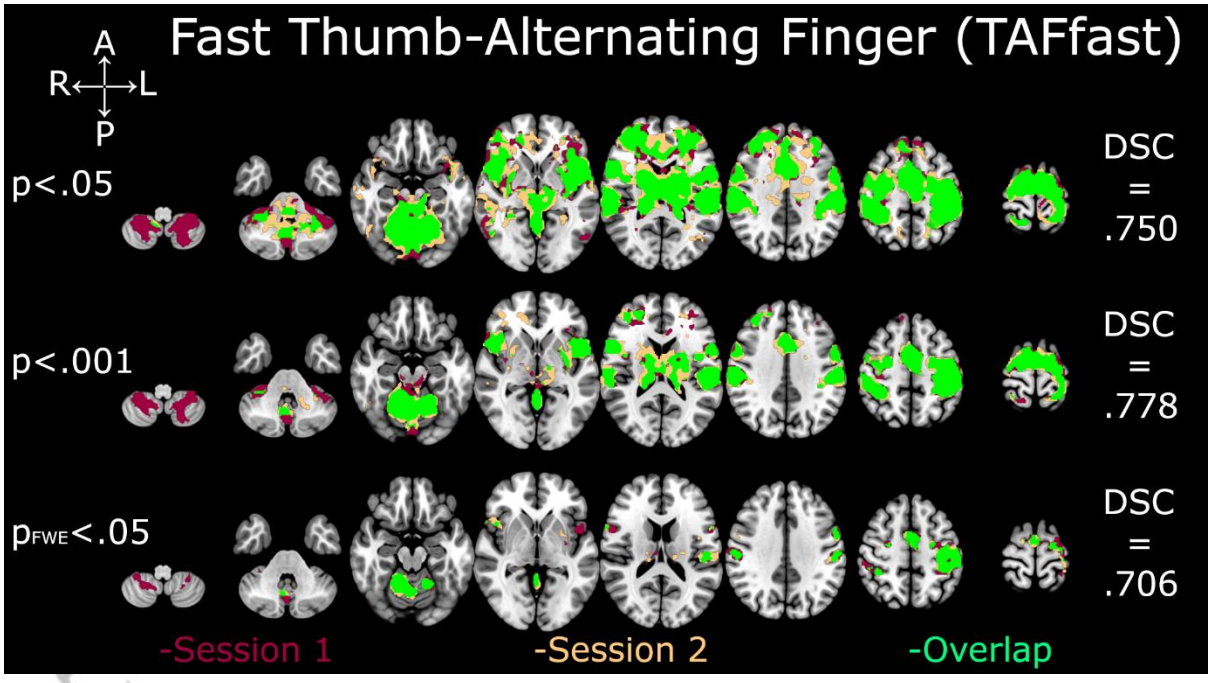


Figure 5: Depiction of Dice Similarity Coefficients for unpaced Thumb-Alternating Finger opposition (TAFfast). Single sessions and overlapping activation of TAFfast at the three examined thresholds. DSC: Dice Similarity Coefficient; FWE: Family-Wise Error.

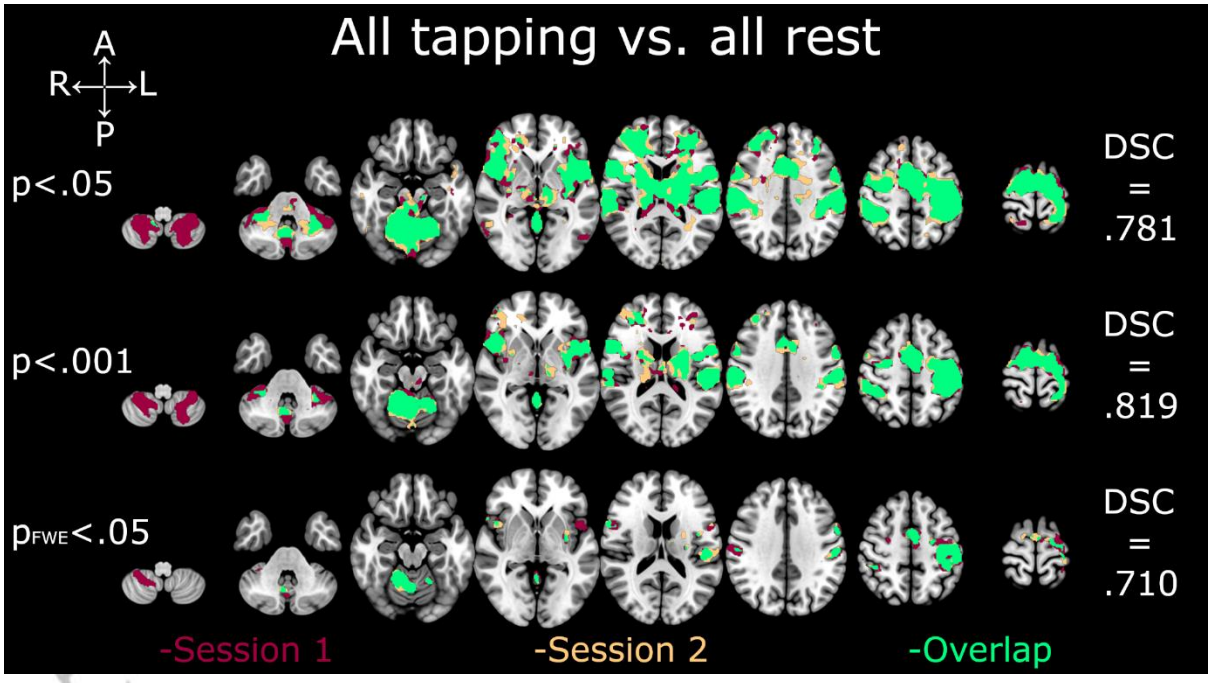


Figure 6: Depiction of Dice Similarity Coefficients for all tapping conditions vs. all rest conditions. Single session and overlapping activations of all tapping vs. all rest conditions at the three examined thresholds. DSC: Dice Similarity Coefficient; FWE: Family-Wise Error.

## Graphical Abstract - Text

This study reports on the group-level test-retest reliability of an fMRI finger-tapping task. We examined overlap of activations across two sessions using Dice Similarity Coefficients and investigated amplitudes of activations by calculation of ROI-based Intraclass Correlation Coefficients in three sets of ROIs. The task included four different tapping conditions and 31 healthy adults were included in the analyses. We found good to excellent overlap and fair to good amplitude agreement in most contrasts and ROI-sets.