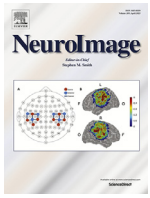
Stimulus-specific random effects inflate false-positive classification accuracy in multivariate-voxel-pattern-analysis: a solution with generalized mixed-effects modelling

NeuroImage 269 (2023) 119901

Contents lists available at ScienceDirect

NeuroImage

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



Stimulus-specific random effects inflate false-positive classification accuracy in multivariate-voxel-pattern-analysis: A solution with generalized mixed-effects modelling

Shogo Kajimura[a,*], Takahiro Hoshino[b], Kou Murayama[c,d]

[a] Faculty of Information and Human Science, Kyoto Institute of Technology, Matsugasakihashigami-cho, Sakyo-ku, Kyoto-shi, Kyoto 606-8585, Japan
[b] Faculty of Economics, Keio University, Tokyo, Japan
[c] Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Germany
[d] School of Psychology and Clinical Language Sciences, University of Reading, UK

ARTICLE INFO

Keywords:
Type-1 error
Group-level analysis
Multivariate-voxel-pattern-analysis
Generalized mixed-effects modelling
Random stimulus effect

ABSTRACT

When conducting multivariate-voxel pattern analysis (MVPA), researchers typically compute the average accuracy for each subject and statistically test if the average accuracy is different from the chance level across subjects (by-subject analysis). We argue that this traditional by-subject analysis leads to inflated Type-1 error rates, regardless of the type of machine learning method used (e.g., support vector machine). This is because by-subject analysis does not consider the variance attributed to the idiosyncratic features of the stimuli that have a common influence on all subjects (i.e., the random stimulus effect). As a solution, we proposed the use of generalized linear mixed-effects modelling to evaluate average accuracy. This method only requires post-classification data (i.e., it does not consider the type of classification methods used) and is easily implemented in the analysis pipeline with common statistical software (SPSS, R, Python, etc.). Using both statistical simulation and real fMRI data analysis, we demonstrated that the traditional by-subject method indeed increases Type-1 error rates to a considerable degree, while generalized mixed-effects modelling that incorporates random stimulus effects can indeed maintain the nominal Type-1 error rates.

1. Introduction

In recent years, there has been increased popularity in predictive modelling of functional MRI-based neuroimaging data, addressing the questions of whether patterns of brain activations encode sufficient information to discriminate different external outputs or mental states (Chavez and Wagner 2020; Taschereau-Dumouchel et al., 2020). One of the most commonly-used techniques is multi-voxel (or multi-variate, depending on the context) pattern analysis (MVPA), which was proposed as a powerful alternative to traditional univariate analyzes (Snoek et al., 2019; Stelzer et al., 2013). MVPA is a broad term and encompasses different classes of statistical analyzes but the current paper focuses on the techniques that aim to classify patterns of brain activations into categories (Weaverdyck et al., 2020). Specifically, using machine learning algorithms, MVPA considers the activation pattern of a set of voxels as features and examines whether the features can correctly classify externally defined categories such as the type of stimuli (Haxby et al., 2001), task conditions (Kamitani and Tong 2005) and memory contents (Harrison and Tong 2009).

The primary interest of MVPA is usually the statistical significance of the accuracy at the group level rather than the accuracies of individual subjects — i.e. "Is the mean classification accuracy significantly different from chance?" Importantly, although there are numerous algorithms to conduct pattern classification and compute accuracy scores, including support vector machine (Woo et al., 2014), logistic regression (Fairhall and Caramazza 2013), and Gaussian Naïve Bayes (Johnson et al., 2009), the way researchers conduct a group-level analysis is surprisingly uniform. Specifically, group-level statistical significance has been predominantly assessed by conducting a *t*-test on summary statistics, i.e., average percentage classification accuracy at the subject level (Haxby et al., 2001; Holroyd et al., 2018; Kliemann et al., 2018; Martin et al., 2016). However, there has been little awareness, at least in the literature of MVPA, that this conventional method to test group-level significance can cause serious inflation of Type-1 error rates (regarding the same problem in the context of univariate neuroimaging analysis, see Westfall et al. (2017)).

To explain the problem, let us describe a typical experiment employing MVPA. The experiment aims to examine whether the activation pat-

* Corresponding author.
E-mail address: kajimura.shogo.1204@gmail.com (S. Kajimura).

https://doi.org/10.1016/j.neuroimage.2023.119901.
Received 20 August 2022; Received in revised form 28 November 2022; Accepted 23 January 2023
Available online 24 January 2023.

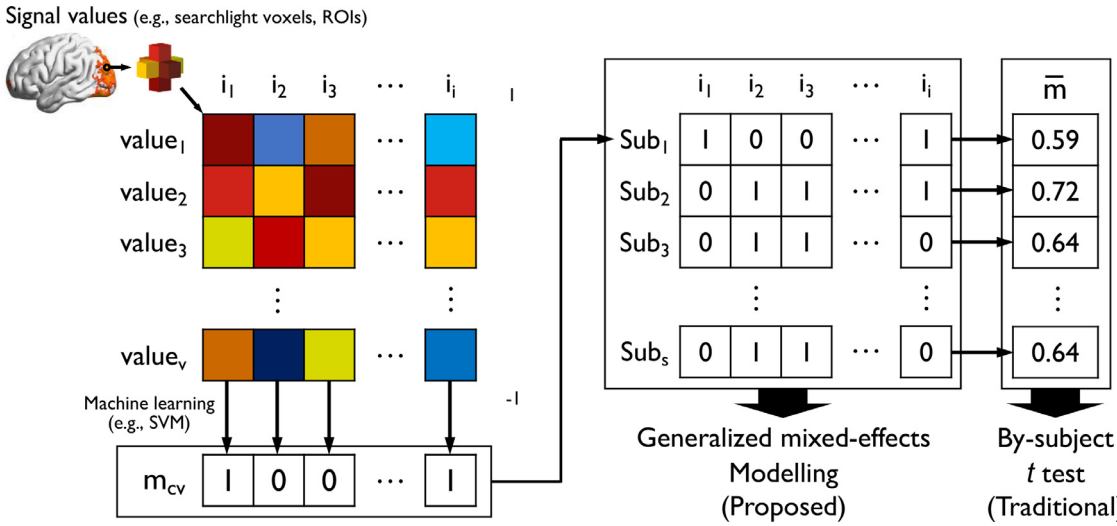


Fig. 1. Schematic procedure of the current analysis with real data. The left-hand side displays the procedure of searchlight analysis of one subject. The right-hand side displays the differences of generalized mixed-effects modelling and by-subject *t*-test. i_i: each stimulus, m_{cv}: result of machine learning classification (1: correct; 0: incorrect); $\bar{m}$: mean accuracy for each subject.

terns of visual areas can discriminate between natural scene categories (e.g., beaches vs. mountains). Subjects passively view the photographs of these natural scene categories; each category typically consists of multiple different stimuli (e.g., Hawaiian beach, Phuket beach, etc. and Mt. Aconcagua, Mt. Fuji, etc.), and all subjects view the same set of stimuli. MVPA tests whether the different categories can be classified by activation patterns of voxels or regions of interest (ROIs) for each stimulus using a machine learning algorithm (e.g., support vector machine) with cross-validation. More specifically, regardless of the types of machine learning algorithms used, the following steps are normally taken (Fig. 1). First, for each subject, using a cross-validation procedure, classification result is evaluated for each of the individual stimuli (correct classification = 1, misclassification = 0). Second, percentage classification accuracy is computed for each subject by aggregating the classification accuracy over the stimuli. Third, group-level effect, i.e., whether the classification accuracy is significantly above chance level, is tested by a one-sample *t*-test using subjects as the unit of analysis. We refer to this procedure as a *by-subject t*-test as the *t*-test was conducted over subjects. Note that voxel-level information is no longer needed to evaluate the group-level effect (Fig. 1, right).

What is the problem with this common procedure? The critical problem is that this procedure does not take into account the variability of stimuli in the group-level analysis (i.e., stimuli and subjects are crossed), large value of which causes underestimation of overall variability in the *t*-test and results in inflation of type-1 error rate. Unless the stimulus set is extremely homogeneous, it is likely that classification accuracy varies across stimuli: Averaged across subjects, some stimuli may have higher classification accuracy than other stimuli. Such inter-stimulus variations or stimulus effect may reflect random variation of the category exemplars, which is irrelevant to the essential features of the categories. For example, when distinguishing pictures of beaches and mountains, let us suppose that beach pictures happened to include the sun more frequently than the mountain pictures when randomly selecting pictures from the categories. The sun is irrelevant to whether a picture represents the mountain or beach, but because of this random incidence, MVPA is more likely to judge pictures with the sun as “beach”. As a result, beach pictures with the sun have higher accuracy than other beach pictures (and mountain pictures with the sun have lower accuracy than other mountain pictures). Of course, there are numerous other irrelevant features that happen to be included more frequently in one category than in the other, and it is almost impossible to experimentally control for them. The collective effects of such idiosyncratic (i.e. stimulus-specific) features of the stimuli on classification accuracy (or the extent to which a picture becomes more likely to be classified in one category) are called *random stimulus effects*.

Importantly, the by-subject *t*-test only makes use of the information about the average accuracy of each subject aggregated across stimuli, and is unable to take into account the random stimulus effect. This means that, the by-subject *t*-test cannot consider the possibility that MVPA picked up some idiosyncratic properties that happened to be present in the stimulus set used in the experiment (e.g., the sun happened to be more frequently observed in beach pictures than in mountain pictures). Statistical consequence of this negligence is not trivial. When the random stimulus effect is not appropriately modelled (i.e., the by-subject *t*-test is applied), the average classification accuracy itself is not biased, as random stimulus effects work both positively and negatively, cancelling each other out (Usami and Murayama, 2018). However, standard errors are underestimated, and as a result, the Type-1 error rate increases as the sample size increases, asymptotically reaching 100%. In fact, in previous simulation studies of behavioral data, with some realistic parameter settings, Type-1 error rate was inflated as high as 50% (Judd et al., 2012; Westfall et al., 2017; Murayama et al., 2014; Wickens and Keppel 1983; Clark 1973; Donnellan et al., 2022). This issue may be explained as follows: the by-subject *t*-test considers subjects as independently sampled; however, the random stimulus effect creates correlation between subjects (because subjects are presented with the same set of stimuli), which violates the critical assumption of the *t*-test. It is well known that the *t*-test is vulnerable to this assumption and violation of the independence assumption could increase Type-1 error rates to a considerable degree (Kenny and Judd 1986).

One intuitive way of addressing the issue is to conduct a *by-stimulus t*-test. Specifically, researchers can compute the average accuracy for each stimuli (not for each subject) over subjects, and conduct a *t*-test using the stimuli as the unit of analysis. This analysis assumes that stimuli are randomly sampled from the population and that the average accuracy includes sampling errors in the picture selection. Consequently, the analysis results can be generalized to the stimulus population. However, the problem with the by-stimulus *t*-test is that it does not consider the fact that subjects also have sampling errors, thus limiting the generalizability to the population of subjects (Clark 1973). The fundamental problem here is that accuracy data include two sources of sampling errors (in addition to random measurement errors): sampling variability

of subjects (the random subject effect) and sampling variability of stimuli (the random stimulus effect), and neither *t*-test can estimate both at the same time.

Recent advances in mixed-effects modelling allow researchers to model both sampling errors simultaneously, providing us with a way to make an appropriate generalization to the subject population as well as the stimulus population (Yu et al., 2022). Mixed-effects modelling can be seen as an extension of the conventional analysis of variance or *t*-test (McNabb and Murayama 2021), with considerable flexibility in modelling different sources of random effects (Baayen et al., 2008; Barr et al., 2013). Although the application of mixed-effects modelling with random stimulus effect has become increasingly popular in other fields such as psychology (Brauer and Curtin 2018; Meteyard and Davies 2020), the neuroimaging community was not well aware of the issue of random stimulus effect until recently (Westfall et al., 2017; Bedny et al., 2007). In fact, potential implications of random stimulus effects in the context of MVPA has little been discussed. We found it especially problematic because machine learning algorithms like the ones used in MVPA are deemed to be strongly influenced by idiosyncratic characteristics within the given stimulus set, thus potentially inflating Type-1 error rates to a considerable degree when a random stimulus effect is present. Some may argue that the cross-validation procedure can address this issue because the procedure examines whether the model can predict the outcome of out-of-sample stimuli. However, this is not the case. The random stimulus effect can be appropriately estimated and handled only when researchers use the information about *both* subjects and stimuli. When conducting machine learning methods in the context of MVPA, the analysis is normally conducted for each subject independently (or it is done for each stimulus independently), and does not make use of the information that stimuli are common across subjects.

In the current paper, we first use statistical simulation to demonstrate that presence of a random stimulus effect could considerably increase Type-1 error rates when a conventional by-subject *t*-test is used to test group-level statistical significance. We also show that generalized mixed-effects modelling with the correct random effect structure can prevent such an inflation of Type-1 error rate. Then we randomly relabel the existing fMRI data (i.e., we create a null dataset based on the empirical data) to demonstrate that ignoring the random stimulus effect can potentially lead to misleading conclusions about the performance of MVPA. We also show that the issue can be addressed by using mixed-effects modelling with random stimulus effect. As part of the online supplementary materials, we provide R code using the *lme4* package to implement the model that we used in this paper. It is worth noting that group-level analysis is needed only at the very end of the MVPA pipeline - after researchers have conducted all the pattern classification using a machine learning algorithm for each subject and stimulus (Fig. 1). Therefore, the implementation of mixed-effects modelling to test group-level statistical significance of classification accuracy is straightforward and easy to use for applied researchers. In fact, researchers can perform the analysis only with a few lines of code with standard software such as *SPSS, R*, and Python. Relatedly, the proposed method does not consider the type of classification method (e.g., support vector machine, elastic-net regression, deep learning, etc.), as long as the method produces the participant X stimulus post-classification result matrix (Fig. 1). What researchers should do is apply mixed-effects modeling to this matrix.

2. Methods

2.1. Simulation

We simulated data for a hypothetical experiment using a common subject × stimulus MVPA design in which each subject responds to several stimuli that are identical across subjects. Specifically, the hypothetical experiment had 20 or 40 subjects and 20 or 40 stimuli. The purpose was to examine whether brain activation patterns can be used to classify categories of stimuli. There were two stimulus categories (A and B) which had the same number of stimuli. We supposed that a machine learning algorithm classified each stimulus as either category A or category B and was given a binary value that represented whether the classification was correct or not (1 if correct and 0 if incorrect). In fact, we generated data from a model in which the result value was randomly generated, i.e., the true classification accuracy was at chance level. Then, significance of the overall classification performance was tested using a conventional by-subject *t*-test and a generalized mixed-effects model. Because the data were generated from the null model (i.e., chance classification accuracy), the proportion of statistically significant effects (where $p \leq 0.05$) observed in this statistical simulation can be interpreted as the Type-1 error rate (in terms of generalizing the results to the subject and stimulus populations).

There are different ways to generate data based on generalized mixed-effects modelling. Here, we generated the dataset from the following latent variable model:

$$y^*_{si} = S_s + I_i + e_{si}$$
$$y_{si} = \begin{cases} 1\ (Correct), & if\ y^*_{si} > 0 \\ 0\ (Incorrect), & if\ y^*_{si} \leq 0 \end{cases} \quad (1)$$

Where y_{si} represents the observed classification result of the *i*th stimulus of *s*th subject (1 = correct classification; 0 = misclassification). This observed result is a function of the latent continuous variable y^*_{si}, which represents the degree to which the *i*th stimulus of *s*th subject is correctly predicted. S_s is a random subject effect and I_i is a random stimulus effect, where $S_s \sim N(0,\ \tau^2)$ $(\tau = 0.3,\ 0.6,\ 0.9)$, and $I_i \sim N(0,\ \omega^2)$ $(\omega = 0.2,\ 0.4,\ 0.6)$. e_{si} is a random error term that follows a normal distribution, $e_{si} \sim N(0,\ \sigma^2)$ $(\sigma = 1)$. When y^*_{si} goes over the threshold (0), the stimulus is correctly classified into the true category. The equation essentially means that the correct classification of the *i*th stimulus of *s*th subject (y^*_{si}) depends on a sum of (a) the extent to which the stimulus is generally easy/difficult to predict (random stimulus effect), (b) the extent to which the subject generally provides good/bad result (random subject effect), and (c) random errors. Importantly, as the model does not contain any intercepts, the correct response is expected to be at chance level (i.e., 50%). In other words, this is a null model.

Note that this is not the only way to consider the accuracy metric (or random stimulus effect) in MVPA. For example, in the equation above, y^*_{si} may be considered as the extent to which the stimulus is more like the "first" category (e.g., "beach") out of two (e.g., "beach" and "mountain"). When y^*_{si} goes above a threshold (e.g., 0), the MVPA judges the stimulus as the first category ($y^{cat}_{si} = 1$ (First category); i.e., "beach"). If it is below the threshold, it judges the stimulus as the second category ($y^{cat}_{si} = 0$ (Second category); i.e., "mountain"). In this case, the result y_{si} should be determined in relation to the true category label. Specifically,

$$y^{cat}_{si} = \begin{cases} 1\ (First\ category), & if\ y^*_{si} > 0 \\ 0\ (Second\ category), & if\ y^*_{si} \leq 0 \end{cases}$$
$$y_{si} = \begin{cases} 1\ (Correct), & if\ y^{cat}_{si}\ matches\ the\ right\ category \\ 0\ (Incorrect), & if\ y^{cat}_{si}\ does\ not\ match\ the\ right\ category \end{cases} \quad (2)$$

If the true category is determined at random (e.g., even number stimuli are "beach" and odd number stimuli are "mountain"), y_{si} is expected be at the chance level. As we do not know which model is correct in reality, the following simulations generate data from both of these models. We manipulated the magnitude of random subject effect and random stimulus effect while fixing the variance of the random error term to 1.

For each of the generated data, we first computed the mean accuracy rate for each subject and applied a one-sample *t*-test to examine whether the mean is significantly different from 50% at the group level. We also applied the following generalized mixed-effects model with probit link function to the data using the lme4 package in *R*:

$$P(y_{si} = 1) = \Phi(\beta + S_s + I_i) \quad (3)$$

Where $\Phi()$ represents the cumulative standard normal distribution function, and $P(y_{si} = 1)$ is the probability of correct classification of the ith stimulus of sth subject. If the classifier has significantly better or worse performance than the chance level (50%), the absolute z-value of the intercept β should become significantly different from 0. This model is in line with the data generation model in Eq. (1). When the second data generation model (Eq. (2)) is correct, this analysis model does not accurately reflect the data generation model — the data may be better explained by a slightly more complicated model including true category membership as a fixed predictor with random slopes. Nevertheless, our simulation results below show that this simple analysis model still seems to protect the inflation of Type-1 error rates to a considerable degree in comparison to the by-subject t-test.

We repeated the data generation 1000 times and assessed Type-1 error rates both with a conventional by-subject t-test and mixed-effects model with random stimulus and random subject effects. It is important to note that the by-subject t-test is expected to inflate Type-1 error rates but it is not expected to bias parameter estimates. In other words, we can expect false-positive findings that are below as well as above the chance level (50%). As we are typically interested in accuracy above the chance level, Type-1 errors are counted only when accuracy was significantly more than 50%.

2.2. Assessment with a real dataset

We also examined the performance of the model with random stimulus effects (in comparison to the by-subject t-test) using a real fMRI dataset, the WU-Minn HCP Retest Data from the Human Connectome Project (HCP). The data contains 45 healthy subjects (14 males) in which subjects underwent an N-back working memory task while undergoing an fMRI scan (voxel size = 2.0 × 2.0 × 2.0, slice number = 72, time repetition [TR] = 720 ms, echo time [TE] = 33.1 ms, flip angle = 52°, 405 frames; see connectome database[1] for full description of fMRI data acquisition). Subjects responded "target" whenever the current stimulus was the same as the one presented N trials before (the data included a 0-back condition and a 2-back condition), and the target cue had 4 different stimulus types (pictures of faces, places, tools and body parts). For full description of the task design, see Barch et al. (2013). In the current study, for the purpose of simplicity, we only focused on the classification of the face and place trials. Trials for the tools and body parts were discarded.

2.2.1. Whole-brain MVPA with ROIs

The first analysis we conducted with the real dataset was whole-brain MVPA using the average activations of the ROIs as input features. We defined 90 ROIs based on automated anatomical labelling (Tzourio-Mazoyer et al., 2002). We computed average time-series activations for each ROI and standardize them over time. Then we further averaged the time-series to calculate the average signal for each trial (i.e., stimulus), obtaining a 90 (ROIs) × 40 (stimuli) feature matrix. We applied the linear support vector machine (SVM, soft margin parameter = 1) to classify the labels of stimulus types (place or face). For each subject, leave-one-out cross-validation was performed and a binary value (1 = correct classification; 0 = incorrect classification) was obtained for each stimulus to assess the classification performance with and without considering the random stimulus effect. We then applied a one-sample t-test and mixed-effects modelling with random stimulus effect (Eq. (3)) to examine whether the classifier can discriminate these categories better than the chance level (i.e., 50%).

Importantly, to evaluate the Type-1 error rates of these analyzes, we randomly permutated the labels of categories and applied SVM to classify these randomly permutated categories before conducting statistical tests (i.e., t-test and mixed-effects modelling), and repeated this procedure 1000 times. Because labels were randomly permutated, for each replication, a statistically significant result against the chance level can be considered as a Type-1 error. This allowed us to assess the impact of considering random stimulus effect on the potential inflation of Type-1 error rates with the real dataset. We also manipulated the number of subjects (22 or 45) and stimuli (20 or 40) to examine the influence of these factors. Note that this analysis is *not* purported to test whether the MVPA results for the original (i.e., pre-permutated) data are false positives. We do not know the ground truth of the original data; therefore, we cannot judge whether the analysis we applied to the data is a false positive. However, by permutating the original data, we can create datasets in which we know that category labels are assigned by chance. Using these permutated data, we can evaluate the performance of the proposed method.

2.2.2. Whole-brain voxel-level MVPA with searchlight analysis

We also examined the implications of not considering random stimulus effect in the whole-brain, voxel-level classification using searchlight analysis. In searchlight analysis, sets of voxels are defined by small "searchlight" regions centered on each voxel, and the activations of this small set of voxels are used as an input to a machine learning algorithm. The method has been proven to be useful in identifying locally informative areas with greater power and flexibility than univariate analyzes (Kriegeskorte et al., 2006; Stelzer et al., 2013) and thus has been widely applied (e.g., Kamitani and Tong 2005; Peelen et al. 2010; Ren et al. 2020). In the current analysis, like the previous analysis, we applied a linear SVM (soft margin parameter = 1) to classify the labels of stimulus types (places or faces). The searchlight was defined as a 3 mm radius sphere that contained 7 voxels and the estimated statistical value (z value for mixed-effects modelling and t value for t-test) using these voxels was assigned to the center voxel. To assess the classification performance, leave-one-out cross validation was performed for each subject and a voxel-wise binary value that represented the result of classification (1 = correct, 0 = incorrect) was obtained for each stimulus. Then, for each voxel, the group-level classification accuracy was evaluated using a one-sample t-test and mixed-effects modelling. The obtained statistical value maps from these two analyzes were then thresholded so that the family-wised error remained 0.05. The number of subjects (22 or 45) and stimuli (20 or 40) were again manipulated.

3. Results

3.1. Simulation analysis

The Type-1 error rates for the models with and without the random stimulus effect and for the by-subject t-test are summarized in Fig. 2. Note that the nominal Type-1 error rate is 2.5%, as we only counted false-positive findings that are above (but not below). The results for the data generated by Eqs. (1) and (2) are presented separately, but the results are consistent. It clearly shows the significant inflation of Type-1 error by ignoring the random stimulus effect: By-subject t-test generally showed Type-1 errors significantly beyond the nominal level. The inflation of Type-1 error rates increased when (1) the number of subjects increased, (2) the number of stimuli decreased, and (3) the SD of random stimulus effect increased. The SD of random subject effect is also inversely related to the increase in the Type-1 error rates (especially when the data are generated by Eq. (1)), because the increase of random subject effect masks the relative contribution of the random stimulus effect.

On the other hand, the Type-1 error rate of the mixed-effects model remains low, regardless of the number of subjects, number of stimuli, and the size of random stimulus or subject effects. These results indicate the effectiveness of mixed-effects modelling to test the significance of accuracy at the group level. There are two further observations. First, Type-1 error rates of mixed-effects modelling seem to remain low even if the data are generated from a model that has a different conceptualization of accuracy from that of the mixed-effects model we used (i.e.,

[1] https://db.humanconnectome.org/app/template/Index.vm

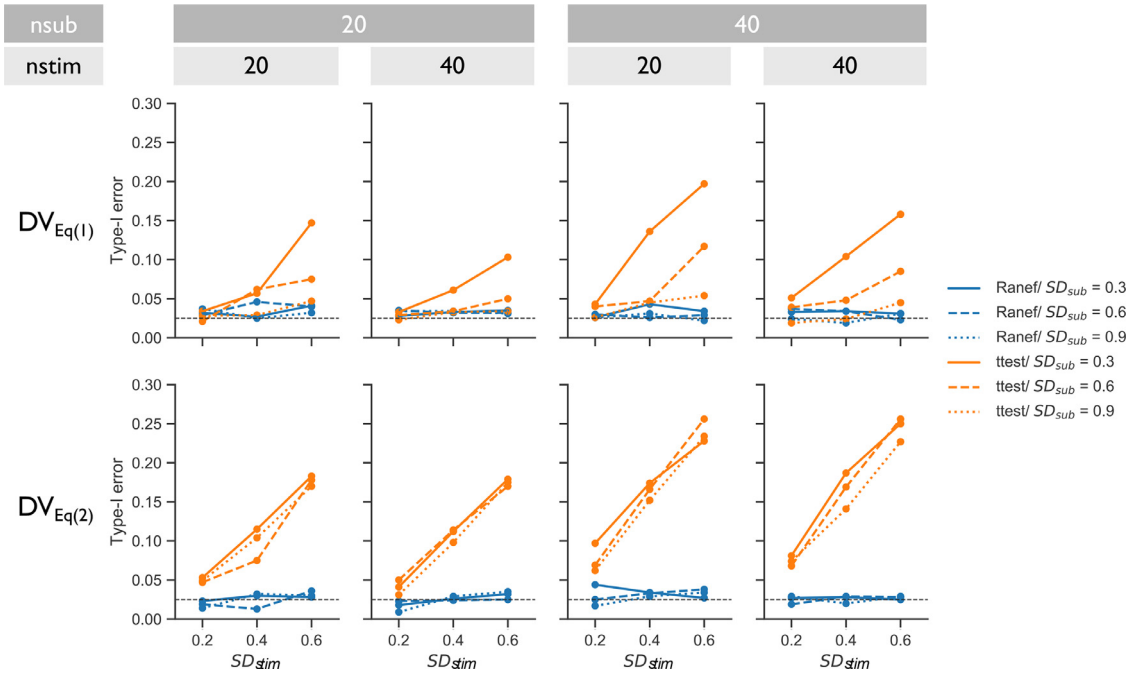


Fig. 2. Simulated Type-1 error ratio of mixed-effects modelling and *t*-test. $DV_{Eq(1)}$: dependant variables generated by Eq. (1), $DV_{Eq(2)}$: dependant variables generated by Eq. (2), Ranef: the model with the random stimulus effect, SD_{sub}: SD of random subject effect, SD_{stim}: SD of random stimulus effect.

Eq. (2)). These results demonstrate the robustness of the analysis model in Eq. (3) for potential model misspecification. Second, even though mixed-effects modelling generally keeps Type-1 error rate low, it still exhibits Type-1 error rate slightly higher than the nominal level (2.5%). We believe this is because sample sizes were generally small in our simulation (both in terms of number of subjects and stimuli), as is typically observed in neuroimaging studies. Previous studies showed that models with multiple random effects tend to underestimate standard errors (thus increasing Type-1 error rates) when sample size is small (McNeish 2017).

We also conducted an additional simulation to compare the statistical power of the by-subject *t*-test and mixed-effects modeling. Specifically, we set the true correct classification rate at 60% (when there are no random subject or stimuli effects) and examined whether the two methods can correctly indicate that the observed accuracy is significantly different from the chance level (50%). Because Type-1 and Type-2 error rates are inversely related, we expected a higher statistical power with the by-subject *t*-test than with the mixed-effects modeling. This expectation was supported (Fig. S1). The results showed that the *t*-test generally had higher power when the random stimulus effect variance or number of participants was large. However, under these conditions, we have a substantial increase in Type-1 error rates, and should not be seen as an advantage of the *t*-test approach. The increased power with the by-subject *t*-test is at the cost of the inflated Type-1 error rates, and, more importantly, without applying the mixed-effects modelling, researchers cannot know the extent of the inflation. However, these observations indicate that more samples are needed to apply MVPA to detect the effects typically observed in previous studies.

3.2. Assessment with a real dataset – ROI-level MVPA

Table 1 presents the Type-1 error rates of ROI-level MVPA based on the real data set with randomly permutated labels. With the traditional by-subject one-sample *t*-test, many of the permutated (randomized) data showed false positive statistically significant performance and the type-1 error rate significantly increased (error rate range: 0.382–0.421). The false-positive rate increased when sample size was larger (45 as opposed to 22) or the number of stimuli was smaller (20 as opposed to 40). This is consistent with the simulation results described above. Mixed-effects modelling, on the other hand, kept the Type-1 errors close to a nominal rate (alpha = 0.025, as we only counted statistically significant results that were above 50%, not below 50%), although the rate is still slightly anti-conservative (error rate range: 0.039–0.048). The slight anticonservatism is consistent with the simulation results. Nevertheless, mixed-effects modelling offers more protection from the inflation of Type-1 error rates, indicating the usefulness of these methods. Fig. 3 also showed the distribution of the *p* values from the two analyzes (with number of participants = 45 and number of stimuli = 40). Mixed-effects modelling showed a uniform distribution of *p* values, which is expected when the null hypothesis is correct (Wang et al., 2019); on the other hand, the conventional one-sample *t*-test showed a strongly skewed distribution where most of the *p* values clustered below the significance level (i.e. $p < .05$), despite category labels being randomly defined.

Table 1
Type-1 error of each number of Stimulus/Subject and method.

	Stim\Sub	22	45
Ranef	20	0.048	0.044
	40	0.041	0.039
t-test	20	0.392	0.421
	40	0.382	0.413

3.3. Assessment with a real dataset – voxel-level MVPA

The result of searchlight analysis to classify stimulus type conditions (places or faces) based on the real data set is summarized in Fig. 4. With the conventional one-sample *t*-test, most of the voxels were statistically significant, except in somatosensory and posterior cingulate regions, irrespective of the number of stimuli or subjects, meaning that almost all of the brain areas were able to classify these two categories. Using mixed-effects modelling with random stimulus effect, on the other hand, the classification performance was only statistically significant in visual

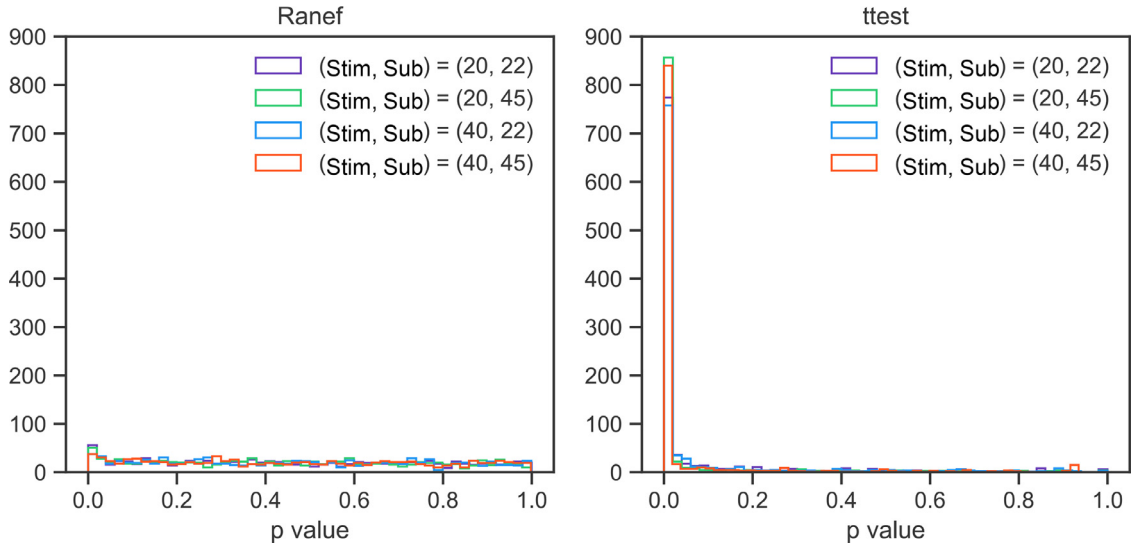


Fig. 3. Distributions of *p*-value for the permutation tests.

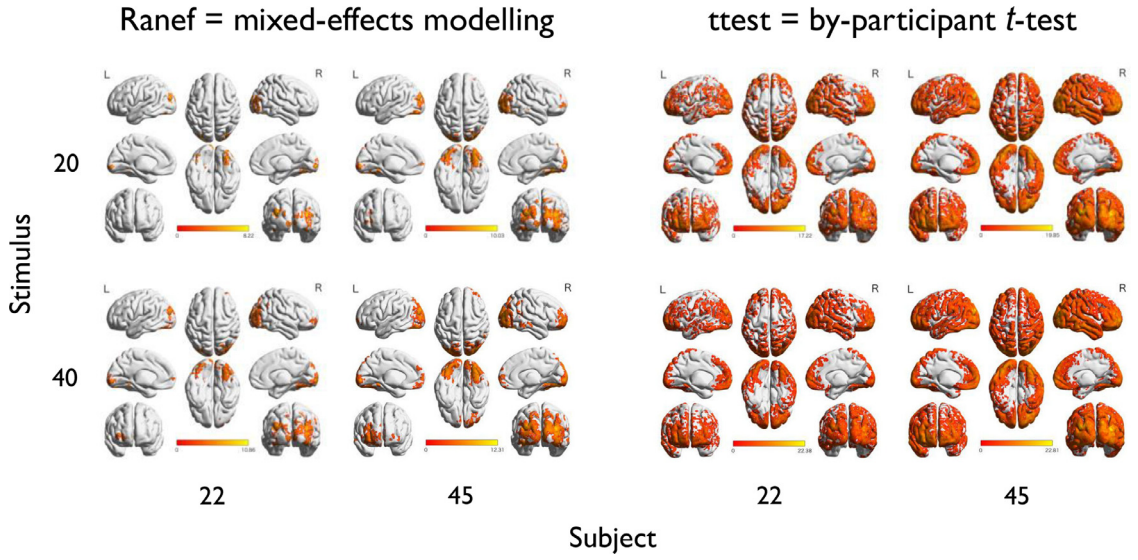


Fig. 4. The result of the searchlight analysis with and without the random stimulus effect. Ranef: the model with the random stimulus effect.

areas after family-wise error (FWE) correction ($\alpha = 0.05$, Bonferroni corrected, cluster threshold = 5), irrespective of the number of stimuli and subjects. In addition, the prefrontal regions reached statistical significance when the number of stimuli and/or subjects was larger. Again, as we do not know the ground truth (i.e., whether the searchlight voxels can truly distinguish place vs. face), we cannot determine whether these significant activations are false positives or not. However, these results indicate that researchers could draw completely different conclusions from the identical MVPA analysis depending on the way they evaluate group-level statistical significance.

4. Discussion

Although the importance of incorporating random stimulus effect has been well recognized in psychology and linguistics (Baayen et al., 2008; Barr et al., 2013), the neuroimaging community seems less aware of this issue, except only for a few instances (e.g., Westfall et al. 2017). Using simulated and real fMRI data, we demonstrated the potential deleterious influence of random stimulus effects when they are not appropriately modelled while testing group-level statistical significance of classification accuracy in MVPA. Specifically, we showed that presence of a random stimulus effect could considerably increase Type-1 error rates for the conventional by-subject *t*-test, especially when the number of subjects is large, the number of stimuli is small, or random stimulus effect is large (relative to random subject effect). We also demonstrated that generalized mixed-effects modelling with random stimulus effect could be an effective solution to this issue. In fact, in both simulation and real-data studies, mixed-effects modelling substantially prevented such an inflation of Type-1 error rate (although it was still slightly above the nominal level in some cases).

There are a few previous studies which criticized classification accuracy for the group-level significance test of MVPA (Gilron et al., 2017; Stelzer et al., 2013; Todd et al., 2013; Wang et al., 2020). A primary problem that has been identified is that confounds (e.g., random individual differences in experimental condition preference, familiarity, or difficulty) can artificially inflate significant effects when testing a group-level effect with summary statistics such as classification accuracy (Todd et al., 2013). This is because summary statistics discard the sign or direction of the underlying effects; thus, information regarding the sign or direction of the underlying effects is not utilized in the group-

level statistical test (Todd et al., 2013; Gilron et al., 2017). Another problem is that probability distribution of classification accuracy is non-Gaussian because of the low number of observations; as a result, several assumptions of the t-statistic are not met, rendering the procedure invalid (Stelzer et al., 2013). Our paper highlights an additional critical issue that has been overlooked in the literature.

Westfall et al. (2017) stated the influence of ignoring the random stimulus effect on traditional univariate fMRI studies and developed a method that can solve the problem of Type-1 error inflation (Westfall et al., 2017). Important differences between Westfall et al. (2017) and our paper is not only the unit of target variable (i.e., univariate and multivariate) but also the applicability of the proposed method in future studies. Most of the univariate fMRI studies generally employ a "two-step" procedure. In the two-step procedure, individual-level parameters are first estimated followed by the integration of these parameters (this is often conducted using a by-subject *t*-test). The two-step approach makes the estimation much less demanding (McNabb and Murayama 2021), but the method essentially precludes the possibility of incorporating random stimulus effect, as information about the stimulus is lost in the first step. Thus, the current predominance of the two-step approach creates a big challenge to apply a method proposed by Westfall et al. (2017). On the other hand, in the context of MVPA, mixed-effects modelling is applied to the subject × stimulus matrix of accuracy (Fig. 1), which is the output of machine learning classification. In other words, mixed-effects modeling does not need to be incorporated in the machine learning classification itself, and it also does not concern the type of machine learning models used in MVPA (e.g., support-vector machine, deep learning, etc.). These features make the implementation of the proposed method much easier (e.g., it can be performed with a few lines of additional code). In addition, as mixed-effects modeling is applied to a relatively small data matrix, it is computationally cheap (i.e., the results are obtainable in a few seconds).

Our research has also demonstrated counter-intuitive effects of sample size (i.e., number of subjects) and the inflation of Type-1 error rates when random stimulus effect is not appropriately modeled. Specifically, increasing the sample size substantially increases Type-1 error rates. When irrelevant features (e.g., "sun" in the example described in the introduction) appear to be present in one category more frequently than in the other category, even if the brain does not contain information to distinguish the categories (e.g., beach vs. mountain), MVPA (falsely) selects the irrelevant feature to make the classification look correct. With a larger sample size, this artefact increased correct classification because the irrelevant features are more likely to be deemed as statistically significant with the by-subject *t*-test. Mixed effects models, conversely, consider this artificial inflation of correct classification as part of the sampling error due to the selection of stimulus, thus preventing the increase in Type-1 error rate. Large sample size is generally encouraged to ensure sufficient statistical power and reproducibility (Calin-Jageman et al., 2019). However, unless the random stimulus effect is appropriately controlled, the collection of many subjects ironically results in the inflation of Type-1 error. However, this does not mean that researchers should stop collecting more data—the implication of our findings is simply that researchers should use appropriate statistical models to evaluate their data.

One critical question is, how common are large random stimulus effects? This depends completely on the design and stimuli of the experiments. If homogeneous stimuli are used or the same stimuli are used between the conditions, random stimulus effects are relatively small and would have a minimal impact on the conclusion drawn from a by-subject *t*-test. Our demonstration with a real fMRI dataset is just one instance in this respect and does not necessarily mean that the degree of impact is the same as in other fMRI studies. This point must be considered when interpreting the permutation analysis of the real fMRI dataset (Table 1). In each permutation, face and place pictures were intermixed in one category label and in the other category label, which were randomly determined. This indicates that the heterogeneity of pictures within a category label is relatively large (because there are pictures of both faces and places), likely resulting in the large underestimation of standard errors when a by-subject *t*-test was used. This is also the case for the following whole-brain analysis: even within face and place pictures, there are large heterogeneities. This dramatic change in the picture may reflect the heterogeneity of the stimuli used in the analysis.

At the end, we would like to make a few notes. First, while it is true that mixed-effects modelling can prevent the excessive inflation of Type-1 error rates, we should be careful not to interpret our results as suggesting the general prevalence of false-positive findings in MVPA studies. Our results simply showed that conventional modelling increases (false) significant results when the true classification accuracy is at the chance level. However, our results do not tell anything about the base rate of the situation when true classification accuracy is indeed at the chance level. Second, it should also be noted that the proposed method is only effective to address random stimulus effects in MVPA. Haynes (2015) for example, discussed several other limitations of MVPA such as accuracy interpretation, overfitting, and circular inference. However, these are separate issues, and researchers should take other appropriate measures to address them.

Finally, it is practically useful to consider when the conventional by-participant analysis does not inflate Type-1 error rates. Our simulation (Fig. 2) and previous studies have shown (Murayama et al., 2014) that using a larger number of stimuli reduces Type-1 error rates. While this sounds like a good solution, it is often difficult to increase the number of stimuli because of the limited time for scanning. Another, perhaps more realistic, alternative is to prepare a large number of stimuli and randomly distribute a small set of stimuli to different participants. Although not common in the literature, such a stimulus selection procedure reduces the dependency between participants due to the shared stimuli, minimizing the risk of by-participant analysis.

Data availability statement

Python and R script for *Assessment with a real dataset* is available as supplementary materials.

Declaration of Competing Interest

The authors declare no conflict of interest.

Credit authorship contribution statement

Shogo Kajimura: Software, Formal analysis, Resources, Data curation, Writing – original draft, Visualization. **Takahiro Hoshino:** Writing – review & editing, Supervision. **Kou Murayama:** Conceptualization, Software, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Data availability

I have shared the files as supplementary files.

Acknowledgments

This research was supported by the Leverhulme Trust (Grant Number RL-2016-030); Jacobs Foundation Research Fellowship; and the Alexander von Humboldt Foundation (the Alexander von Humboldt Professorship endowed by the German Federal Ministry of Education and Research).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2023.119901.

References

Baayen, R.H., Davidson, D.J., Bates, D.M., 2008. Mixed-effects modeling with crossed random effects for subjects and items. J. Mem. Lang. 59 (4), 390–412. doi:10.1016/j.jml.2007.12.005.

Barr, D.J., Levy, R., Scheepers, C., Tily, H.J., 2013. Random effects structure for confirmatory hypothesis testing: keep it maximal. J. Mem. Lang. 68 (3), 255–278. doi:10.1016/j.jml.2012.11.001.

Bedny, M., Aguirre, G.K., Thompson-Schill, S.L., 2007. Item analysis in functional magnetic resonance imaging. NeuroImage 35 (3), 1093–1102. doi:10.1016/j.neuroimage.2007.01.039.

Brauer, M., Curtin, J.J., 2018. Linear mixed-effects models and the analysis of nonindependent data: a unified framework to analyze categorical and continuous independent variables that vary within-subjects and/or within-items. Psychol. Methods 23 (3), 389–411. doi:10.1037/met0000159.

Calin-Jageman, R.J, Cumming, G., Greenland, S., Katz, P.S, Krafnick, A., Lakens, D., Mcshane, B., Peters, G.J., Pliske, R., Wagen, E.J., 2019. Novel tools and methods estimation for better inference in neuroscience significance statement. eNeuro 6 (4), 205–224. doi:10.1523/ENEURO.0205-19.2019.

Chavez, R.S., Wagner, D.D., 2020. The neural representation of self is recapitulated in the brains of friends: a round-robin FMRI study. J. Pers. Soc. Psychol. 118 (3), 407–416. doi:10.1037/pspa0000178.

Clark, H.H., 1973. The language-as-fixed-effect fallacy: a critique of language statistics in psychological research. J. Verbal Learn. Verbal Behav. 12, 335–359 file:///Users/baumann/Documents/Mendeley Desktop/Clark_1973_LanguageAsAFixedEffectFallacy.pdf.

Donnellan, E., Usami S., and Murayama K. 2022. "Random item slope regression: examining both similarities and differences in the association with individual items." PsyArXiv. doi:10.31234/osf.io/s6erz.

Fairhall, S.L., Caramazza, A., 2013. Brain regions that represent amodal conceptual knowledge. J. Neurosci. 33 (25), 10552–10558. doi:10.1523/JNEUROSCI.0051-13.2013.

Gilron, R., Rosenblatt, J., Koyejo, O., Poldrack, R.A., Mukamel, R., 2017. What's in a pattern? Examining the type of signal multivariate analysis uncovers at the group level. NeuroImage 146, 113–120. doi:10.1016/j.neuroimage.2016.11.019, September 2016.

Harrison, S.A., Tong, F., 2009. Decoding reveals the contents of visual working memory in early visual areas. Nature 458 (7238), 632–635. doi:10.1038/nature07832.

Haxby, J.V., Gobbini, M.I., Furey, M.L., Ishai, A., Schouten, J.L., Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal corten. Science 293 (28), 2425–2430. doi:10.4324/9780203496190.

Haynes, J.D., 2015. A primer on pattern-based approach to fMRI: principles, pitfalls, and perspectives. Neuron 87 (2), 257–270. doi:10.1016/j.neuron.2015.05.025.

Holroyd, C.B., Ribas-Fernandes, J.J.F., Shahnazian, D., Silvetti, M., Verguts, T., 2018. Human midcingulate cortex encodes distributed representations of task progress. Proc. Natl. Acad. Sci. U.S.A. 115 (25), 6398–6403. doi:10.1073/pnas.1803650115.

Johnson, J.D., McDuff, S.G.R., Rugg, M.D., Norman, K.A., 2009. Recollection, familiarity, and cortical reinstatement: a multivoxel pattern analysis. Neuron 63 (5), 697–708. doi:10.1016/j.neuron.2009.08.011.

Judd, C.M., Westfall, J., Kenny, D.A., 2012. Treating stimuli as a random factor in social psychology: a new and comprehensive solution to a pervasive but largely ignored problem. J. Pers. Soc. Psychol. 103 (1), 54–69. doi:10.1037/a0028347.

Kamitani, Y., Tong, F., 2005. Decoding the visual and subjective contents of the human brain. Nat. Neurosci. 8 (5), 679–685. doi:10.1038/nn1444.

Kenny, D.A., Judd, C.M., 1986. Consequences of violating the independence assumption in analysis of variance. Psychol. Bull. 99 (3), 422–431.

Kliemann, D., Richardson, H., Anzellotti, S., Ayyash, D., Haskins, A.J., Gabrieli, J.D.E., Saxe, R.R., 2018. Cortical responses to dynamic emotional facial expressions generalize across stimuli, and are sensitive to task-relevance, in adults with and without autism. Cortex 103, 24–43. doi:10.1016/j.cortex.2018.02.006.

Kriegeskorte, N., Goebel, R., Bandettini, P., 2006. Information-based functional brain mapping. Proc. Natl. Acad. Sci. 103 (10), 3863–3868. doi:10.1073/pnas.0600244103.

Martin, C.B., Cowell, R.A., Gribble, P.L., Wright, J., Köhler, S., 2016. Distributed category-specific recognition-memory signals in human perirhinal cortex. Hippocampus 26 (4), 423–436. doi:10.1002/hipo.22531.

McNabb, C.B., Murayama, K., 2021. Unnecessary reliance on multilevel modelling to analyse nested data in neuroscience: when a traditional summary-statistics approach suffices. Curr. Res. Neurobiol. 2, 100024. doi:10.1016/j.crneur.2021.100024, October.

McNeish, D., 2017. Small sample methods for multilevel modeling: a colloquial elucidation of Reml and the Kenward-Roger correction. Multivar. Behav. Res. 52 (5), 661–670.

Meteyard, L., Davies, R.A.I., 2020. Best practice guidance for lmms: best practice guidance for LMMs. J. Mem. Lang. 112, 104092.

Murayama, K., Sakaki, M., Yan, V.X., Smith, G.M., 2014. Type I error inflation in the traditional by-participant analysis to metamemory accuracy: a generalized mixed-effects model perspective. J. Exp. Psychol. Learn. Mem. Cogn. 40 (5), 1287–1306. doi:10.1037/a0036914.

Peelen, M. V., Atkinson, A. P., Vuilleumier, P., 2010. Supramodal representations of perceived emotions in the human brain. J. Neurosci. 30 (30), 10127–10134. doi:10.1523/JNEUROSCI.2161-10.2010.

Ren, J., Huang, F., Zhou, Y., Zhuang, L., Xu, J., Gao, C., … Luo, J., 2020. The function of the hippocampus and middle temporal gyrus in forming new associations and concepts during the processing of novelty and usefulness features in creative designs. Neuroimage 214, 116751. doi:10.1016/j.neuroimage.2020.116751.

Snoek, L., Miletić, S., Scholte, H.S., 2019. How to control for confounds in decoding analyses of neuroimaging data. NeuroImage 184, 741–760. doi:10.1016/j.neuroimage.2018.09.074, September 2018.

Stelzer, J., Chen, Y., Turner, R., 2013. Statistical inference and multiple testing correction in classification-based multi-voxel pattern analysis (MVPA): random permutations and cluster size control. NeuroImage 65, 69–82. doi:10.1016/j.neuroimage.2012.09.063.

Taschereau-Dumouchel, V., Kawato, M., Lau, H., 2020. Multivoxel pattern analysis reveals dissociations between subjective fear and its physiological correlates. Mol. Psychiatry 25 (10), 2342–2354. doi:10.1038/s41380-019-0520-3.

Todd, M.T., Nystrom, L.E., Cohen, J.D., 2013. Confounds in multivariate pattern analysis: theory and rule representation case study. NeuroImage 77, 157–165. doi:10.1016/j.neuroimage.2013.03.039.

Wang, B., Zhou, Z., Wang, H., Tu, X.M., Feng, C., 2019. The P-value and model specification in statistics. Gen. Psychiatry 32 (3), 1–4. doi:10.1136/gpsych-2019-100081.

Wang, Q., Cagna, B., Chaminade, T., Takerkart, S., 2020. Inter-subject pattern analysis: a straightforward and powerful scheme for group-level MVPA. NeuroImage 204 (2014), 1–26. doi:10.1016/j.neuroimage.2019.116205.

Weaverdyck, M.E., Lieberman, M.D., Parkinson, C., 2020. Tools of the trade multivoxel pattern analysis in FMRI: a practical introduction for social and affective neuroscientists. Soc. Cogn. Affect Neurosci. 15 (4), 487–509. doi:10.1093/scan/nsaa057.

Westfall, J., Nichols, T.E., Yarkoni, T., 2017. Fixing the stimulus-as-fixed-effect fallacy in task FMRI. Wellcome Open Res. 1, 1–24. doi:10.12688/wellcomeopenres.10298.1, May.

Wickens, T.D., Keppel, G., 1983. On the choice of design and of test statistic in the analysis of experiments with sampled materials. J. Verbal Learn. Verbal Behav. 22 (3), 296–309.

Woo, C.W., Koban, L., Kross, E., Lindquist, M.A., Banich, M.T., Ruzic, L., Andrews-Hanna, J.R., Wager, T.D., 2014. Separate neural representations for physical pain and social rejection. Nat. Commun. 5. doi:10.1038/ncomms6380, May.

Yu, Z., Guindani, M., Grieco, S.F., Chen, L., Holmes, T.C., Xu, X., 2022. Beyond t Test and ANOVA: applications of mixed-effects models for more rigorous statistical analysis in neuroscience research. Neuron 110 (1), 21–35. doi:10.1016/j.neuron.2021.10.030.

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.

Usami, S., Murayama, K., 2018. Time-specific Errors in Growth Curve Modeling: Type-1 Error Inflation and a Possible Solution with Mixed-Effects Models. Multivariate Behav. Res. 53 (6), 876–897. doi:10.1080/00273171.2018.1504273.