Idiosyncratic relation between human brain activity and behavior

Johan Nakuci[1*], Jiwon Yeon[2,3], Kai Xue[1], Ji-Hyun Kim[4], Sung-Phil Kim[4] and Dobromir Rahnev[1]

[1]School of Psychology, Georgia Institute of Technology, Atlanta, Georgia, 30332, USA.

[2]Department of Psychology, Stanford University, Stanford, California, 94305, USA.

[3]Department of Ophthalmology, Stanford University, School of Medicine, Stanford, California, 94305, USA.

[4]Department of Biomedical Engineering, Ulsan National Institute of Science and Technology, Ulsan, South Korea.

*Corresponding author. Email: jnakuci3@gatech.edu

Keywords: second-level analyses, general linear model, fMRI, perceptual decision making, mental health, individual differences

Abstract

Human behavior is known to be idiosyncratic, yet research in neuroscience typically assumes a universal brain-behavior relationship. Here we test this assumption by estimating the level of idiosyncrasy in individual brain-behavior maps obtained using human neuroimaging. We first show that task-based activation maps are both stable within an individual and similar across people. Critically, although behavior-based activation maps are also stable within an individual, they strongly diverge across people. A computational model that jointly generates brain activity and behavior explains these results and reveals that within-person factors have much larger effect than group factors in determining behavior-based activations. These findings demonstrate that unlike task-based activity that is mostly similar among people, the relation between brain activity and behavioral outcomes is largely idiosyncratic. Thus, contrary to popular assumptions, group-level behavior-based maps reveal relatively little about each individual.

Introduction

Human behavior is idiosyncratic: what elicits a certain behavior in one person is often very different from what elicits that same behavior in another (Eysenck, 1953; Forkosh et al., 2019). For example, one person may shout at strangers but never at home, while another may shout at home but never in public. Yet, research in neuroscience often tacitly assumes that behavior is homogeneous in the population and that the same neural correlates of a given behavior should emerge across individuals (Friston et al., 1999). This assumption is implicit in the common practice – enshrined in popular tools for functional MRI (fMRI) analyses such as SPM, AFNI and FSL – of performing and reporting second-level results as the true neural correlate of a given behavioral outcome.

Here we directly test this assumption by comparing the idiosyncrasy of brain activity maps obtained using two different categories of analyses: (1) task-based analyses such as comparisons of different stimuli, tasks, or internal states (Buckner et al., 1996; Kanwisher et al., 1997; Morrone et al., 2000; Rosenberg et al., 2020; Singer et al., 2004), and (2) behavior-based analyses such as comparisons between left/right choices, fast/slow responses, or high/low confidence (Fleming et al., 2012; Morrone et al., 2000; Yarkoni et al., 2009). Both task- and behavior-based analyses are routinely performed in neuroscience research and the distinction between them is rarely even noticed. In fact, no study to date has hinted that these staple analyses may fundamentally differ in their consistency across individuals.

We collected a unique fMRI dataset (N = 50) that allowed us to jointly estimate average behavior and brain activity for short blocks of trials. Subjects judged whether a briefly presented display featured more red or blue dots and provided a confidence rating (**Figure 1A**). The experiment was organized in 96 blocks of 8 trials each (**Figure 1B**; see Materials and Methods for full details). For each block, we computed average reaction time (RT), average confidence, and per voxel beta values corresponding to the total activation in that voxel over the course of each block (**Figure 1B**).

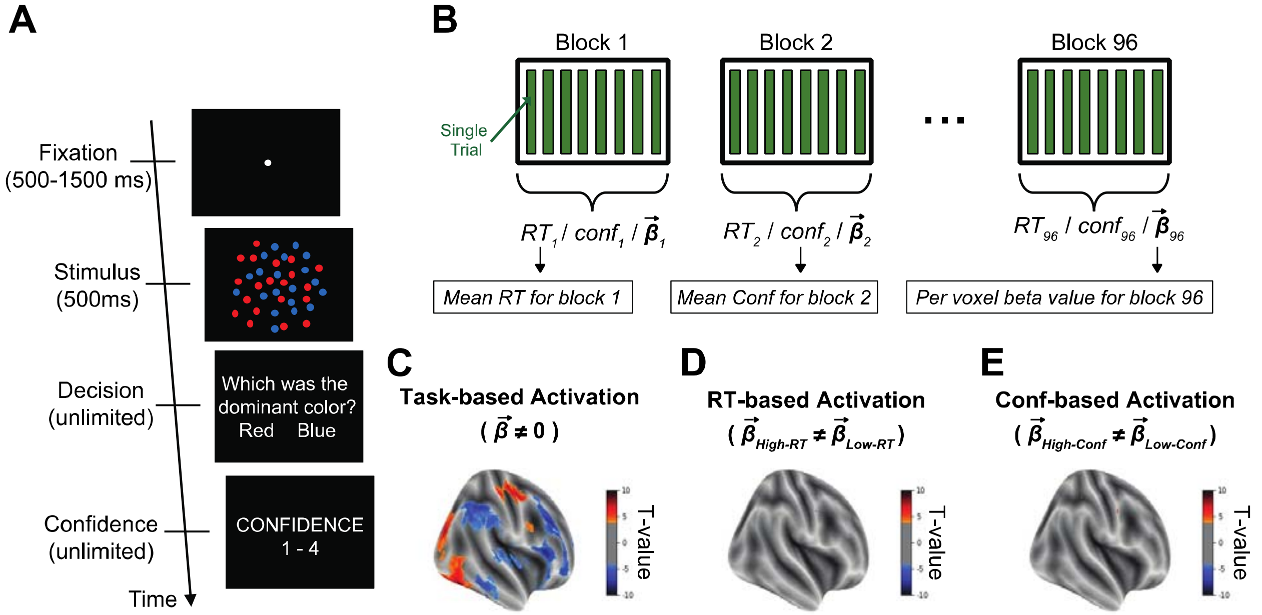

A
Fixation
(500-1500 ms)
Stimulus
(500ms)
Decision
(unlimited)
Which was the dominant color?
Red Blue
Confidence
(unlimited)
CONFIDENCE
1 - 4
Time
B
Block 1
Block 2
Block 96
Single Trial
...
RT_1 / $conf_1$ / $\vec{\beta}_1$
RT_2 / $conf_2$ / $\vec{\beta}_2$
RT_{96} / $conf_{96}$ / $\vec{\beta}_{96}$
Mean RT for block 1
Mean Conf for block 2
Per voxel beta value for block 96
C
Task-based Activation
($\vec{\beta} \neq 0$)
D
RT-based Activation
($\vec{\beta}_{High\text{-}RT} \neq \vec{\beta}_{Low\text{-}RT}$)
E
Conf-based Activation
($\vec{\beta}_{High\text{-}Conf} \neq \vec{\beta}_{Low\text{-}Conf}$)
10
5
0
-5
-10
T-value

Figure 1. Task and results of standard group analyses. (A) Task. Subjects performed a simple perceptual decision-making task that required them to judge the dominant color in a display of colored dots and rate their confidence. (B) Task structure and analysis. The experiment was organized in 96 blocks of 8 trials each (total of 768 trials). For each block, we computed the average RT, average confidence, and per voxel activation (beta value). (C) Standard task-based group analyses. We compared the voxel activations obtained across subjects against zero, and found that our task induced consistent increases and decreases in activation across several brain regions. (D-E) Standard behavior-based group analyses. We also performed classic RT- and confidence-based analyses that compared activations between blocks of high vs. low average RT or high vs. low average confidence. Unlike the task-based analysis in panel C, these analyses revealed no activations after whole-brain correction. A common interpretation of these results is that brain activations for high vs. low RT or confidence do not differ from each other. All maps thresholded at $p < 0.001$ uncorrected for display purposes.

Results

We first performed standard group fMRI analyses by conducting t-tests across all subjects for each voxel. A task-based analysis compared the obtained beta values with zero and revealed several regions of activation and de-activation (**Figure 1C**). On the other hand, two behavior-based analysis compared the beta values for blocks with higher- vs. lower-than-median average RT, as well as higher- vs. lower-than-median average confidence. Both comparisons revealed no activations anywhere in the brain after whole-brain correction (**Figure 1D-E, S1**).

The customary interpretation of these standard group analyses would be that brain activity does not differ for blocks of high vs. low average RT (or high vs. low confidence) in our study. However, examination of individual subject maps demonstrates that such conclusion would be misguided. For example, we inspected the activations for the task-based and the two behavior-based analyses in Subjects 1-3 (**Figure 2**). We found that the task-based maps for all three subjects were similar to each other, with clear activations in visual and somato-motor regions, as well as de-activations in areas of the default mode network (**Figure 2A**). However, very different results emerged for the two behavior-based analyses. Those analyses still revealed areas that consistently tracked RT (**Figure 2B**) and confidence (**Figure 2C**) but the individual maps were highly dissimilar. For example, high-RT blocks were associated with stronger visual cortex activations in Subject 1 but weaker visual cortex activations in Subject 2, whereas the opposite pattern of results was observed in the somato-motor cortex of these two subjects. One may want to dismiss such subject-by-subject variability as being purely due to noise. However, examining the equivalent analyses for odd and even blocks only (see smaller brain maps in **Figure 2A-C**) reveals that each individual map is highly reliable and therefore the stark differences in subjects' behavior-based maps do not appear to be simply due to noise.

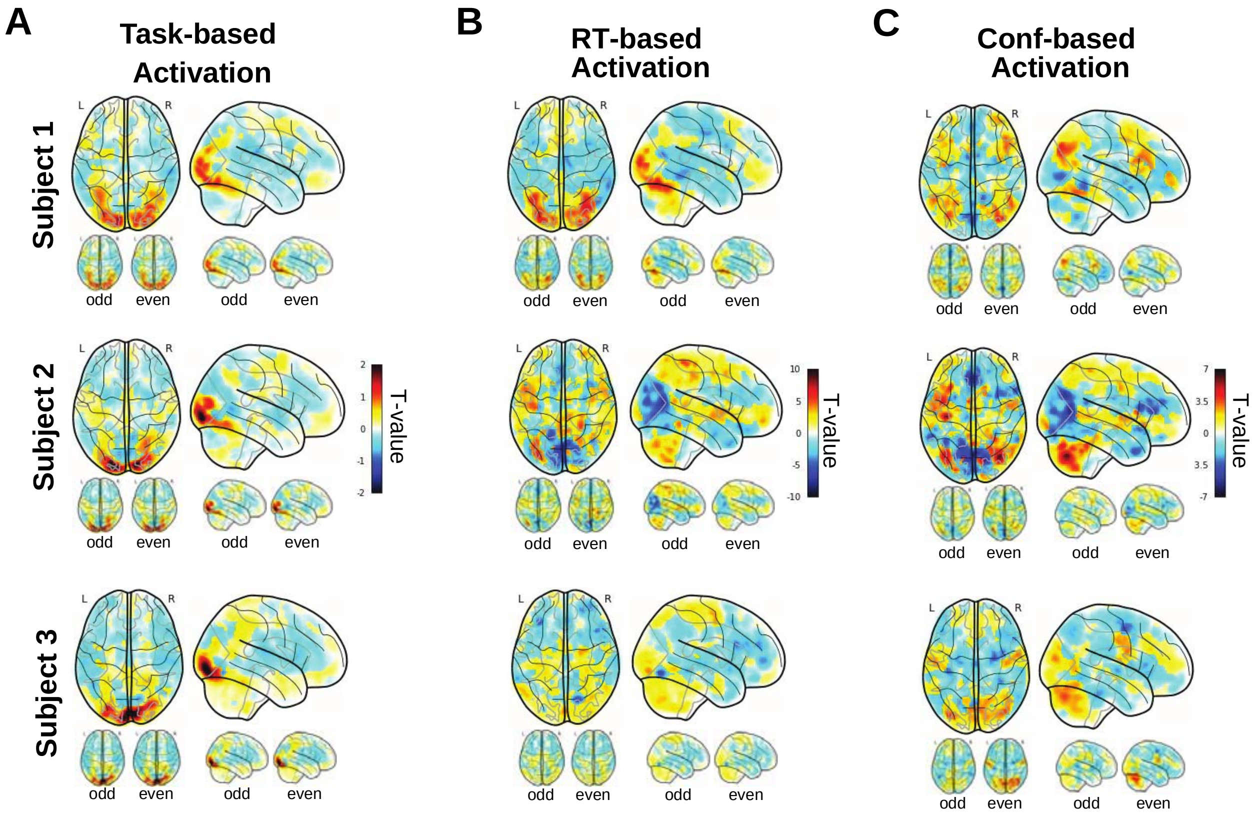


Figure 2. Task- and behavior-based maps for three example subjects. The maps obtained for the first three subjects are displayed for (A) task-based, (B) RT-based, and (C) confidence-based analyses. Task-based activations are computed by averaging the 96 beta values across all blocks, whereas behavior-based activations are the result of t-tests comparing the beta values for blocks with lower- vs. higher-than-median RT or lower- vs. higher-than-median confidence. Small brains underneath represent the same analyses conducted only on odd or even blocks. Similar activations for all three subjects appear for the task-based but not for the two behavior-based analyses, despite the high within-subject consistency of all maps.

To formally test these impressions, we first examined the within-subject reliability of the whole-brain maps produced by the task-, RT-, and confidence-based analyses. We computed this reliability by performing, for the top-10% most activated voxels of each subject, a Pearson correlation between the activations obtained when examining only the odd or only the even blocks. We found that the within-subject reliability for the task-based maps was near perfect (average correlation between odd and even block maps, r = 0.99) and was significantly higher than the reliability for the two behavior-based maps (RT: r = 0.68, t(49) = 9.82, p = 3.62 x 10^{-13}; Confidence: r = 0.49; t(49) = 10.99, p = 7.79 x 10^{-15}; equivalent results were obtained when the top 5, 25, 50, 75, or 100% of voxels were considered; **Figure 3A**). Therefore, to equate the reliability of task- and behavior-based analyses, we computed the task-based maps based on the activation produced by two blocks only (instead of using all 96 blocks). We found that the reliability in these 2-block task-based analyses (average r = 0.39) was well matched to the confidence-based maps (p = 0.12) and was actually significantly lower than the RT-based maps (p = 3.81 x 10^{-6}). Thus, the 2-block task-based analyses provide a fair way to compare across-subject task- and behavior-based maps without the worry that the task-based maps are inherently less noisy.

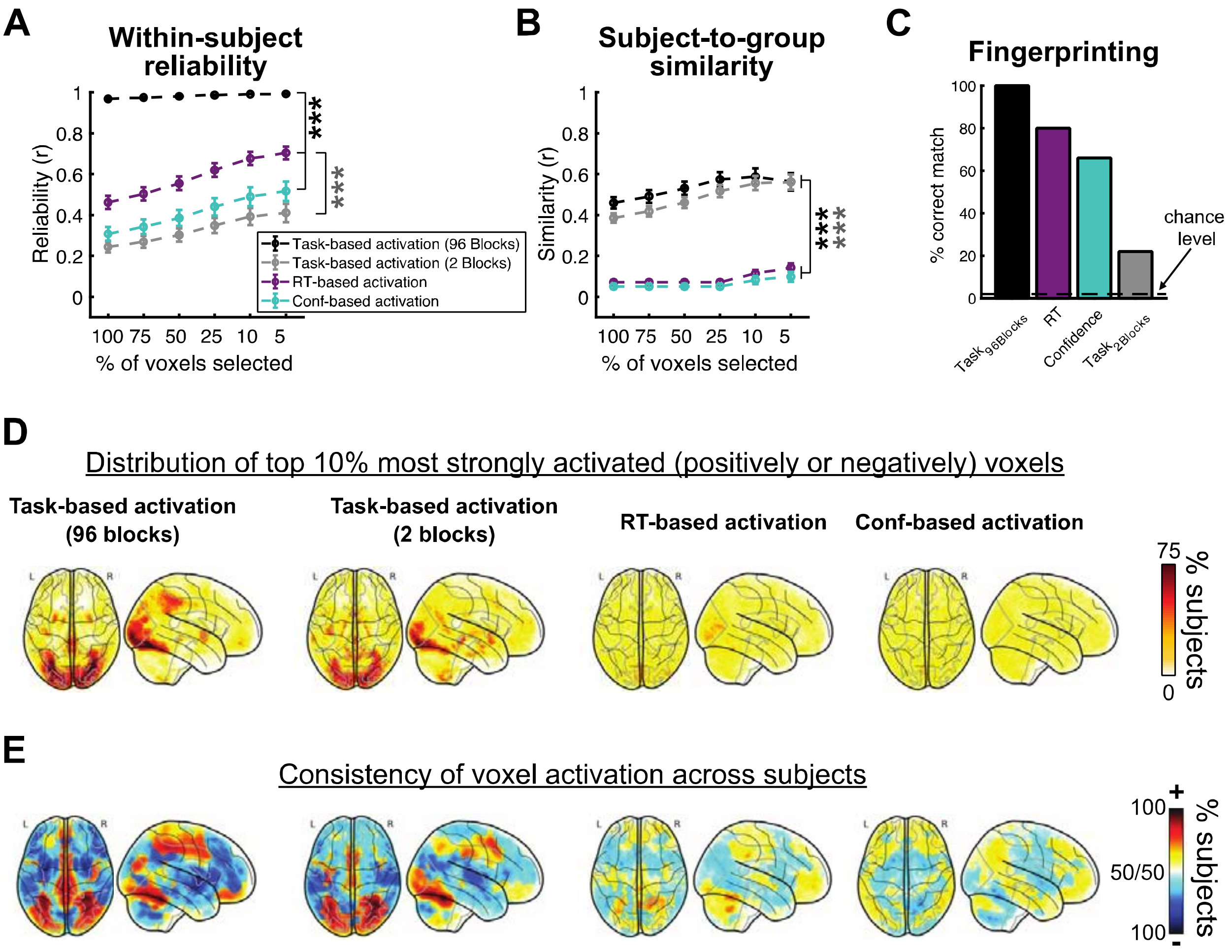


Figure 3. Idiosyncratic brain maps for behavior-based but not task-based analyses. A) Within-subject reliability values as a function of the percent of most activated voxels selected. For all voxel selection levels, the within-subject reliability was highest for the 96-block task-based analysis and lowest for the 2-block task-based analyses. B) Subject-to-group similarity computed as the average similarity between the maps of each person and the group map of the remaining subject. Both the 96- and 2-block task maps exhibited high subject-to-group similarity, whereas both the behavior-based maps exhibited very low (and significantly smaller than the task-based) subject-to-group similarity. C) Fingerprinting results based on the maps produced by the odd and even blocks for each subject. Fingerprinting success clearly depends on both high within-subject reliability and low across-subject similarity. D) Maps of the distribution of the top-10% most activated voxels showing strong areas of consistency for task-based but not for behavior-based analyses. E) Maps of voxel consistency computed as the proportion of subjects showing a positive or negative relationship between voxel activity and behavior. Task-based maps again reveal areas of much higher consistency than behavior-based maps. Error bars show SEM; ***, $p < 0.001$. All p-values are Bonferroni corrected.

Critically, we examined the subject-to-group similarity in the maps produced by task- and behavior-based analyses (**Figure 3B**). For each subject, we correlated their brain map with the group map obtained by averaging the maps of the remaining 49 subjects. Echoing the qualitative impressions from **Figure 2A**, for task-based analyses we found a high degree of subject-to-group similarity that was surprisingly almost identical when examining all 96 blocks (average r = 0.59) or just 2 blocks (average r = 0.57). However, reflecting the results in **Figure 2B-C**, for behavior-based analyses we found very low similarity between the activation maps of each individual and the corresponding group map (RT: mean r = 0.11; Confidence: mean r = 0.08). Both behavior-based similarities were significantly lower than both task-based ones (all pairwise tests: $p < 10^{-14}$). Again, equivalent results were obtained when the top 5, 25, 50, 75, or 100% of voxels were considered.

To gain further intuition for the underlying effects, we conducted three additional analyses. First, we conducted fingerprinting analyses on the odd- and even-blocks maps of each subject (100 maps total) (Finn et al., 2015). We expected that fingerprinting success will be positively related to both high within-subject reliability (making the maps less noisy) and low across-subject similarity (making each individual more distinct from the rest). For each map, we checked which of the remaining 99 maps was closest to it; successful fingerprinting for that map occurred if the closest match came from the other map from the same subject (chance level = 1.01%). We found very high fingerprinting success rate for task-based maps computed from all 96 blocks (100%), presumably driven by their extremely high within-subject reliability (**Figure 3C**). We also found very high fingerprinting success rate for RT- (80%) and confidence-based (66%) maps, presumably driven by their very low subject-to-group similarity. However, the combination of high subject-to-group similarity and the absence of very high within-subject reliability led to substantially lower fingerprinting success for the 2-block task-based maps (22%, significantly lower than all three other maps, all p’s < 10^{-7}). These results confirm the critical role of both within-subject reliability and across-subject similarity for fingerprinting success, and further support the finding of strong idiosyncrasy for behavior- but not task-based analyses.

Second, we further tested the difference between task- and behavior-based analyses by examining the distribution of the locations of the top-10% most strongly activated voxels for each subject (both positive and negative activations were considered). Predictably, we found areas of very high overlap for the 96-block task-based analysis in the visual and parietal cortex with up to 76% of subject showing activation for the same voxel (**Figure 3D**). Critically, despite its low within-subject reliability, the 2-block task-based analysis also showed similar areas of high overlap (maximal overlap: 72%). On the other hand, we found only minimal overlap for RT- or confidence-based analyses (maximal overlap: 32% and 30%, respectively), with maximal overall values that were only slightly higher than the maximum expected value in random data (28%).

Third, we examined the consistency of the sign of voxel activations (whether they were positive or negative) across subjects and again found a stark difference between task- and behavior-based analyses. Specifically, both the 96-block and the 2-block task-based analyses showed areas of very high across-subject consistency (96-block: maximal overlap of 96% and 94% for positive and negative activations; 2-block: maximal overlap of 98% and 92% for positive and negative activations; **Figure 3E**). On the other hand, behavior-based analyses showed substantially lower overlap with most voxels of the brain showing roughly equal proportions of positive and negative activations (RT: maximal overlap of 82% and 80% for positive and negative activations; Confidence: maximal overlap of 76% and 82% for positive and negative activations; expected values in random data: 80% and 80% for positive and negative activations). Overall, both the 96- and 2-block task-based consistency values showed much a much wider range than both of the behavior-based ones (all p's < 10^{-200}). Altogether, each of these three additional analyses further underscores the very high level of idiosyncrasy in behavior-based analyses.

The maps of Subjects 1-3 (**Figure 2**) suggest that the low subject-to-group similarity in behavior-based analyses are likely due to large-scale, rather than fine-grained, differences in the activation maps. To confirm this impression, we repeated the same analyses above with a wide range of smoothing levels (from 5 to 20 mm) and obtained very similar results (**Figure S2-5**). Further, rather than performing the correlations on a voxel-per-voxel level, we did so on the level of 200 large regions of interest obtained from the Schaefer atlas(Schaefer et al., 2018) and still obtained the same results (**Figure S6**). These findings clearly indicate that the low subject-to-group similarity in behavior-based analyses is due to genuine, large-scale differences in the maps rather than issues of misalignment of individual brain maps (Haxby et al., 2011, 2020; Nieto-Castañón and Fedorenko, 2012).

Having established the existence of a stark difference between the level of idiosyncrasy of task- and behavior-based analyses, we sought to precisely quantify these differences by building a simple computational model that jointly generates behavior and brain activity maps. The model produces activation maps for each individual block based on three group-level factors (group task map, group RT map, and group confidence map), three subject-level factors (subject-specific task map, subject-specific RT map, and subject-specific confidence map), and one noise factor (**Figure 4A**). To keep the model simple, both behavior and individual voxel activation for each group- and task-level factor were generated randomly by ignoring known temporal and inter-regional dependencies. The weight of the noise factor was fixed to 1, leaving the model with a total of six free parameters (one for the weight of each group- and subject-level factor). We then fit the model to the observed within-subject reliability and subject-to-group consistency values computed using 100% of the voxels. Despite its simplicity, the model was able to provide excellent fit to the data from **Figure 3A-B** by capturing closely the observed patterns of within-subject reliability (**Figure 4B**) and subject-to-group similarity (**Figure 4C**). By further considering the number of independent voxels simulated as a free parameter, the model could even fit the observed fingerprinting success rates for the different conditions (**Figure 4D**). The only notable deviation between the data and model fits concerned the size of the difference in subject-to-group similarity between the 2-block and 96-block task-based analyses. Nevertheless, the model was able to capture all other patterns of the data remarkably well despite its simplicity.

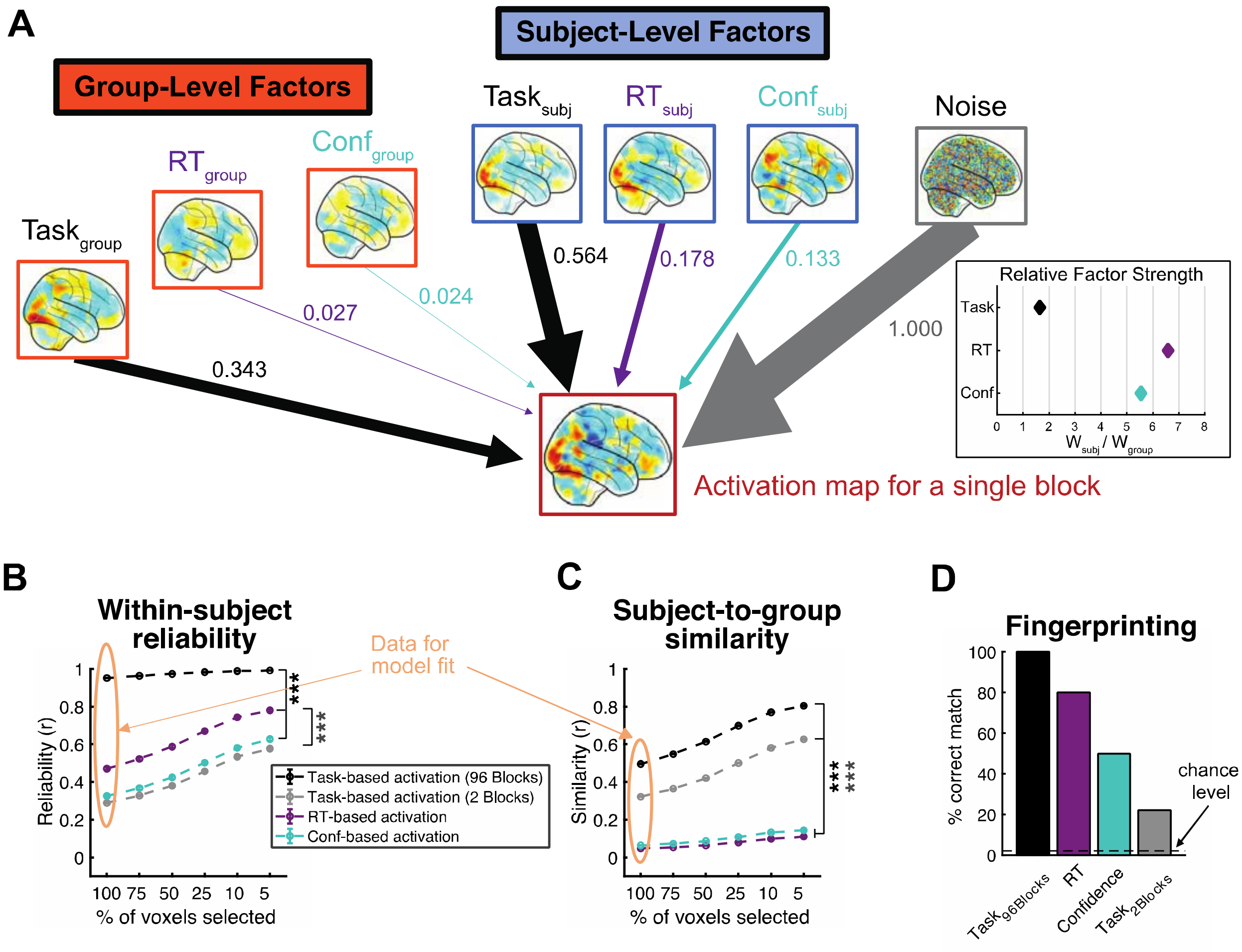


Figure 4. Model structure and predictions. (A) Model structure. The activation map on a single block of trials is modeled as the confluence of three group-level, three subject-level, and one noise factor. The thickness of the arrows and associated numbers correspond to the weights obtained from fitting the model to the data. The inset shows the relative weights of the subject-level and corresponding group-level factors. The brain maps displayed in the figure were produced as follows: the group brain maps show the actual maps obtained in our data, the subject-level brain maps show the maps for subject 1, and the activation map at the bottom shows the activation map for block 1 of subject 1 (B-D) Model predictions for within-subject reliability, subject-to-group reliability, and fingerprinting values. The peach ovals highlight the data points the model was fit to. Despite its simplicity, the model is able to reproduce the empirical data from Fig. 3A-C very well.

Critically, the model allowed us to examine the weights of the group- and subject-level factors, thus providing insight into the relative contribution of each. We found that the weights for subject-level task factor was only a little higher than the group-level task factor (subject-level = 0.564, group-level: 0.343, ratio = 1.64). On the other hand, the weights for RT and confidence subject-level factors were about six times higher than the weights for the corresponding group-level factors (RT: subject-level weight = 0.178, group-level weight = 0.027, ratio = 6.6; Confidence: subject-level weight = 0.133, group-level weight = 0.024, ratio = 5.5). In other words, our model suggests that group- and subject-level factors have relatively similar influence on brain activity, which corresponds well to recent findings about group- and subject-level influences on brain connectivity(Gratton et al., 2018). However, the model reveals that the influence of subject-level behavior-based factors is about six times larger than the influence of group-level behavior-based factors (**Figure 4A inset**). We further repeated the model fitting on data with 5 to 20 mm smoothing, as well as on the data from the 200 Schaefer atlas ROIs and obtained similar results again: the weights ratio between the subject- and group-level task factors was between 1 and 2.5 in all cases, whereas the same ratio for the RT- and confidence-based factors was between 5 and 9.5 (**Figure S7**).

Finally, to establish the replicability of our results, we repeated all analyses on a completely different dataset. We used data from a recently published study (Mazor et al., 2020) where subjects (N = 46) performed a different perceptual decision-making task (Gabor orientation discrimination) and completed a total of up to 200 trials (**Figure S8**). We estimated beta values for mini-blocks of either 2 trials (100 blocks total) or 5 trials (40 blocks total), and then performed the same analyses as above. Even though the smaller quantity of data in that study resulted in slightly lower within-subject reliability and subject-to-group similarity values, the same stark difference between task- and behavior-based analyses emerged again (**Figure S9,10**). Our model was again able to fit the observed data very well and again showed a large difference between task-based factors ratios and behavior-based factor ratios (**Figure S7**). Thus, all of our results were replicated in this independent dataset, demonstrating the generality and replicability of our findings.

Discussion

A major goal of neuroscience research has been to understand the neural correlates of behavior. Behavior is a complex phenomenon that is often idiosyncratic to a person (Eysenck, 1953; Forkosh et al., 2019). Idiosyncratic behavior is ubiquitous in social situations(Durlauf, 2001), economic decisions (Kable and Glimcher, 2007), judgments of beauty (Martinez et al., 2020), confidence ratings (Navajas et al., 2017), response bias (Rahnev, 2021), and even low-level perception (Afraz et al., 2010). However, an implicit assumption in much of neuroscience research is that the neural correlates of behavior are the same across individuals (Friston et al., 1999). Here, we test and reject this assumption. Across two different studies, we find that subject-level behavior-based brain maps are very consistent within an individual, and yet remarkably different across subjects. On the other hand, task-based maps are both consistent within an individual and similar across subjects. A computational model explains these results and suggests that for task-based analyses, the influence of subject-level factors is only slightly stronger than the influence of group-level factors, whereas for behavior-based analyses, the influence of subject-level factors is about six times larger than the influence of group-level factors.

Although these results were unexpected at first, we believe that they are commonsensical in retrospect. Take, for example, the task of running a long-distance race. A “task-based” analysis would certainly reveal many similarities between different runners: when running (as opposed to walking), all people increase their heart rate and exhibit changes in gait and speed (Cappellini et al., 2006). On the other hand, a “behavior-based” within-subject analysis that compares better vs. worse performances within each runner is likely to be fundamentally different. Better performance for one runner may be explained by a faster-than-average heart rate, smaller-than-average step sizes, and higher-than-average levels of hydration. However, for another runner, due to differences in either anatomy or running style, better performance may be predicted by the exact opposite factors. This is not to say that there won't be any consistency in such within-subject behavior-based analyses – there will be – but that the importance of the group-level factors that are the same for everyone is likely to pale in comparison to the importance of the subject-level factors that are idiosyncratic to an individual.

Similarly, it makes sense that task-based brain analyses would reveal strong similarities between subjects: given that large-scale brain anatomy is very similar across people (Hagoort, 2019; Sanes et al., 1995), the same brain areas will likely be involved for everyone in a given task. However, variations in performance from one trial or block to the next would logically depend on different mechanisms for different people (Saleri Lunazzi et al., 2021). For example, someone whose attention is captured too strongly by the visual stimuli may show lower RT with lower visual cortex activity and higher motor cortex activity. On the other hand, another person who fails to attend the visual stimuli consistently may show lower RT when they successfully devote more attention to the screen display resulting in higher visual but lower motor cortex activity.

Our results have strong implications about the common way of reporting the result of fMRI studies. The majority of papers in the field only report second-level, group maps (Bandettini and Ungerleider, 2001; Belliveau et al., 1991; le Bihan et al., 1993; Poldrack, 2011). The implicit assumption is that the group map represents the true "neural correlates" of a given behavior. However, our results demonstrate that such group-level results only have a small relationship to the neural correlates for any actual person (i.e., they explain just 1% of the variance in single-subject maps). This is not to say that second-level maps for behavioral-based analyses are somehow wrong – instead, they represent our best approach for uncovering what is common across all subjects. The issue is that focusing exclusively on the relatively weak commonality across people can distract from the much stronger but idiosyncratic effects within each individual. Our results thus highlight the need for a renewed focus on investigating the brain-behavior relationship at the level of single subjects (Gilmore et al., 2021; Gordon and Nelson, 2021; Naselaris et al., 2021; Song and Rosenberg, 2021). Perhaps ironically, while thousands of subjects are needed for brain-wide association studies (Marek et al., 2022), revealing the brain correlates of behavior requires us to focus on single individuals.

Our findings also suggest novel ways for finding robust biomarkers for various mental disorders (Elliott et al., 2018; Kaufmann et al., 2017; Li et al., 2020; Parkes et al., 2020). Most research in the field has focused on biomarkers unrelated to behavior such as functional connectivity patterns at rest (Drysdale et al., 2017; Woodward and Cascio, 2015). An exciting possibility is that subject-level activations maps for disease-relevant behaviors could serve as much more powerful biomarkers because of their high reliability and clear differences among people. Focusing directly on the relationship between one's behavior and one's brain activations may help to delineate the intricate relationship between the brain and psychopathology (Gratton et al., 2020). For example, our results imply that mental illness might have neural correlates that are unique to an individual. For example, in one individual, fluctuation in positive mood or thoughts might depend on activity in the frontal cortex, but in another individual, they could depend on the parietal cortex. Similar effects have already been suggested in the context of pain perception

(Kohoutová et al., 2022). Therefore, subject-level effects would be crucial to diagnosing and treating these different individuals. Additionally, an analysis that is focused on subject-level variability might be more informative since between-subject analyses ignore the large degree of within-subject variability (Fisher et al., 2018; Lebreton et al., 2019).

One open problem concerns the quantification of the level of idiosyncrasy for different types of analyses. Here we have provisionally classified all analyses as either “task-based” or “behavior-based.” However, it could be that rather than a binary distinction, there is more of a continuum of analysis types. For example, examining the differential activations of two different tasks (Yeon et al., 2020) may show slightly higher idiosyncrasy levels than examining the activations of a single task in isolation (as in the current analyses). Similarly, analyses that compare internal states (e.g., aroused vs. unaroused, excited vs. bored) (Rosenberg et al., 2020) or the effects of brain stimulation (Chen et al., 2013; Rafiei et al., 2021) may show yet greater levels of idiosyncrasy despite our provisional classification of such analyses as “task-based.”

Our model represents one of very few attempts to build process models that jointly generate behavior and brain activations. There are rich literatures of building process model in cognitive science that focuses exclusively on behavior (Ratcliff, 1978; Rescorla and Wagner, 1972; Zhang and Luck, 2008) and in cognitive neuroscience that focuses exclusively on brain activity (Breakspear, 2017; Coombes et al., 2007; Wilson and Cowan, 1972). Yet, understanding the brain-behavior relationship clearly requires the development of process models that jointly generate both types of data, thus explicitly clarifying the links between the two. Future work in the field should increasingly emphasize process models that specify the mechanisms that generate behavior and brain activity.

In conclusion, we find a stark level of idiosyncrasy in behavior-based analyses such that single-subject maps are remarkably reliable, yet very different across subjects. These results have strong implications about the common practice of only reporting second-level analyses and suggest the need to examine individual-subject results in all behavior-based analyses.

Acknowledgments: This work was supported by the National Institutes of Health (grant R01MH119189 to DR) and the Office of Naval Research (grant N00014-20-1-2622 to DR).

Author contributions:

Conceptualization: JN, DR

Methodology: JN, JY, KX, DR

Data Curation: JY, JHK, SPK

Visualization: JN, DR

Funding acquisition: DR

Writing – original draft: JN, DR

Writing – review & editing: JN, JY, KX, JHK, SPK, DR

Competing interests: Authors declare that they have no competing interests.

References

Acerbi, L., and Ma, W.J. (2017). Practical Bayesian Optimization for Model Fitting with Bayesian Adaptive Direct Search. In Advances in Neural Information Processing Systems *30*, 1834–1844. .

Afraz, A., Pashkam, M.V., and Cavanagh, P. (2010). Spatial heterogeneity in the perception of face and form attributes. Current Biology *20*. https://doi.org/10.1016/j.cub.2010.11.017.

Bandettini, P.A., and Ungerleider, L.G. (2001). From neutrons to BOLD: new connections. Nature Neuroscience *4*.

Belliveau, J.W., Kennedy, D.N., McKinstry, R.C., Buchbinder, B.R., Weisskoff, R.M., Cohen, M.S., Vevea, J.M., Brady, T.J., and Rosen, B.R. (1991). Functional mapping of the human visual cortex by magnetic resonance imaging. Science (1979) *254*. https://doi.org/10.1126/science.1948051.

le Bihan, D., Turner, R., Zeffiro, T.A., Cuenod, C.A., Jezzard, P., and Bonnerot, V. (1993). Activation of human primary visual cortex during visual recall: A magnetic resonance imaging study. Proc Natl Acad Sci U S A *90*. https://doi.org/10.1073/pnas.90.24.11802.

Breakspear, M. (2017). Dynamic models of large-scale brain activity. Nature Neuroscience *20*. https://doi.org/10.1038/nn.4497.

Buckner, R.L., Bandettini, P.A., O'Craven, K.M., Savoy, R.L., Petersen, S.E., Raichle, M.E., and Rosen, B.R. (1996). Detection of cortical activation during averaged single trials of a cognitive task using functional magnetic resonance imaging. Proceedings of the National Academy of Sciences *93*, 14878–14883. https://doi.org/10.1073/pnas.93.25.14878.

Cappellini, G., Ivanenko, Y.P., Poppele, R.E., and Lacquaniti, F. (2006). Motor patterns in human walking and running. Journal of Neurophysiology *95*. https://doi.org/10.1152/jn.00081.2006.

Chen, A.C., Oathes, D.J., Chang, C., Bradley, T., Zhou, Z.W., Williamsa, L.M., Glover, G.H., Deisseroth, K., and Etkin, A. (2013). Causal interactions between fronto-parietal central executive and default-mode networks in humans. Proc Natl Acad Sci U S A *110*. https://doi.org/10.1073/pnas.1311772110.

Coombes, S., Venkov, N.A., Shiau, L., Bojak, I., Liley, D.T.J., and Laing, C.R. (2007). Modeling electrocortical activity through improved local approximations of integral neural field equations. Physical Review E - Statistical, Nonlinear, and Soft Matter Physics *76*. https://doi.org/10.1103/PhysRevE.76.051901.

Drysdale, A.T., Grosenick, L., Downar, J., Dunlop, K., Mansouri, F., Meng, Y., Fetcho, R.N., Zebley, B., Oathes, D.J., Etkin, A., et al. (2017). Resting-state connectivity biomarkers define neurophysiological subtypes of depression. Nature Medicine *23*, 28–38. https://doi.org/10.1038/nm.4246.

Durlauf, S.N. (2001). A framework for the study of individual behavior and social interactions. Sociological Methodology *31*. https://doi.org/10.1111/0081-1750.00089.

Elliott, M.L., Romer, A., Knodt, A.R., and Hariri, A.R. (2018). A Connectome-wide Functional Signature of Transdiagnostic Risk for Mental Illness. Biological Psychiatry *84*, 452–459. .

Eysenck, H.J. (1953). The structure of human personality. (New York, NY, US: Methuen).

Finn, E.S., Shen, X., Scheinost, D., Rosenberg, M.D., Huang, J., Chun, M.M., Papademetris, X., and Todd Constable, R. (2015). Functional connectome fingerprinting: Identifying individuals based on patterns of brain connectivity HHS Public Access. Nat Neurosci *18*, 1664–1671. https://doi.org/10.1038/nn.4135.

Fisher, A.J., Medaglia, J.D., and Jeronimus, B.F. (2018). Lack of group-to-individual generalizability is a threat to human subjects research. Proc Natl Acad Sci U S A *115*, 6106–6115. .

Fleming, S.M., Huijgen, J., and Dolan, R.J. (2012). Prefrontal Contributions to Metacognition in Perceptual Decision Making. The Journal of Neuroscience *32*, 6117. https://doi.org/10.1523/JNEUROSCI.6489-11.2012.

Forkosh, O., Karamihalev, S., Roeh, S., Alon, U., Anpilov, S., Touma, C., Nussbaumer, M., Flachskamm, C., Kaplick, P.M., Shemesh, Y., et al. (2019). Identity domains capture individual differences from across the behavioral repertoire. Nature Neuroscience *22*, 2023–2028. https://doi.org/10.1038/s41593-019-0516-y.

Friston, K.J., Holmes, A.P., Price, C.J., Büchel, C., and Worsley, K.J. (1999). Multisubject fMRI Studies and Conjunction Analyses. Neuroimage *10*, 385–396. .

Gilmore, A.W., Nelson, S.M., and McDermott, K.B. (2021). Precision functional mapping of human memory systems. Current Opinion in Behavioral Sciences *40*. https://doi.org/10.1016/j.cobeha.2020.12.013.

Gordon, E.M., and Nelson, S.M. (2021). Three types of individual variation in brain networks revealed by single-subject functional connectivity analyses. Current Opinion in Behavioral Sciences *40*. https://doi.org/10.1016/j.cobeha.2021.02.014.

Gratton, C., Laumann, T.O., Nielsen, A.N., Greene, D.J., Gordon, E.M., Gilmore, A.W., Nelson, S.M., Coalson, R.S., Snyder, A.Z., Schlaggar, B.L., et al. (2018). Functional Brain Networks Are Dominated by Stable Group and Individual Factors, Not Cognitive or Daily Variation. Neuron https://doi.org/10.1016/j.neuron.2018.03.035.

Gratton, C., Kraus, B.T., Greene, D.J., Gordon, E.M., Laumann, T.O., Nelson, S.M., Dosenbach, N.U.F., and Petersen, S.E. (2020). Defining Individual-Specific Functional Neuroanatomy for Precision Psychiatry. Biological Psychiatry *88*, 28–39. .

Hagoort, P. (2019). The neurobiology of language beyond single-word processing. Science (1979) *366*. https://doi.org/10.1126/science.aax0289.

Haxby, J. v., Guntupalli, J.S., Connolly, A.C., Halchenko, Y.O., Conroy, B.R., Gobbini, M.I., Hanke, M., and Ramadge, P.J. (2011). A Common, High-Dimensional Model of the Representational Space in Human Ventral Temporal Cortex. Neuron *72*, 404–416. .

Haxby, J. v, Guntupalli, J.S., Nastase, S.A., and Feilong, M. (2020). Hyperalignment: Modeling shared information encoded in idiosyncratic cortical topographies. Elife *9*, e56601. .

Kable, J.W., and Glimcher, P.W. (2007). The neural correlates of subjective value during intertemporal choice. Nature Neuroscience *10*. https://doi.org/10.1038/nn2007.

Kanwisher, N., McDermott, J., and Chun, M.M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. The Journal of Neuroscience *17*, 4302 LP – 4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997.

Kaufmann, T., Alnæs, D., Doan, N.T., Brandt, C.L., Andreassen, O.A., and Westlye, L.T. (2017). Delayed stabilization and individualization in connectome development are related to psychiatric disorders. Nature Neuroscience *20*, 513–515. .

Kohoutová, L., Atlas, L.Y., Büchel, C., Buhle, J.T., Geuter, S., Jepma, M., Koban, L., Krishnan, A., Lee, D.H., Lee, S., et al. (2022). Individual variability in brain representations of pain. Nature Neuroscience *25*, 749–759. https://doi.org/10.1038/s41593-022-01081-x.

Lebreton, M., Bavard, S., Daunizeau, J., and Palminteri, S. (2019). Assessing inter-individual differences with task-related functional neuroimaging. Nature Human Behaviour *3*, 897–905. .

Li, A., Zalesky, A., Yue, W., Howes, O., Yan, H., Liu, Y., Fan, L., Whitaker, K.J., Xu, K., Rao, G., et al. (2020). A neuroimaging biomarker for striatal dysfunction in schizophrenia. Nature Medicine *26*, 558–565. .

Marek, S., Tervo-Clemmens, B., Calabro, F.J., Montez, D.F., Kay, B.P., Hatoum, A.S., Donohue, M.R., Foran, W., Miller, R.L., Hendrickson, T.J., et al. (2022). Reproducible brain-wide association studies require thousands of individuals. Nature *603*, 654–660. https://doi.org/10.1038/s41586-022-04492-9.

Martinez, J.E., Funk, F., and Todorov, A. (2020). Quantifying idiosyncratic and shared contributions to judgment. Behavior Research Methods *52*. https://doi.org/10.3758/s13428-019-01323-0.

Mazor, M., Friston, K.J., and Fleming, S.M. (2020). Distinct neural contributions to metacognition for detecting, but not discriminating visual stimuli. Elife *9*, e53900. https://doi.org/10.7554/eLife.53900.

Morrone, M.C., Tosetti, M., Montanaro, D., Fiorentini, A., Cioni, G., and Burr, D.C. (2000). A cortical area that responds specifically to optic flow, revealed by fMRI. Nature Neuroscience *3*, 1322–1328. https://doi.org/10.1038/81860.

Naselaris, T., Allen, E., and Kay, K. (2021). Extensive sampling for complete models of individual brains. Current Opinion in Behavioral Sciences *40*. https://doi.org/10.1016/j.cobeha.2020.12.008.

Navajas, J., Hindocha, C., Foda, H., Keramati, M., Latham, P.E., and Bahrami, B. (2017). The idiosyncratic nature of confidence. Nature Human Behaviour *1*. https://doi.org/10.1038/s41562-017-0215-1.

Nieto-Castañón, A., and Fedorenko, E. (2012). Subject-specific functional localizers increase sensitivity and functional resolution of multi-subject analyses. Neuroimage *63*, 1646–1669. .

Parkes, L., Satterthwaite, T.D., and Bassett, D.S. (2020). Towards precise resting-state fMRI biomarkers in psychiatry: synthesizing developments in transdiagnostic research, dimensional models of psychopathology, and normative neurodevelopment. Current Opinion in Neurobiology *65*, 120–128. .

Poldrack, R.A. (2011). Inferring mental states from neuroimaging data: From reverse inference to large-scale decoding. Neuron *72*. https://doi.org/10.1016/j.neuron.2011.11.001.

Rafiei, F., Safrin, M., Wokke, M.E., Lau, H., and Rahnev, D. (2021). Transcranial magnetic stimulation alters multivoxel patterns in the absence of overall activity changes. Human Brain Mapping *42*. https://doi.org/10.1002/hbm.25466.

Rahnev, D. (2021). Response Bias Reflects Individual Differences in Sensory Encoding. Psychological Science *32*. https://doi.org/10.1177/0956797621994214.

Ratcliff, R. (1978). A theory of memory retrieval. Psychological Review *85*. https://doi.org/10.1037/0033-295X.85.2.59.

Rescorla, R.A., and Wagner, A.R. (1972). A Theory of Pavlovian Conditioning: Variations in the Effectiveness of Reinforcement and Nonreinforcement. Clasical Conditioning II: Current Research and Theory.

Rosenberg, M.D., Scheinost, D., Greene, A.S., Avery, E.W., Kwon, Y.H., Finn, E.S., Ramani, R., Qiu, M., Constable, R.T., and Chun, M.M. (2020). Functional connectivity predicts changes in attention observed across minutes, days, and months. Proceedings of the National Academy of Sciences *117*, 3797–3807. https://doi.org/10.1073/pnas.1912226117.

Saleri Lunazzi, C., Reynaud, A.J., and Thura, D. (2021). Dissociating the Impact of Movement Time and Energy Costs on Decision-Making and Action Initiation in Humans. Frontiers in Human Neuroscience *15*. https://doi.org/10.3389/fnhum.2021.715212.

Sanes, J.N., Donoghue, J.P., Thangaraj, V., Edelman, R.R., and Warach, S. (1995). Shared neural substrates controlling hand movements in human motor cortex. Science (1979) *268*. https://doi.org/10.1126/science.7792606.

Schaefer, A., Kong, R., Gordon, E.M., Laumann, T.O., Zuo, X.-N., Holmes, A.J., Eickhoff, S.B., and Yeo, B.T.T. (2018). Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. Cerebral Cortex *28*, 3095–3114. .

Singer, T., Seymour, B., O'Doherty, J., Kaube, H., Dolan, R.J., and Frith, C.D. (2004). Empathy for Pain Involves the Affective but not Sensory Components of Pain. Science (1979) *303*, 1157–1162. .

Song, H., and Rosenberg, M.D. (2021). Predicting attention across time and contexts with functional brain connectivity. Current Opinion in Behavioral Sciences *40*. https://doi.org/10.1016/j.cobeha.2020.12.007.

Wilson, H.R., and Cowan, J.D. (1972). Excitatory and Inhibitory Interactions in Localized Populations of Model Neurons. Biophysical Journal *12*. https://doi.org/10.1016/S0006-3495(72)86068-5.

Woodward, N.D., and Cascio, C.J. (2015). Resting-State Functional Connectivity in Psychiatric Disorders. JAMA Psychiatry *72*, 743–744. https://doi.org/10.1001/jamapsychiatry.2015.0484.

Yarkoni, T., Barch, D.M., Gray, J.R., Conturo, T.E., and Braver, T.S. (2009). BOLD Correlates of Trial-by-Trial Reaction Time Variability in Gray and White Matter: A Multi-Study fMRI Analysis. PLOS ONE *4*, e4257-. .

Yeon, J., Shekhar, M., and Rahnev, D. (2020). Overlapping and unique neural circuits are activated during perceptual decision making and confidence. Scientific Reports *10*. https://doi.org/10.1038/s41598-020-77820-6.

Zhang, W., and Luck, S.J. (2008). Discrete fixed-resolution representations in visual working memory. Nature *453*. https://doi.org/10.1038/nature06860.

Materials and Methods

Subjects

Fifty-two healthy subjects were recruited for this study. Two subjects were excluded because one had metal braces in their teeth and one decided to stop the experiment after the second run. All analyses were thus based on the remaining 50 subjects (25 females; Mean age = 26; Age range = 19-40; Compensated 20,000 KRW or approximately 18 USD). All subjects were screened for any history of neurological disorders or MRI contraindications. The study was approved by Ulsan National Institute of Science and Technology Review Board (UNISTIRB-20-30-C) and all subjects gave written consent.

Task

Subjects had to determine which set of colored dots (red vs. blue) is more frequent in a cloud of dots (Fig. 1A). Each trial began with a white fixation dot presented for a variable amount of time between 500-1500 ms at the center of the screen on a black background. Then, the stimulus was shown for 500 ms, followed by untimed decision and confidence screens. The stimulus consisted of between 140 and 190 red- and blue-colored dots (dot size = 5 pixels) dispersed randomly inside an imaginary circle with a radius of 3° from the center of the screen. Four different dot ratios were used – 80/60, 80/70, 100/80, and 100/90, where the two numbers indicate the number of dots from each color. The experiment was organized in blocks of 8 trials each (Fig. 1B), with each dot ratio presented twice in a random order within a block. The more frequent color was pseudo randomized so that there were equal number of trials where red and blue were the correct answer within a run (consisting of 16 blocks). Subjects used an MRI-compatible button box with their right hand to indicate their decision and confidence responses. For the decision response, the index finger was used to indicate a “red” response and the middle finger for a “blue” response. Confidence was given on a 4-point scale, where 1 is the lowest and 4 is the highest, with the rating of 1 mapped to the index finger and the rating of 4 mapped to the little finger.

Subjects performed 6 runs each consisting of 16 blocks of 8 trials (for a total of 768 trials per subject). Three subjects completed only half of the 6th run and another three subjects completed only the first 5 runs due to time constraints. The remaining 44 subjects completed the full 6 runs. Subjects were given 5 seconds of rest between blocks, and self-paced breaks between runs.

MRI recording

The MRI data was collected on a 64-channel head coil 3T MRI system (Magnetom Prisma; Siemens). Whole-brain functional data were acquired using a T2*-weighted multi-band accelerated imaging (FoV = 200 mm; TR = 2000 ms; TE = 35 ms; multiband acceleration factor = 3; in-plane acceleration factor = 2; 72 interleaved slices; flip angle = 90°; voxel size = 2.0 x 2.0 x 2.0 mm^3). High-resolution anatomical MP-RAGE data were acquired using T1-weighted imaging (FoV = 256 mm; TR = 2300 ms; TE = 2.28 ms; 192 slices; flip angle = 8°; voxel size = 1.0 x 1.0 x 1.0 mm^3).

MRI preprocessing and general linear model fitting

MRI data were preprocessed with SPM12 (Wellcome Department of Imaging Neuroscience, London, UK). We first converted the images from DICOM to NIFTI and removed the first three volumes to allow for scanner equilibration. We then preprocessed with the following steps: de-spiking, slice-timing correction, realignment, segmentation, coregistration, normalization, and spatial smoothing with 10 mm full width half maximum (FWHM) Gaussian kernel. In control analyses, we used 5 and 20 mm FWHM smoothing to investigate whether the results are due to fine-grained differences in the activations maps between subjects, given that local differences would be substantially reduced by larger smoothing kernels. Despiking was done using the 3dDespike function in AFNI. The preprocessing of the T1-weighted structural images involved skull-removal, normalization into MNI anatomical standard space, and segmentation into gray matter, white matter, and cerebral spinal fluid, soft tissues, and air and background.

We fit a general linear model (GLM) that allowed us to estimate the beta values for each voxel in the brain for each block of the experiment. The model consisted of regressors for each individual block, inter-block rest periods, as well as linear and squared regressors for six head movement (three translation and three rotation), five tissue-related (gray matter, white matter, cerebrospinal fluid, soft tissues, and air and background), and a constant term per run.

Standard group-level analyses

We first performed a standard group analysis by conducting t-tests across all subjects for each voxel. A task-based analysis compared the obtained beta values with zero to identify regions of activation and de-activation. Two behavior-based analyses compared the beta values for blocks with higher- vs. lower-than-median average reaction times (RT) and higher- vs. lower-than-median average confidence. Significance was assessed using $p < 0.05$ after Bonferroni correction for multiple comparisons. For display purposes, Fig. 1 and Fig. S1 used the more liberal threshold of $p < 0.001$ uncorrected.

Within-subject reliability analyses

We examined the within-subject reliability of the whole-brain maps produced by the task- and behavior-based analyses. To do so, we first re-did the task- and behavior-based analyses by only using the odd blocks, as well as by only using the even blocks. We then compared the similarity between the maps obtained for odd and even blocks using Pearson correlation. We performed the analysis six times based on the top 5, 10, 25, 50, 75, or 100% of most strongly activated voxels in the following way. We first identified the X% most strongly activated voxels (i.e., the voxels with highest absolute activation values) when only examining the data from the odd blocks. The activation values used were the average beta value for task-based analyses, and the t-value (obtained by using a t-test to compare the beta values for blocks with above- vs. below-median RT or confidence) for the behavior-based analyses. This selection procedure ensured that both positively and negatively activated voxels were selected and that an equal number of voxels were selected each time. The activations in the selected top X% of voxels from the odd blocks were then correlated with the activations in the same voxels in the even blocks, thus obtaining an “odd-to-even” correlation value. Then, using an equivalent procedure, we identified the top X% of most activated voxels in the even blocks, and correlated their activations with the activations in the corresponding voxels in the odd blocks, thus obtaining an "even-to-odd" correlation value. Finally, we computed the overall within-subject reliability as the average of the odd-to-even and even-to-odd correlation values.

In addition to the three analyses above that were performed on all collected data, we performed a task-based analysis based on data from just two blocks. We selected blocks 1 and 49 for these analyses because they were the first blocks of the first and second half of the experiment, respectively. The within-subject reliability for this 2-block task-based analysis was computed in the same way as the 96-block task-based analysis above by treating block 1 as the "odd blocks" and block 49 as the "even blocks." Examining data from just two blocks produced lower within-subject reliability values and thus allowed for a fair comparison between task-based and behavior-based analyses.

Subject-to-group similarity analyses

Critically, we examined the subject-to-group similarity in the maps produced by task- and behavior-based analyses. For each subject, we correlated their individual task-, RT-, and confidence-based activation maps with the corresponding group map obtained by averaging the maps of the remaining 49 subjects. We conducted the task-based analyses both on the average of all 96 blocks (96-block analysis) and the average of blocks 1 and 49 (2-block analysis). Similar to the within-subject reliability analyses, we conducted these analyses separately for the top 5, 10, 25, 50, 75, or 100% of most activated voxels. These voxels were selected in the same way as for the within-subject reliability analyses using all of the data in a given subject (except for the 2-block task-based analysis where both blocks 1 and 49 were used); the activations in the voxels identified for a given subject were then correlated with the average activations in the same voxels for the remaining 49 subjects.

Fingerprinting analyses

As another test of the strength of within-subject reliability and across-subject similarity (Finn et al., 2015), we conducted fingerprinting on the brain maps produced by the task- and behavior-based analyses. Specifically, we considered the odd- and even-block maps of each of our 50 subjects (100 maps total). We compared the magnitude of the similarity of each map to each of the remaining 99 maps. For a given map, if the most similar other map was the second map from the same subject (chance level 1/99 = 1.01%), then we counted that as successful fingerprinting for that specific map. We performed this fingerprinting analysis on the 96-block task-based activation maps, 2-block task-based activation maps, RT-based activation maps, and confidence-based activation maps. Finally, we tested for statistical differences in fingerprinting success rate between task- and behavior-based activation maps using a Z-test for proportions.

Distribution of top-10% most strongly activated voxels

As another test of the across-subject similarity of the task- and behavior-based results, we sought to identify the consistency of the location of the most strongly activated brain regions across subjects. For each subject, we selected the top-10% most strongly activated voxels by considering the absolute value of either the average beta value (for task-based analyses) or t-value (for behavior-based analyses). Note that this procedure selected positive and negative activations. We then estimated, for each voxel, the percent of subjects for which the voxel was selected as one of the top-10% most strongly activated voxels. The analysis was performed separately for the 96-block task-based activation maps, 2-block task-based activations maps, RT-based activation maps, and confidence-based activation maps.

Low across-subject similarity in these analyses would result in most voxels being selected about 10% of the time. However, due to chance, some voxels are bound to be selected more than 10% of the time. Therefore, to enable the appropriate interpretation of the obtained results, we computed the expected level of maximal overlap in the maps of 50 subjects whose maps have no relationship to each other. Specifically, for each of the 50 subjects, we generated a random set voxel activation values. We then selected the top-10% of the highest absolute values from each subject and calculated the overlap across subjects. The expected value from random data was computed as the average maximal overlap after 1000 iterations. This analysis revealed that completely random data would produce a maximal overlap of 28% given the number of voxels and number of subjects that we had, which was only a little less than the empirically observed values for behavior-based analyses (32% for RT-based analyses and 30% for confidence-based analyses).

Consistency in activation analysis

As a final test of the across-subject similarity of the task- and behavior-based results, we computed the consistency in the sign of activation. Our main analyses relied on taking correlations, but it is possible that just considering the sign of activation (rather than the strength of activation) would produce different results. To investigate this possibility, we examined the consistency of the sign of voxel activations (positive or negative) across subjects. To do so, we first set all voxels values that were equal to zero to not-a-number value (NaN). This applied to regions that are outside the brain. We then binarized the voxel activation values $activation_i$ such that:

$$binary_i = \begin{cases} 1, & activation_i \geq 0 \\ 0, & activation_i < 0 \end{cases}$$

The consistency of the sign of a voxel's activation across subjects (C_i) was then calculated as percentage of subjects for which a voxel i was positively or negatively activated using the formula:

$$C_i = 100 \times \frac{1}{50} \sum_{j=1}^{50} binary_i$$

As defined, C_i goes from 0 (all subjects having negative activation for that voxel) to 100 (all subjects having positive activations for that voxel), with a value of 50 indicating that half of the subjects had positive and half had negative activation. However, when reporting the values of C_i, we flipped values under 50 using the formula $C_{i,flipped} = 100 - C_i$, so that these values represent the percent of subjects with negative activations. As before, the analysis was performed separately for the 96-block task-based activation maps, 2-block task-based activations maps, RT-based activation maps, and confidence-based activation maps. The activation values were the average beta value (for task-based analyses) or t-value (for behavior-based analyses).

Low across-subject similarity in these analyses would result in most voxels having consistency, C_i, values close to 50 (corresponding to the voxel activation having positive sign in half the subjects and negative sign in the other half). However, due to chance, the consistency values are bound to sometimes be higher. Therefore, to enable the appropriate interpretation of the obtained results, we computed the expected consistency values in the maps of 50 subjects whose maps have no relationship to each other. Specifically, we generated a random set of voxel activation values for each of 50 sample subjects. Maximal consistency from the random data was calculated in the same manner as the empirical values and the procedure was repeated 1000 times. This analysis revealed that completely random data would produce a maximal consistency of 80% (for both positive and negative activations) given the number of voxels and number of subjects that we had, which was close to the empirically observed values for behavior-based analyses.

ROI-based within-subject reliability and subject-to-group similarity analyses

All of the above analyses were performed at the level of individual voxels. However, to ensure that the results obtained were not due to the fine-grained misalignment of individual maps, we performed the within-subject reliability and subject-to-group similarity analyses on the level of 200 large regions of interest (ROIs) obtained from the Schaefer atlas (Schaefer et al., 2018). For these analyses, we averaged all the beta values within an ROI and repeated both the within-subject reliability and subject-to-group similarity analyses in the same way as for the voxel-based analyses above. These ROI-based analyses produced very similar results to the main voxel-based analyses (see Fig. S6), indicating that the low subject-to-group similarity in behavior-based analyses is due to genuine, large-scale differences in the maps rather than issues of misalignment of individual brain maps.

Model specification

Our analyses revealed that task-based activation maps are largely consistent across subjects, whereas behavior-based maps are largely idiosyncratic. We sought to precisely quantify the contributions of group- and subject-level factors in both the task- and behavior-based activations maps by building and fitting a simple computational model.

The model jointly generates behavior and brain activity maps using minimal assumptions. Specifically, RT and confidence for a block of trials were generated randomly from a standard normal distribution (with a mean of 0 and standard deviation of 1). Critically, the model assumes that the activation map for each block is a function of seven different factors. The first three are group-level factors (i.e., factors common to all subjects) for the task itself, the influence of the block-specific RT, and the influence of the block-specific confidence. The next three factors are subject-level factors (i.e., factors specific to each subject) for the task itself, the influence of the block-specific RT, and the influence of the block-specific confidence. Finally, the 7th factor is simply Gaussian noise. Critically, each factor is weighed by a corresponding factor weight that determines the strength of influence of that factor to the final voxel activation values, such that the activation strength (β) for a given voxel on a given block is:

$$\begin{aligned}\beta = w_{task_{group}} * f_{task_{group}} + w_{rt_{group}} * f_{rt_{group}} * RT + w_{conf_{group}} * f_{conf_{group}} * conf \\ + \ w_{task_{subj}} * f_{task_{subj}} + w_{rt_{subj}} * f_{rt_{subj}} * RT + w_{conf_{subj}} * f_{conf_{subj}} \\ * conf + w_{noise} * f_{noise}\end{aligned}$$

where RT and $conf$ are the reaction time and confidence on that block, the w's are the weights associated with each factor, and the f's are the factors that influence the voxel activity for a given block. Without loss of generality, the weight of the noise factor (w_{noise}) was fixed to 1. The value of each factor f was randomly sampled from a standard normal distribution such that group-level factors were randomly sampled for each voxel, subject-level factors were randomly sampled for each voxel and subject, and the noise factor was randomly sampled for each voxel, subject, and block.

We note that, in the model, both the behavioral measures (RT and $conf$) and the factors (f) that control the individual voxel activations were sampled from a standard normal distribution, which makes the simulated values of RT, confidence, and voxel activations (β's) not match their observed values. Additional parameters could be easily used to match the actual distributions of the empirical RT, confidence, and β values, but this added level of complexity would not affect the model's ability to explain the quantities of interest here, which are the within-subject reliability values, subject-to-group similarity values, and fingerprinting results. This is because all three of these analyses are based on correlations, and correlations are insensitive to additive and multiplicative changes of the underlying variables. Therefore, we chose not to fit the actual observed values of RT, confidence, and voxel β values so as to keep the model as simple as possible.

Model fitting

We first fit the model to the empirically observed within-subject reliability and subject-to-group similarity values. The model had six free parameters corresponding to the weights, w, of the group- and subject-level factors that determined the simulated β value for each voxel in each block. For a given set of weights, we simulated a complete experimental dataset by generating simulated data for 50 subjects with 96 blocks per subjects (each block had corresponding RT, confidence, and per voxel beta value). Based on these data, we then computed the within-subject reliability and subject-to-group similarity values in the same way as for the empirical data. When simulating the model, we observed that the exact number of voxels used made no systematic difference to the observed values of the obtained within-subject reliability and subject-to-group similarity values. Therefore, we used 10,000 voxels, which allowed for stable values to be obtained on different iterations. The fitting minimized the mean squared error (MSE) between the simulated and empirically observed within-subject reliability and subject-to-group similarity values calculated using the top-100% most activated voxels (that is, using all voxels). Once the fitting was completed, we also generated the predictions of the best-fitting model for the within-subject reliability and subject-to-group similarity values calculated using the top 5, 10, 25, 50, and 75% most activated voxels. The fitting itself was carried out using the

Bayesian Adaptive Direct Search (BADS) toolbox(Acerbi and Ma, 2017). We fit the model 10 times are reported the best fitting model among the 10 iterations.

In addition to fitting the model to the within-subject reliability and subject-to-group similarity values, we further fit it to the observed fingerprinting success rate. Predictably, we found that the number of simulated voxels had a large effect on the fingerprinting success rate of the simulated data (simulating more voxels leads to more robust fingerprinting due to the availability of more overall data). Therefore, in fitting the model to the fingerprinting success rate, we used the weights (w) obtained from the initial fit above (when fitting the model to the within-subject reliability and subject-to-group similarity values) and then systematically varied the number of simulated voxels in the model from 10 to 100. For each simulation, we computed the fingerprinting success rates as in the analyses of the empirical data, and then chose the number of voxels that minimized the MSE between the simulated and the empirical results. We conducted the simulations 10 times for each number of simulated voxels and we averaged the obtained MSE values across repetitions. We found that simulating 62 voxels led to the best model fit and therefore report these fits in the main paper. It is notable that this number is substantially smaller than the number of voxels in the brain, but it is important to appreciate that voxels in the brain exhibit a very large degree of covariance such that the actual dimensionality of the voxel activations could well be of a similar order of magnitude as the number that our model fitting arrived at.

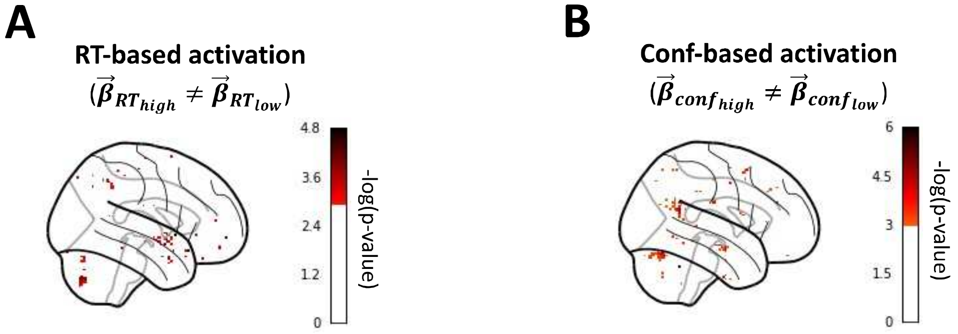
Data and Code Availability

All preprocessed data and code are available at https://osf.io/xe8r5/ and https://osf.io/8dbq2/.

Supplementary Materials for

Idiosyncratic Relation Between Human Brain Activity and Behavior

Johan Nakuci*, Jiwon Yeon, Kai Xue, Ji-Hyun Kim, Sung-Phil Kim and Dobromir Rahnev


A
RT-based activation
$(\vec{\beta}_{RT_{high}} \neq \vec{\beta}_{RT_{low}})$
4.8
3.6
2.4
1.2
0
-log(p-value)
B
Conf-based activation
$(\vec{\beta}_{conf_{high}} \neq \vec{\beta}_{conf_{low}})$
6
4.5
3
1.5
0
-log(p-value)

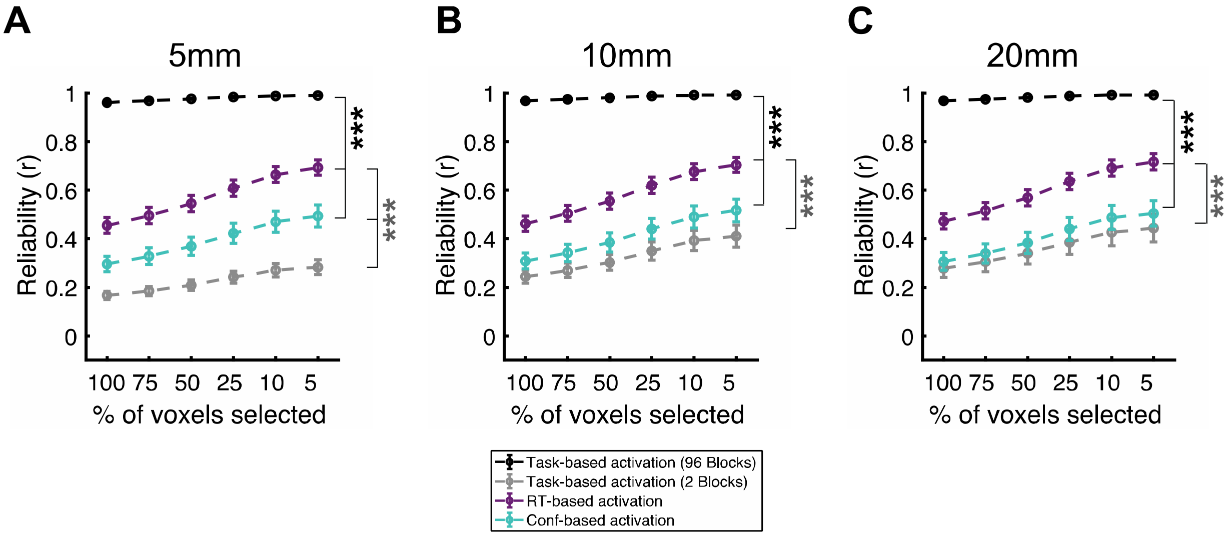


Figure S2. Effect of smoothing on within-subject reliability of the whole-brain maps produced by the task-, RT-, and confidence-based analyses. The fMRI data were spatially smoothed with A) 5 mm, B) 10 mm, or C) 20 mm full width half maximum (FWHM) Gaussian kernel. Panel B is the same as in the manuscript and is shown for comparative purposes. As can be observed, very similar results are obtained for different levels of smoothing, indicating that the results obtained are likely due to large-scale rather than small-grained differences in the maps. Error bars show SEM; ***, p < 0.001. All p-values are Bonferroni corrected.

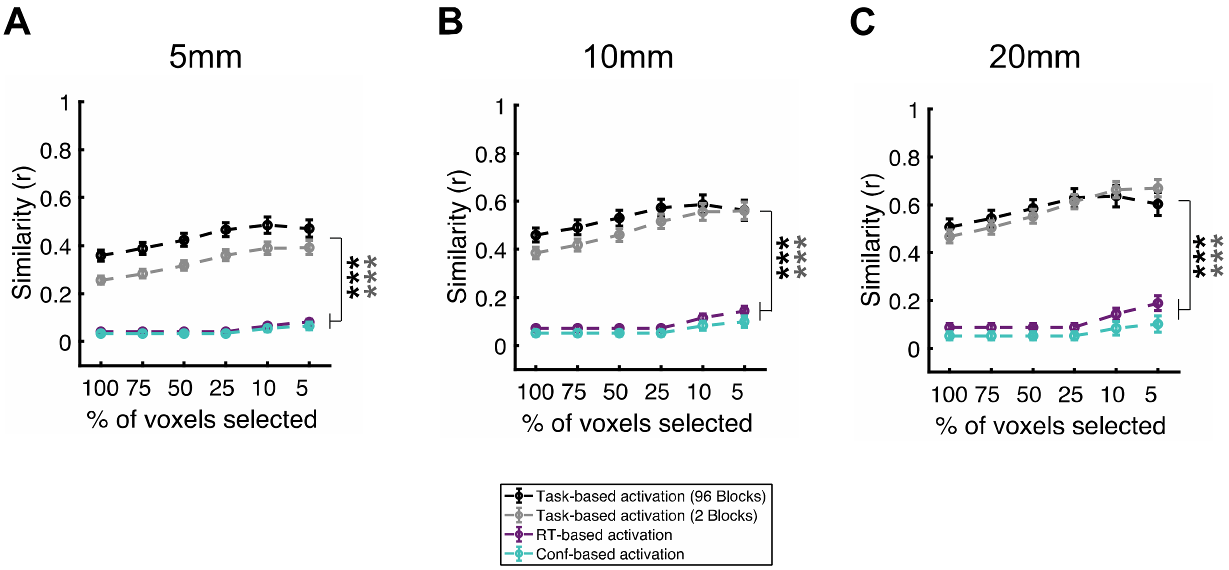


Figure S3. Effect of smoothing on the subject-to-group similarity values. The fMRI data were spatially smoothed with A) 5 mm, B) 10 mm, or C) 20 mm full width half maximum (FWHM) Gaussian kernel. Panel B is the same as in the manuscript and is shown for comparative purposes. As can be observed, very similar results are obtained for different levels of smoothing, indicating that the results obtained are likely due to large-scale rather than small-grained differences in the maps. Error bars show SEM; ***, $p < 0.001$. All p-values are Bonferroni corrected.

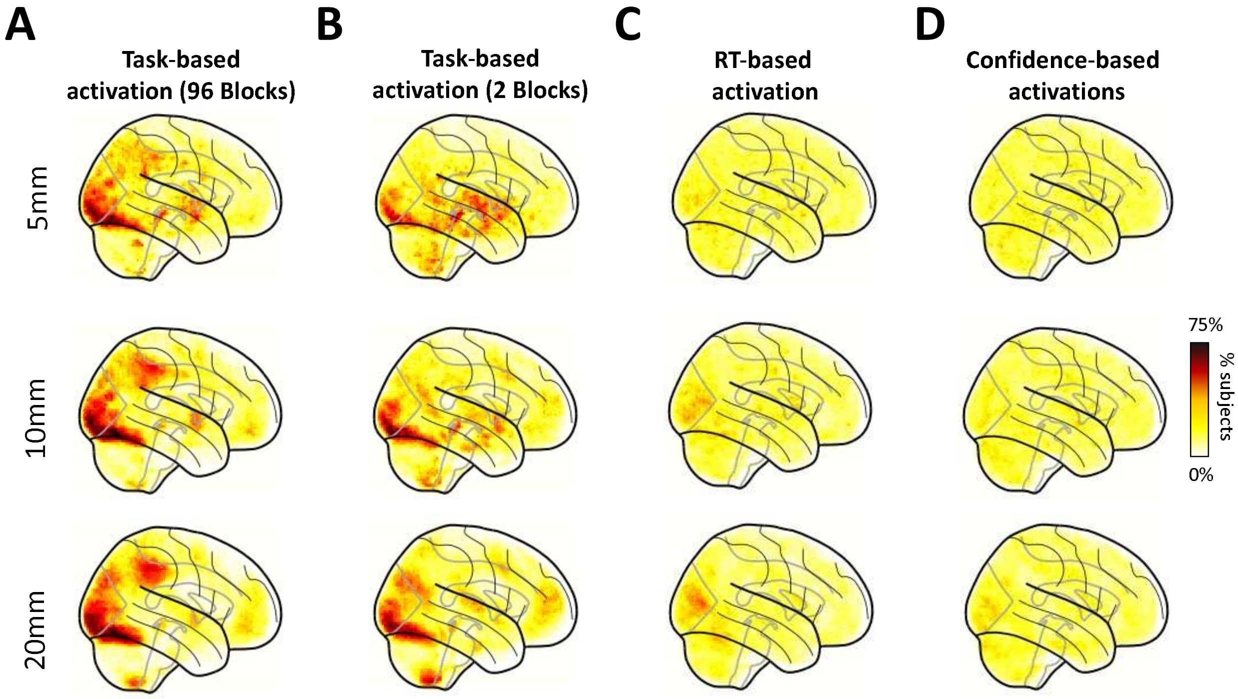

A
Task-based activation (96 Blocks)
B
Task-based activation (2 Blocks)
C
RT-based activation
D
Confidence-based activations
5mm
10mm
20mm
75%
% subjects
0%

Figure S4. Maps of the distribution of the top-10% most activated voxels. A) Task-based activation calculated from all 96 blocks. B) Task-based activation calculated from 2 blocks only. C) RT-based activation. D) Confidence-based activation. Task-based activations computed for both 96 and 2 blocks exhibited strong areas of consistency compared to the behavior-based activations. Analysis was conducted on fMRI data smoothed with 5, 10, and 20 mm FWHM kernels. The 10 mm results are the same as in the main manuscript and are shown here for comparative purposes. Again, similar results are obtained for different levels of smoothing.

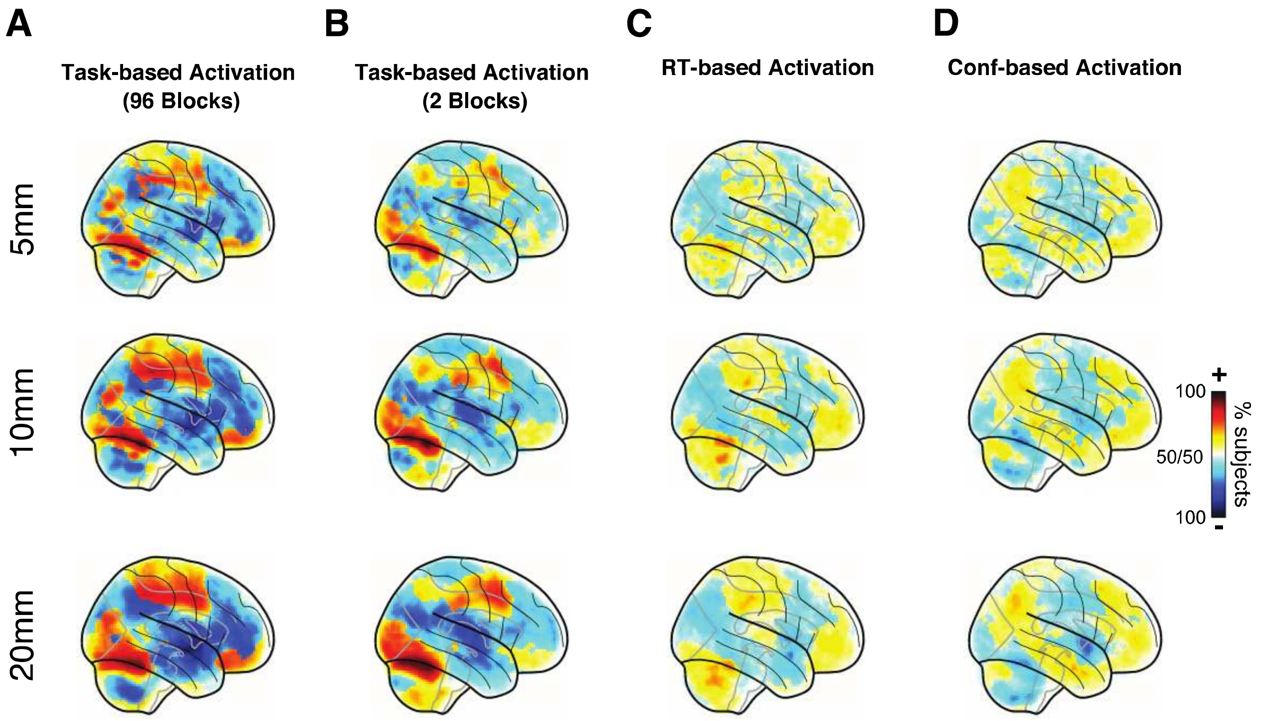


Figure S5. Maps of voxel activation consistency across subjects. A) Task-based activation calculated from all 96 blocks. B) Task-based activation calculated from 2 blocks only. C) RT-based activation. D) Confidence-based activation. Task-based activations computed for both 96 and 2 blocks exhibited strong areas of consistency compared to the behavior-based activations. Analysis was conducted on fMRI data smoothed with 5, 10, and 20 mm FWHM kernels. The 10 mm results are the same as in the main manuscript and are shown here for comparative purposes. Again, similar results are obtained for different levels of smoothing.

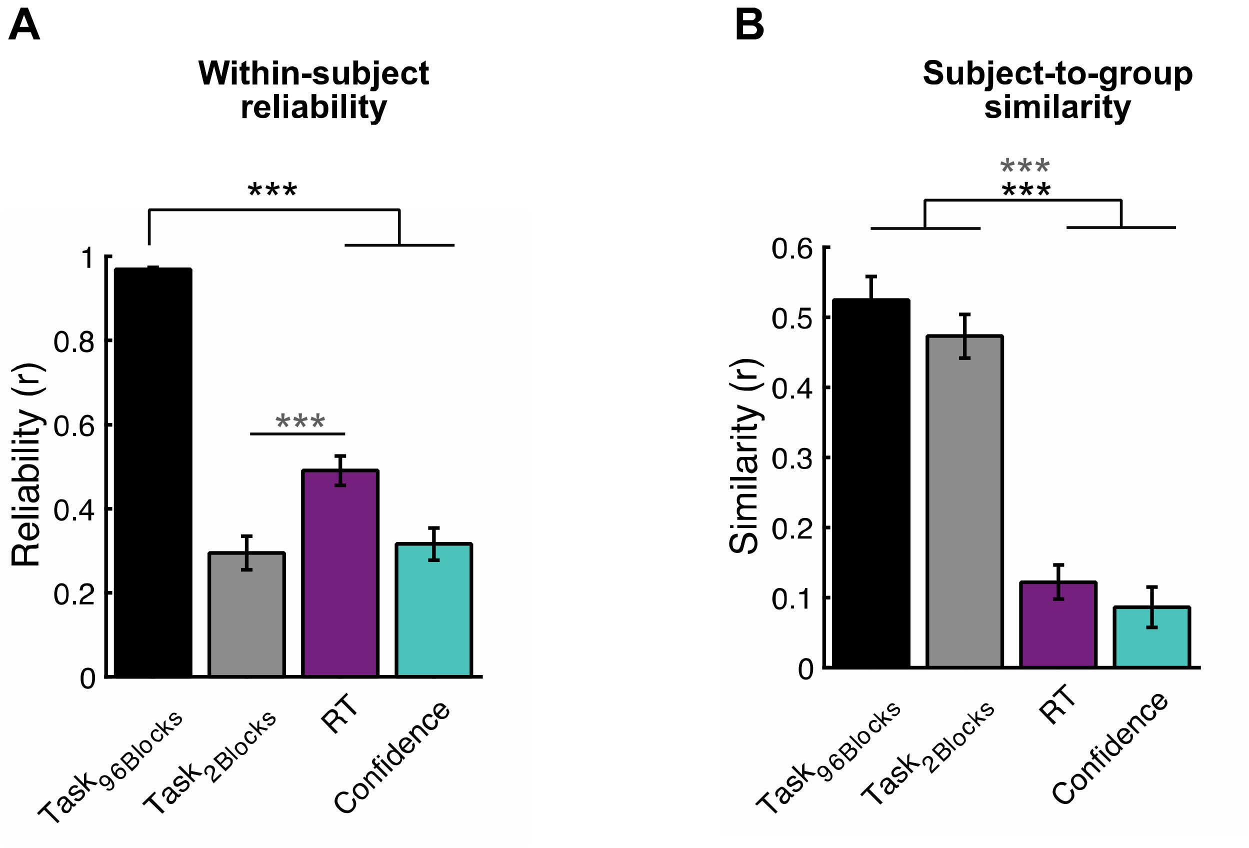


Figure S6. Idiosyncratic relation at the ROI level for behavior-based but not task-based analyses. A) Within-subject reliability values obtained by analyzing activations within the 200 regions-of-interest (ROIs) from the Schaefer atlas. B) Subject-to-group similarity computed based on the activations of the 200 Schaefer atlas ROIs. Analysis is based on fMRI data smoothed with 10 mm FWHM Gaussian kernel. The results obtained at the level of large ROIs in this analysis mimics closely the results obtained using analyses conducted on individual brain voxels (**Fig. 2A,B**). Error bars show SEM; ***, p < 0.001. All p-values are Bonferroni corrected.

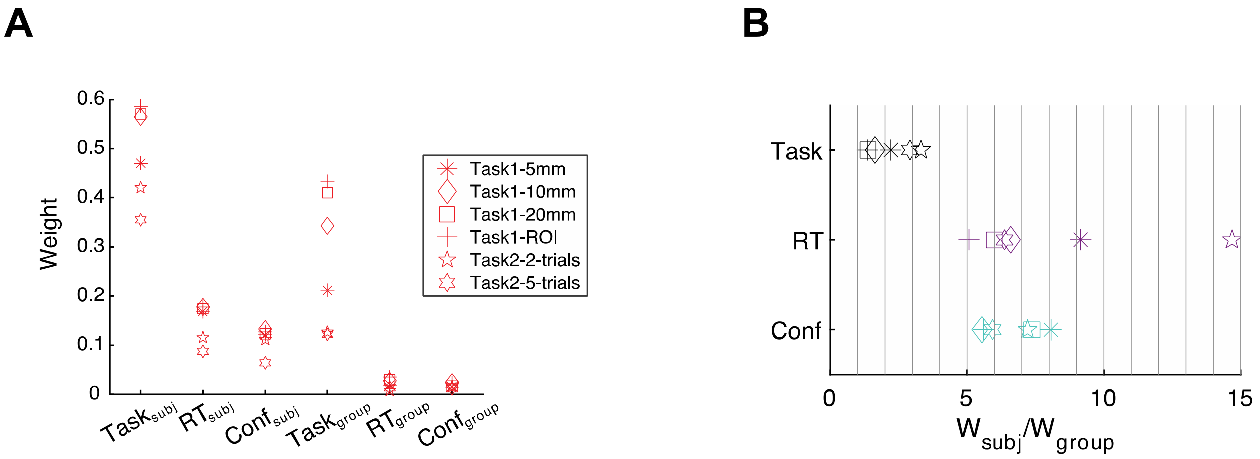


Figure S7. Model weights and ratios. A) Model weights. Subject- and group-level weights obtained from fitting the model separately to each analysis and task. In line with the results in **Figure 4**, the subject-level weights were consistently higher than the group-level weights across all analyses and tasks. Task1 and Task2 refer to the tasks from the main experiment and the data from Mazor et al. (2020) (see Fig. S8), respectively. B) Weight ratios. Relative weights of the subject-level and corresponding group-level factors from each analysis and task. Across most analyses, the weights ratio between the subject- and group-level task factors was between 1 and 3.5, whereas the same ratio for the RT- and confidence-based factors was between 5 and 9.5. The exception was Task2 with blocks composed of 2 trials ("Taks2-2-trials"), with an RT-factor weight ratio of 15.4, which could be a result of higher noise level associated with using only two trials to estimate the beta values.

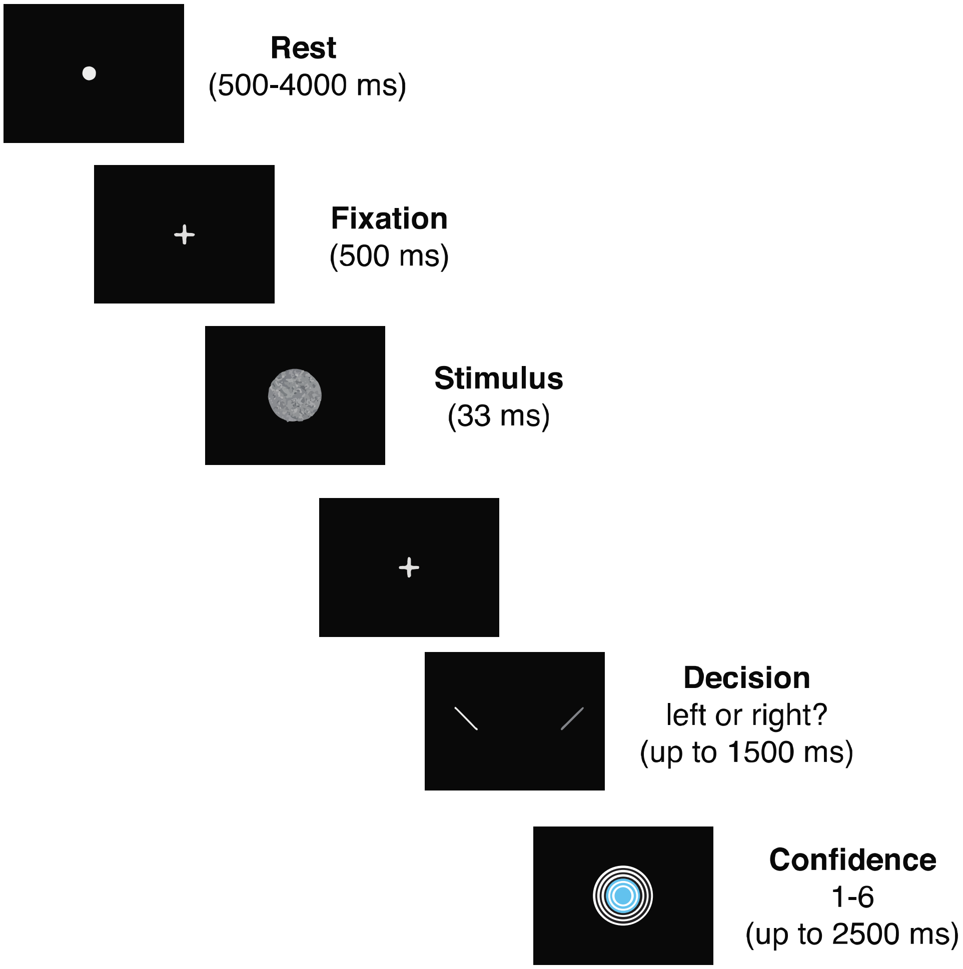


Figure S8. Gabor orientation discrimination task in Mazor et al. (2020). Subject (N=46) performed an in-scanner discrimination task where they had to indicate the orientation of a visual grating (clockwise or counterclockwise). After a temporally jittered rest period that lasted between 500 and 4000 ms, each trial began with a fixation cross (500 ms) followed by the stimulus (33 ms). The subjects had up to 1500 ms to make decision on the orientation of the grating. Immediately after making their decision, subjects rated their confidence on a 6-point scale by increasing or decreasing the size of the color segment of the circle. In total, subjects performed 5 runs with 40 trials per run for a total of 200 trials. They also performed a second task, which was not analyzed here. Scanning was conducted on a 3 Tesla Siemens Prisma MRI scanner at the Wellcome Centre for Human Neuroimaging, London. Structural images were acquired using an MPRAGE sequence (1x1x1 mm voxels, in plane FoV = 256 x256 mm^2). Functional scans were acquired using a 2D EPI sequence (3x3x3 mm voxels, TR = 3.36 s, TE = 30 ms 48 slices) (see Mazor *et. al*, 2020 for further details). Raw data were obtained from the first author upon request and preprocessed in the same manner as the main task in manuscript (see Materials and Methods for details). Preprocessed data and relevant scripts can be obtained from https://osf.io/8dbq2/.

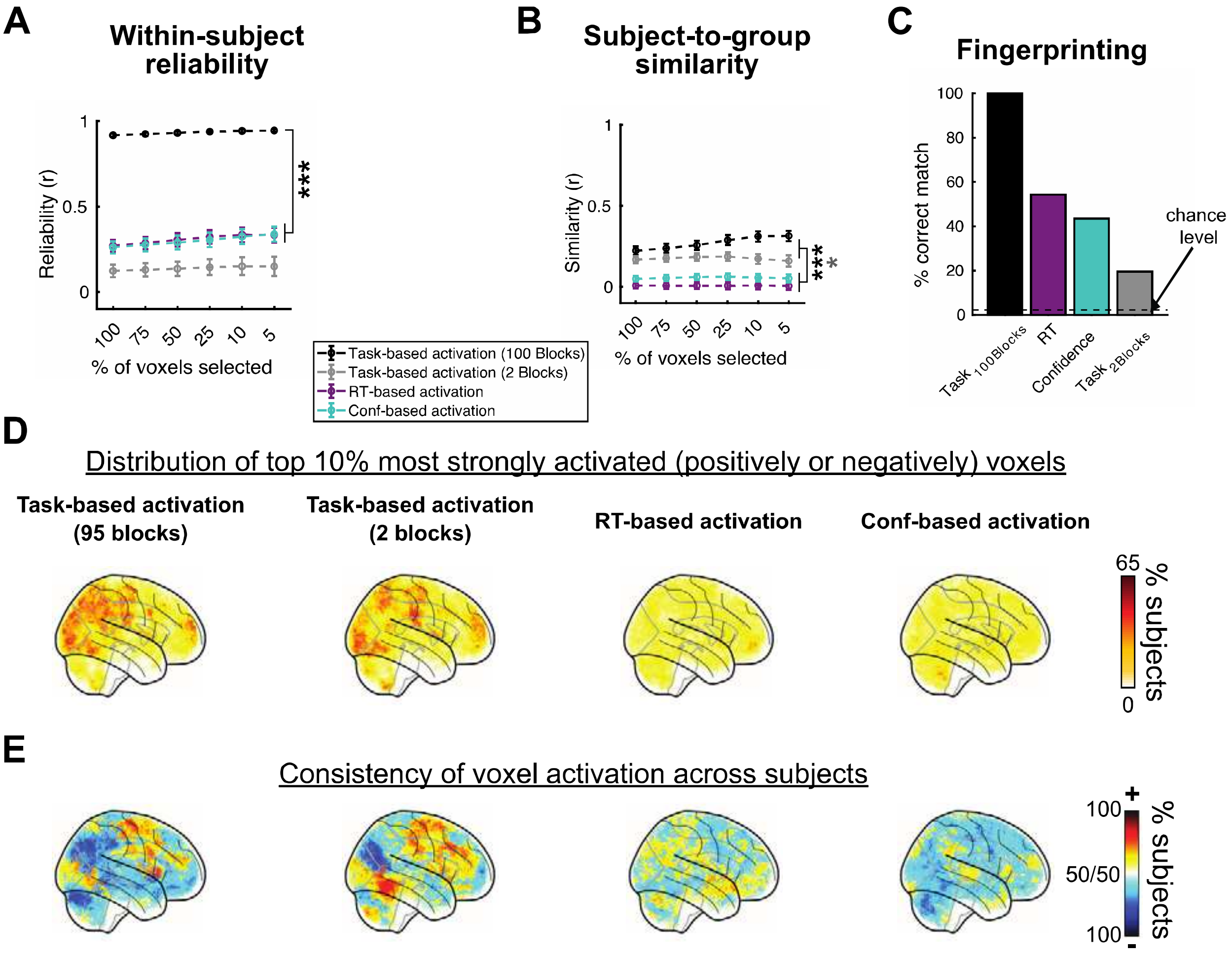


Figure S9. Analysis of the data from Mazor et al. (2020) with 2 trials per blocks. A) Within-subject reliability values as a function of the percent of most activated voxels selected. The analysis here was conducted on blocks composed of 2 trials resulting in 100 total blocks, thus roughly matching the number of blocks used in the analysis in the main text. For all voxel selection levels, the within-subject reliability was highest for the 100-block task-based analysis and lowest for the 2-block task-based analysis. B) Subject-to-group similarity computed in the same way as in **Fig. 3**. Both the 100- and 2-block task maps exhibited higher subject-to-group similarity compared to both the behavior-based maps. C) Fingerprinting results based on the maps produced by the odd and even blocks for each subject. D) Maps of the distribution of the top-10% most activated voxels showing strong areas of consistency for task-based but not for behavior-based analyses. E) Maps of voxel consistency computed as the proportion of subjects showing a positive or negative relationship between voxel activity and behavior. Task-based maps reveal areas of much higher consistency than behavior-based maps. Error bars show SEM; ***, p < 0.001; *, p < 0.05. All p-values are Bonferroni corrected.

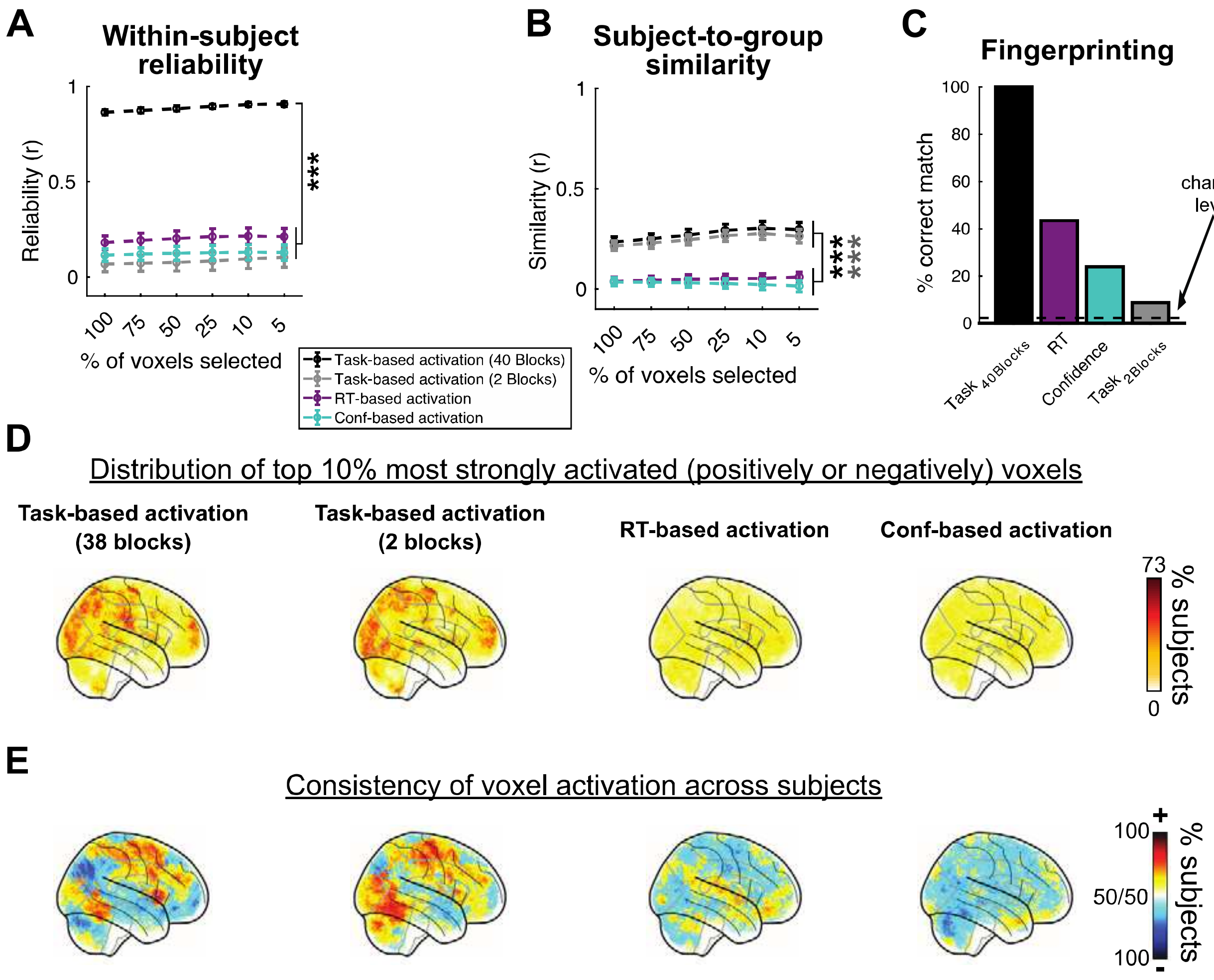


Figure S10. Analysis of the data from Mazor et al. (2020) with 5 trials per block. A) Within-subject reliability values as a function of the percent of most activated voxels selected. To increase the robustness of the estimation of the beta values, the analyses here were conducted on blocks composed of 5 trials resulting in 40 total blocks. For all voxel selection levels, the within-subject reliability was highest for the 40-block task-based analysis and lowest for the 2-block task-based analyses. B) Subject-to-group similarity computed the same as in **Fig. 3**. Both the 40- and 2-block task maps exhibited higher subject-to-group similarity compared to both the behavior-based maps. C) Fingerprinting results based on the maps produced by the odd and even blocks for each subject. D) Maps of the distribution of the top-10% most activated voxels showing strong areas of consistency for task-based but not for behavior-based analyses. E) Maps of voxel consistency computed as the proportion of subjects showing a positive or negative relationship between voxel activity and behavior. Task-based maps reveal areas of much higher consistency than behavior-based maps. Error bars show SEM; ***, p < 0.001. All p-values are Bonferroni corrected.