NeuroImage

www.elsevier.com/locate/ynimg
NeuroImage 35 (2007) 1093 – 1102

Rapid Communication

Item analysis in functional magnetic resonance imaging

Marina Bedny,[a,b,c,*] Geoffrey K. Aguirre,[a,d] and Sharon L. Thompson-Schill[a,b]

[a] *Center for Cognitive Neuroscience, University of Pennsylvania, USA*
[b] *Department of Psychology, University of Pennsylvania, USA*
[c] *Center for Non-Invasive Brain Stimulation, Beth Israel Deaconess Medical Center, Harvard Medical School, USA*
[d] *Department of Neurology, University of Pennsylvania, USA*

Received 25 May 2006; revised 4 January 2007; accepted 12 January 2007
Available online 8 February 2007

Behavioral and neuroimaging studies of cognition frequently test hypotheses regarding mental processing of different stimulus categories (e.g. verbs, faces, animals, scenes, etc.). The conclusions of such studies hinge upon the generalizability of their findings from the specific stimuli used in the experiment to the category as a whole. This type of generalizability is explicitly tested in behavioral studies, using "item analysis". However, generalizability to stimulus categories has up until now been assumed in neuroimaging studies, without employing item analysis for statistical validation. Here we apply item analysis to a functional magnetic resonance imaging study of nouns and verbs, demonstrating its theoretical importance and feasibility. In the subject-wise analysis, a left prefrontal and a left posterior–temporal region of interest showed putative grammatical class effects. An item-wise analysis revealed, however, that only the left posterior–temporal effect was generalizable to the stimulus categories of nouns and verbs. Taken together, the findings of the subject- and item-wise analyses suggest that grammatical-class effects in the left prefrontal cortex depend on the particular word stimuli used, rather than reflecting categorical differences between nouns and verbs. This empirical example illustrates that item analysis not only is sufficiently powered to detect task relevant changes in BOLD signal but also can make theoretically important distinctions between findings that generalize to the item populations, and those that do not.

In its most recent guide for authors, the influential psychological publication *Journal of Memory and Language* states: "Statistical analyses [are] normally expected to demonstrate generalizability of the results over both participants and items." These standards are similar to those of most psychology journals, but differ strikingly from the standards applied to functional neuroimaging research. Unlike behavioral studies of cognition, neuroimaging studies do not demonstrate the generalizability of results to item populations. In fact, generalizing findings in neuroimaging experiments from a *subject* sample to a population did not become common practice until the late 1990s. Previously, subjects were treated as a fixed-effect in neuroimaging studies. Fixed-effect analyses allow limited inference, as they do not permit findings to be generalized from a sample to a population of subjects (Friston et al., 1999; Holmes and Friston, 1998).

Most neuroimaging studies now employ random-effects analysis with respect to subjects, comparing the effect size to inter-subject variability in order to evaluate whether a finding generalizes to the subject population. However, neuroimaging studies are still victim to the "fixed-effects fallacy" (Clark, 1973) with respect to items. Here we demonstrate that neuroimaging studies can and should treat items as a random effect, when it is desirable to generalize findings to an item population.

When an independent variable is treated as "fixed" it is assumed that its levels are exhaustively sampled and are the only ones of interest. Thus, the levels of a fixed independent variable in one study will be the same as in all other studies, and the findings of a study need not be generalized to untested levels of that independent variable. In contrast, an independent variable is treated as "random" when its levels are randomly sampled from a larger population; the goal of the study is to make conclusions about the population of levels for this variable based on the results obtained from the sample of levels.

For example, the independent variable "gender" is typically treated as fixed. When studying gender differences, researchers test males and females and in doing so exhaust all the levels of the "gender" variable that are of interest. In studies of gender, researchers wish to infer that males and females differ in some way and do not wish to extrapolate to any untested levels of the gender factor. In contrast, subjects are generally treated as a random variable in such experiments because researchers wish to make inferences regarding a population of subjects, but only test a sample of that population (Kleinbaum et al., 1998).

The distinction between fixed and random variables has consequences for the ways in which data are analyzed. In treating a variable as random, we estimate the extent to which our effect of interest varies across the population of this random variable. Thus

[*] Corresponding author. Center for Cognitive Neuroscience, University of Pennsylvania, USA.
E-mail address: mbedny@bidmc.harvard.edu (M. Bedny).
Available online on ScienceDirect (www.sciencedirect.com).

1053-8119/$ - see front matter © 2007 Elsevier Inc. All rights reserved.
doi:10.1016/j.neuroimage.2007.01.039

we test the effect of our treatment of interest against the interaction of that treatment with the random variable. Said differently, we compare the size of our treatment effect to the size of the variability of the treatment effect over the different levels of our random variable. In contrast, when we treat a variable as fixed, we do not take into account the variation of our effect of interest across a population and thus are not able to say anything about whether our effect of interest is true of the population as a whole.

Treating the variable *item* as random is exactly parallel to treating the variable *subject* as random. When we treat subjects as a random variable, the effect size is compared to the variability of the effect across subjects (or the treatment by subject interaction). In item analysis we compare the effect size to the variation of the effect across items, this enables us to say whether the effect is true of the item population as a whole. The results of the subject- and item-wise analyses can be reported separately, as well as combined into a quasi-F statistic (F' or $minF'$) which, if significant, indicates generalizability across both participants and items (Clark, 1973; Coleman, 1964).

The use of item analysis in behavioral research was spearheaded by Herb Clark with the publication of his influential paper, "Language-as-a-fixed-effect fallacy: A critique of language statistics in psychological research" (Clark, 1973). Clark pointed out that findings were routinely generalized to stimulus populations without statistical basis. For example, a researcher interested in demonstrating that words with multiple meanings (homonyms such as "bank," that can refer to a place to store money or to the side of a river) are read more slowly than words with a single meaning (non-homonyms such as "cat") would choose a sample of each type of word and present these to a sample of subjects. Upon finding that with this sample of subjects and sample of items, homonyms are indeed read more slowly than non-homonyms, they would then compare the size of this effect to its variability across subjects. If the effect were reliable, they would conclude that homonyms are read more slowly than non-homonyms. Such an inference, however, is unsupported without also conducting an item analysis. Having performed only the subject-wise analysis, the researcher in the homonym example can only support the claim that *this particular set* of homonyms is read slower than *this particular set* of non-homonyms.

Similar instances of studies where item analysis is required are readily found in neuroimaging research. For example, a researcher interested in showing that a brain area is more important for processing animals than tools would select a sample of each type of stimulus (in the form of pictures or words), and present these to a sample of subjects. If, on average, tool stimuli elicited more activity than animal stimuli in a particular brain area, this effect would be statistically evaluated by comparing its size to its variability across subjects. If the effect were reliable across subjects, the researcher would conclude that area X is more important for processing and/or representing the category of tools than the category of animals. However, like the homonym example (above) this conclusion goes beyond what was tested using an estimate of across-subject variability.

All that can be concluded, based on this subject-wise analysis, is that if exactly the same set of animals and tools were tested on a different group of subjects, a similar outcome is likely to be found. One cannot conclude that a different sample of tools and animals tested on the same group of subjects would yield similar results. This is analogous to a fixed-effects analysis across subjects, which does not provide evidence regarding the reliability of an effect across the subject population. Using solely a subject-wise analysis, the researcher is unable to test the hypothesis originally intended: the population of tool stimuli activates area X more than the population of animal stimuli. This example illustrates why a subject-wise analysis is not sufficient for many neuroimaging studies.

To illustrate the theoretical importance and feasibility of item analysis in functional magnetic resonance (fMRI) research, we applied item analysis to an event-related fMRI experiment that examined the neural correlates of nouns and verbs (Bedny and Thompson-Schill, 2006). The frequent comparison of nouns and verbs in neuroimaging studies has yielded equivocal results, and to date none of the conclusions of these classes of items has been supported by relevant item analyses. First, we demonstrate that item analysis can be as sensitive as subject-wise analysis to task-relevant changes in activity. We then describe examples and theoretical implications of findings that are reliable across subjects and items, and contrast these to findings that are reliable across subjects but not items. Our discussion considers both the particulars of this experiment as well as the more general role that item analysis may play in neuroimaging studies of cognition. We also consider those circumstances in which item analysis is not applicable.

Methods

Participants

Thirteen subjects (eight females) participated in the experiment. Their mean age was 26 years (range: 20–31). All participants were right-handed and had spoken only English until at least age 5. None of the participants suffered from psychiatric or neurological disorders or had ever sustained head injury. All subjects gave informed consent to participate in the study and were paid $15 per hour for taking part in the experiment. Subjects came into the laboratory one day before the fMRI scan for a prescreening to ensure they could safely participate in an fMRI study and to become familiarized with the task through a 5-min practice session.

Behavioral procedure

The experiment consisted of 300 trials (200 word trials and 100 non-word trials). The word trials consisted of 100 noun trials and 100 verb trials. Nouns and verbs were matched on letter length, frequency (Francis and Kucera, 1982), and imageability (Bedny and Thompson-Schill, 2006). Imageability ratings for the stimuli were obtained from a web-survey, each word was rated by at least 20 participants. We selected stimuli that were highly biased to be interpreted as either nouns or verbs based on usage in the English Language: stimuli used as verbs occur at least 10 times more frequently as verb than as nouns, and stimuli used as nouns occur at least 10 times more frequently as nouns (Francis and Kucera, 1982).

On each word trial, participants saw a single word followed by a pair of words. The task was to decide which pair member was most similar in meaning to the immediately preceding single word. Nouns appeared with the article "the" and verbs with the article "to." All stimuli were presented visually on a black screen. Each run contained 10 non-word and 20 word trials.

Stimuli appeared in a pseudo-random order with no more than four trials of the same type in a row (word or non-word trials). Across subjects, the order of stimuli within a run and run order was randomized.

Each trial lasted 15 s and consisted of a target word, a word pair and a jittered inter-trial interval (ITI). Subjects were instructed to press the left or the right button to indicate which of the pair words was most similar in meaning to the target word. On each word trial, a target-word appeared for 2, 4 or 6 s. The jitter in the length of the target word presentation allowed us to better separate the hemodynamic response to the target words from that to the pair of words that followed. After the target word was removed from the screen, a pair of words appeared for 2 s or until the subject made a response. If the subject made a response before the 2 s elapsed, the word pair was replaced by a crosshair for the remainder of the 2-s period. The word pair was followed by a jittered ITI that lasted 7, 9 or 11 s. The length of the ITI was yoked to the length of the target-word presentation such that the entire trial duration was always 15 s. During the ITI, a cross hair appeared in the center of the screen. Subjects were instructed to fixate on the crosshair during the ITI. The ITI fixation was used as the baseline for the purposes of data analysis.

Non-word trials (which were randomly interspersed throughout the experiment) were similar in event sequence to the word trials. The non-words were orthographically legal sequences, and were matched to the word stimuli on length in letters (mean=5.8, SD=1.5) and, like the word stimuli, were preceded by “the” or “to”. The non-words appeared in yellow on a black background. The stimuli were presented in 10 runs of 30 trials each. Prior to the experiment, subjects were instructed that yellow font indicated a non-word trial. On non-word trials, subjects were instructed to select the pair member that was identical to the initial non-word target.

Behavioral data were only available for 6 of the 13 subjects (all comparisons were evaluated using the within-subject Wilcoxon's Signed Rank Test due to this small sample). Participants were significantly more accurate ($z=-10.5$, $p<0.05$) and faster ($z=-10.5$, $p<0.05$) to make a decision for non-word (mean=99.8%, SD=0.5%; mean=684 ms, SD=124 ms) than word trials (mean=96.8%, SD=1.6%; mean=1117 ms, SD=157 ms). The noun and verb responses were not different in either accuracy ($\text{mean}_{\text{nouns}}=97\%$, SD=1%; $\text{mean}_{\text{verbs}}=97\%$, SD=3%) or reaction time ($\text{mean}_{\text{nouns}}=1102$ ms, SD=185 ms; $\text{mean}_{\text{verbs}}=1131$ ms, SD=137 ms) ($p>0.40$).

fMRI data acquisition

Structural and functional data were collected on a 3.0 Tesla Siemens Trio scanner using a transmit/receive gradient head coil. High-resolution T1-weighted structural images were collected in 160 axial slices and near isotropic voxels (0.9766 mm × 0.9766 mm × 1.0000 mm; TR=1620 ms, TE=3 ms, TI=950 ms). Functional, blood-oxygenation-level-dependent (BOLD), echoplanar data were acquired in 3 mm isotropic voxels (TR=3000, TE=30). BOLD data were acquired in 42 contiguous axial slices, in an interleaved fashion with 64×64 in-plane resolution using a prospective motion correction (PACE) sequence. The functional data were collected in 10 runs of 7 min and 14 s each. The first 24 s of each run consisted of a “dummy” gradient and radio frequency pulse to allow for steady-state magnetization.

Image processing and data analysis

Common aspects of subject and item analyses

Off-line data analysis was performed using VoxBo (www.voxbo.org) and SPM2 (http://www.fil.ion.ucl.ac.uk/) software. Using VoxBo, data were sync-interpolated in time to correct for the slice acquisition sequence. Data were then motion corrected with a six-parameter, least-squares rigid-body realignment routine using the first functional image as a reference. The data were smoothed with an 8×8×8 mm^3 full-width at half maximum Gaussian smoothing kernel. Data were then normalized in SPM2 to a standard template, in Montreal Neurological Institute (MNI) space. Normalization maintained 3-mm isotropic voxels and used 4th degree B-spline interpolation.

First-level analysis was performed using the modified (for serially correlated error terms) general linear model (Worsley and Friston, 1995; Zarahn et al., 1997a). Covariates of interest were convolved with a standard hemodynamic response function (HRF) (Aguirre et al., 1998). Neural activity was modeled as a brief impulse at stimulus onset (Zarahn et al., 1997a). Nuisance covariates were included for effects of scan and global signal. Time series data were subjected to a high-pass (.0177 Hz) filter, and serial correlation of error terms was modeled as previously described (Zarahn et al., 1997b).

BOLD signal differences between words and non-words as well as nouns and verbs were evaluated through second level (random-effects) analyses. Second-level analyses were performed on the β-values obtained from the first-level analysis. For whole-brain analyses, the false positive rate was controlled ($\alpha<0.05$ corrected for multiple comparisons with a minimum cluster size of 15 voxels) by performing 2000 Monte Carlo permutation tests on the data (Nichols and Holmes, 2002). In anatomical ROI analysis time series were averaged over the entire ROI to assess significance.

Anatomical ROIs were created based on findings from previous neuroimaging studies (Davis et al., 2004; Perani et al., 1999; Tyler et al., 2004; Wise et al., 2000). Perani et al. (1999) reported greater activity for verbs than nouns in a region on the border of the left inferior and middle frontal gyri (LIFG/MFG, $X=-28$, $Y=28$, $Z=28$) during a lexical decision task. Davis et al. (2004) found greater activity for verbs than nouns in the posterior aspect of the left superior temporal gyrus (LSTG, $X=-54$, $Y=-36$, $Z=21$) during a one-back synonym-monitoring task. Anatomical ROIs were created by growing spheres, 5 mm in radius, centered on the reported peaks of activation (converted to MNI if reported in Talairach). The second-level analysis of the Verb–Noun contrast was performed upon the average data vector within these ROIs, as well as upon the whole-brain dataset.

Subject-wise analysis

First-level analysis was performed by modeling BOLD signal for each subject as a function of condition, on each trial. Covariates were created for each event type including: non-word target–word, noun target–word, verb target–word, non-word pair, noun pair, verb pair and baseline. Two additional covariates of interest were used to model the imageability of nouns and verbs separately; each of these was mean centered. These covariates were included to test hypotheses that are not the subject of the present paper and are discussed in detail in Bedny and Thompson-Schill (2006). A covariate was included to model the amount of time that each word was presented (2, 4 or 6 s).

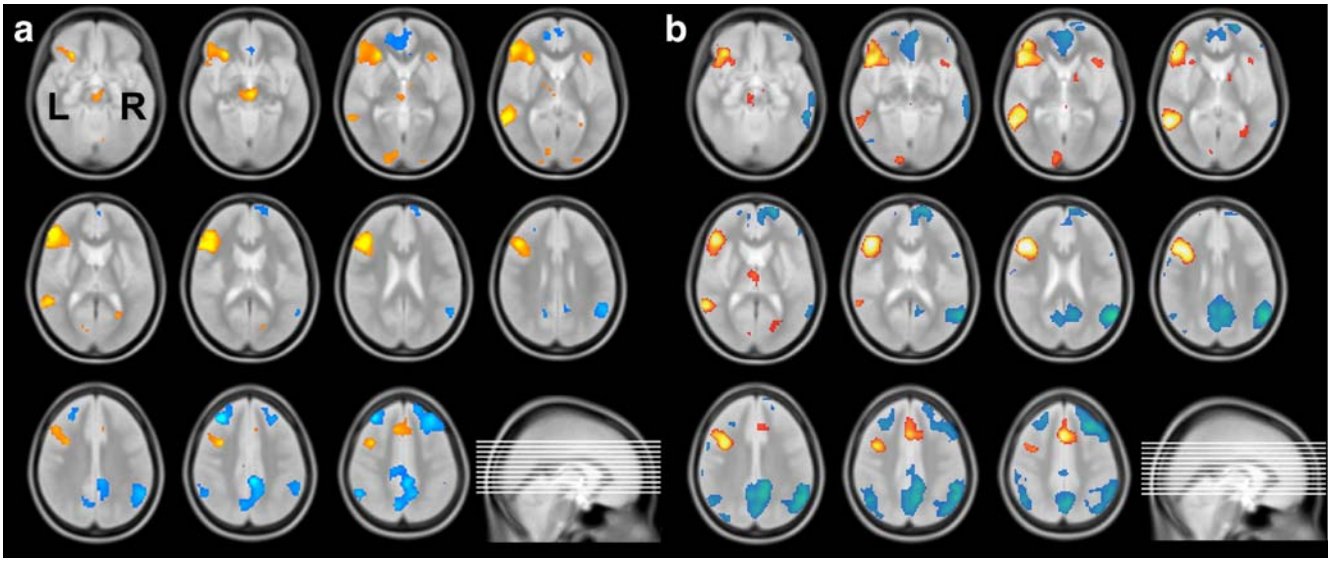

Fig. 1. Words–Non-words contrast for subject- and item-wise analyses. (a) The results of the whole-brain, subject-wise analysis for the Words–Non-words contrast [map-wise thresholds $t(12)=4.59$, $p<0.05$, minimum of 15 contiguous voxels]. (b) Whole-brain, item-wise analysis results for the Words–Non-words contrast [map-wise thresholds $t(297)=3.85$, $p<0.05$, minimum of 15 contiguous voxels]. Warm colors represent voxels more active for words than non-words; cool colors represent voxels more active for non-words than words. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

The second-level model treated subjects as a random variable: the variability of an effect across subjects was used to calculate a single-sample, t-statistic at each voxel (Friston et al., 1999; Holmes and Friston, 1998).

Item-wise analysis

For the first-level analysis we created a condition function with 299 covariates, modeling each word presented during the course of the experiment.[1] This analysis yielded 299 standardized β-maps for each subject. At this point, it is possible to enter these as the dependent variable into a second-level, random-effects item analysis; however, this is computationally burdensome. A mathematically equivalent solution is to average the β-maps for each item across subjects and perform second tier analysis on the results. This procedure consolidated the data from the thirteen subjects into a set of 299 β-maps; each β-map contained an average β-value across subjects at every voxel. These β-maps served as the dependent data for a second-tier analysis, which treated items as a random variable.

To test the significance of the Words–Non-words contrast, a two-sample t-statistic was computed at every voxel. The numerator of the t-value is identical to that used in the subject-wise analysis (the difference between the average β-value for the words and the average β-value of non-words). The denominator is the standard error of the effect across items.[2]

Multiple regression was used to assess the Verbs–Nouns contrast. Covariates included: grammatical class, noun imageability, verb imageability,[3] noun target–word presentation time, and verb target–word presentation time (2, 4 or 6 s) with BOLD response to word items as the dependent measure. We report the significance of the partial correlation coefficient for the Verb–Noun contrast.

Combining the subject- and item-wise analyses

In psychological research it is standard practice to report both the subject- and item-wise analyses. As noted in the introduction, these analyses can also be combined into a quasi-F statistic (*minF′*), which, if significant, allows generalization to items and participants. In applying item analysis to neuroimaging research it is possible to combine the subject- and item-wise analyses into the quasi-F statistic for a region of interest (ROI) analyses. However, calculating a significance threshold for a spatially distributed minF' statistical map is problematic. As the degrees of freedom for the minF' distribution vary as a function of the item- and subject-wise individual F-values, the degrees of freedom of the *minF′* statistic (and thus its distribution) vary across voxels. This violates the assumption of stationarity in Gaussian Random Field (GRF) theory. Nor can the permutation approach to multiple comparisons be readily applied to calculate a minF'. In permutation analysis subject-wise significance is assessed by permuting data over subjects and item-wise significance is assessed by permuting data over items.

[1] The intent was to choose 100 verbs and 100 nouns; however, subsequent to data collection, one noun was found to repeat in the stimulus set.

[2] The standard error is that of conventional t-statistic. The pooled variance of the word and non-word β's across items multiplied by the square root of $1/N_{\mathrm{w}}+1/N_{\mathrm{nw}}$ (where N_{w} is the number of word stimuli and N_{nw} is the number of non-word stimuli). Assuming equal variance of the BOLD effect across words and non-words, with degrees of freedom (df)= $N_{\mathrm{w}}+N_{\mathrm{nw}}-2$.

[3] Consideration of the effect of "imageability" on the BOLD response is outside the scope of the present paper, thus we treated "imageability" as a covariate of no interest (see Bedny and Thompson-Schill, 2006 for discussion of imageability effects). However, it is important note that item analysis is as necessary for continuous predictors as it is for discrete ones such as the noun/verb distinction. Item analysis for continuous variables proceeds in the same fashion as for discrete predictors i.e. by collapsing across participants and performing an item-wise regression to assess significance across items.

Calculating a $minF'$ with the permutation approach is not straightforward as it is not possible to permute simultaneously over subjects and items. Nor can we simply use the $minF'$ formula to derive a joint subject- and item-wise threshold from F_1 and F_2. Doing so would generate a statistic that is always smaller than either F_1 or F_2, and thus fails to ensure significance across both items and subjects.

Thus, we report the quasi-F statistic only for the ROI analyses. We return to the topic of the quasi-F statistic in the Discussion section.

Results

Subject-wise analysis

Words–Non-words contrast

The effect of reading Words versus Non-words was assessed on the whole-brain level across the subject population [$t_1(12)=4.59$, $p<0.05$, with a minimum of 15 contiguous voxels (corrected for multiple comparisons)]. These results are depicted in Fig. 1a and summarized in Table 1. Multiple regions were significantly more active for words than non-words. The largest and most robust areas of activation were in the left inferior frontal gyrus [LIFG, 382.5 cm^3, $t_{1\max}(12)=17.17$, $X=-51$, $Y=33$, $Z=9$], the right and left medial–superior frontal gyri/anterior cingulate [mSFG/AC 93.6 cm^3, $t_{1\max}(12)=8.25$, $X=-9$, $Y=21$, $Z=48$], and the left middle and superior temporal gyri [LMTG/STG 62.1 cm^3, $t_{1\max}$ (12) = 13.44, $X=-63$, $Y=-51$, $Z=3$].

Table 1
Results for whole-brain, subject-wise words vs. non-words contrast

Brain region	Peak voxel t-value	Cluster size (cm^3)	X	Y	Z
Words > Non-words					
Left inferior frontal and middle frontal gyri (47/45/46)	17.17	382.5	−51	33	9
Left middle and superior temporal gyri (37/21/22)	13.44	62.1	−63	−51	3
Brainstem, midbrain	10.21	74.4	3	−21	−9
Bilateral medial superior frontal gyri/anterior cingulate (6/8)	8.25	93.6	−9	21	48
Left lingual gyrus and cuneus (17)	7.46	36.6	−18	−99	0
Right posterior cingulate and occipital lobe (30/31/18)	6.97	21.90	24	−66	12
Right middle occipital gyrus (18)	5.59	15.90	30	−99	3
Non-words > Words					
Left middle and superior frontal gyri (9)	14.97	66.3	−33	33	39
Right posterior cingulate/ precuneus (31/7)	11.30	202.5	6	−42	39
Right middle frontal gyrus (9)	9.38	191.4	33	33	42
Left inferior parietal lobule (40)	8.98	30.3	−54	−57	45
Right supramarginal gyrus (40)	8.30	103.8	54	−54	27
Left inferior temporal gyrus (20)	7.88	19.2	−57	−24	−27
Left anterior cingulate, medial frontal gyrus (32/10)	6.54	115.8	−18	45	−6

Table 2
Results for whole-brain, item-wise words vs. non-words contrast

Brain region	Peak voxel t-value	Cluster size (cm^3)	X	Y	Z
Words > Non-words					
Left inferior frontal gyrus (46/ 45/9)	11.65	396.9	−51	21	24
Left middle temporal gyrus (21/22)	10.79	110.4	−60	−48	6
Left and right medial superior frontal gyrus (6/8)	9.80	132	−3	15	51
Right occipital lobe (18)	7.22	26.7	6	−90	−21
Left occipital lobe (17/18)	5.86	27.3	−15	−93	−3
Right sublobar/insula (47)	5.36	19.5	36	21	0
Right occipito-temporal cortex	5.23	19.2	30	−60	9
Right cerebellum	5.22	11.4	21	−87	−42
Left thalamus	5.09	10.8	−6	−9	12
Right lentiform nucleus	4.88	8.1	12	3	3
Left brainstem	4.78	9	−9	−18	−12
Non-words > Words					
Right superior/middle frontal gyri (8/6)	9.00	202.2	27	32	51
Right superior temporal/ supramarginal gyri (39/40)	8.92	300.3	12	22	32
Right cingulate/precuneus (31)	8.05	336.6	9	−42	39
Right superior frontal gyrus (10)	7.37	205.5	15	66	18
Right middle temporal gyrus (20/21/37)	6.98	96.9	57	−42	−12
Left inferior parietal lobule (40/39)	6.57	112.2	−51	−60	45
Left inferior/middle temporal gyri (20/21)	6.52	42.6	−57	−27	−21
Left superior/middle frontal gyri (8/9)	5.90	65.7	−24	39	48
Right middle frontal gyrus (11)	5.76	7.8	39	51	−9
Right inferior frontal gyrus (46)	5.07	6	51	42	9
Left inferior temporal gyrus (37)	4.37	4.5	−48	−72	−3

Several regions were also more active for Non-words than Words. The most robust areas of activation were in the left middle and superior frontal gyri [66.3 cm^3, $t_{1\max}(12)=14.97$, $X=-33$, $Y=33$, $Z=39$] and the bilateral posterior cingulated and precuneus [202.5 cm^3, $t_{1\max}(12)=11.30$, $X=6$, $Y=-42$, $Z=39$] (see Table 1 for list of active regions).

Verbs–Nouns contrast

The effect of reading verbs as compared to nouns was assessed within anatomically defined regions of interest across the population of subjects. The LSTG ROI showed significantly greater activity for verbs than nouns [$t_1(12)=5.20$, $p<0.001$]. In contrast to previously reported findings, the left inferior frontal/ middle frontal gyri (LIFG/MFG) ROI showed greater activity for nouns than for verbs in the present study [$t_1(12)=-2.68$, $p=0.01$].

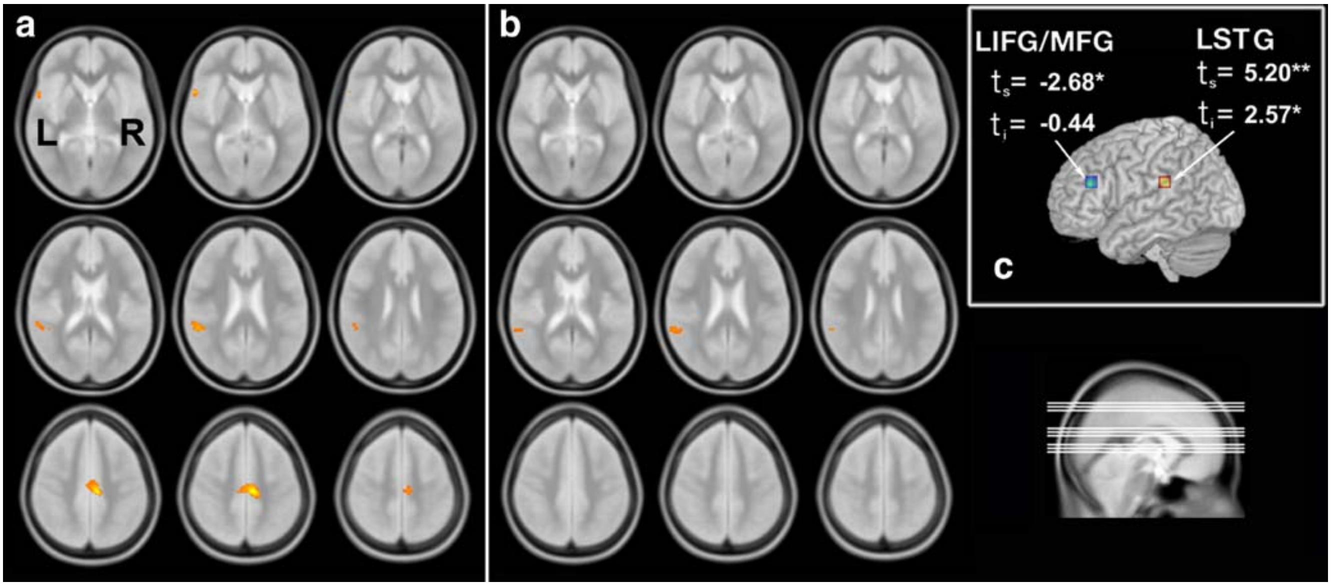

Fig. 2. Verb–Noun contrast for subject- and item-wise analyses. The results of the whole-brain, subject-wise (a) and item-wise (b) analyses for the Verbs–Nouns contrast. Warm colors represent voxels more active for verbs than nouns; cool colors represent voxels more active for nouns than verbs. The map-wise statistical threshold was set to $p<0.05$ for the subject-and item-wise maps [$t_1(12)=4.68$, $t_2(193)=3.9$, cluster=15 voxels] to demonstrate the spatial extent of the points of significant difference between verbs and nouns. The top rows of panels a and b depicts slices $Z=0$, $Z=4$, and $Z=8$, the middle rows depict slices $Z=18$, $Z=22$, and $Z=26$, and the bottom rows depict slices $Z=46$, $Z=50$, and $Z=54$. The results of the Verbs–Noun contrast, in the LMFG/IFG and LSTG anatomical ROIs are depicted in the bottom right-hand corner (c). t_i and t_s are the t-statistic for the Verbs–Nouns contrast by items and by subjects, respectively. Warm colors represent voxels more active for verbs than nouns; cool colors represent voxels more active for nouns than verbs. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

In addition to the focused ROI analyses, a whole-brain analysis of the Verb-Noun contrast across the subject population was also examined [$t_1(12)=4.68$, $p<0.05$, minimum of 15 contiguous voxels]. Four regions were more active for verbs than nouns: the anterior left STG (BA22) [7.8 cm^3, $t_{1\max}(12)=9.03$, $X=-54$, $Y=9$, $Z=0$]; the posterior left STG [19.5 cm^3, $t_{1\max}(12)=7.02$, $X=-54$, $Y=42$, $Z=18$]; the right posterior cingulate gyrus (BA31) [36.0 cm^3, $t_{1\max}(12)=13.22$, $X=9$, $Y=-27$, $Z=45$]; and the left precuneus (BA7) [4.5 cm^3, $t_{1\max}(12)=6.49$, $X=-9$, $Y=-48$, $Z=60$]. No regions were more activated for nouns than verbs at this threshold.

Item-wise analysis

Words–Non-words contrast

The whole-brain effects of reading Words versus Non-words were assessed across items [$t_2(297)=3.85$, $p<0.05$, minimum of 15 contiguous voxels], and the results are depicted in Fig. 1b. Overall, the item-wise analysis revealed a set of regions that were similar to that seen in the subject-wise analysis for the Words>Non-words comparison. Areas with the greatest peak and extent of activation included: LIFG [396.9 cm^3, $t_{2\max}(297)=11.65$, $X=-51$, $Y=21$, $Z=24$], LMTG [110.4 cm^3, $t_{2\max}(297)=10.79$, $X=-60$, $Y=-48$, $Z=6$], and left and right mSFG [132.0 cm^3, $t_{2\max}(297)=9.80$, $X=-3$, $Y=15$, $Z=51$].

Areas showing the strongest Non-word>Words effect included the right SFG/MFG [202.2 cm^3, $t_{2\max}(297)=9.00$, $X=27$, $Y=32$, $Z=51$], the right superior temporal and supramarginal gyri [300.3 cm^3, $t_{2\max}(297)=8.92$, $X=12$, $Y=22$, $Z=32$], and the posterior cingulate and precuneus [336.6 cm^3, $t_{2\max}(297)=8.05$, $X=9$, $Y=-42$, $Z=39$]. The left middle/superior frontal gyri peak of the Non-words–Words in the subject-wise analysis was also significant across items [65.7 cm^3, $t_{2\max}(297)=5.90$, $X=-24$, $Y=39$, $Z=48$] (see Table 2 for the list of activated regions).

Verbs–Nouns

In contrast to the findings of the Words–Non-words contrast, the item-wise analysis of the Verbs–Nouns validated only some of the findings of the subject-wise analysis. The LSTG ROI was significantly more active for verbs than nouns [$t_2(193)^4=2.57$, $p<0.01$] across items, while the LIFG/MFG ROI did not show a grammatical class effect [$t_2(193)=-0.44$, $p>0.10$] (see Fig. 2). Combining the subject- and item-wise analyses into $minF'$, the Verb–Noun contrast was significant in the LSTG ROI [$minF'(1,155)=5.31$, $p<0.05$], but not in the LIFG/MFG ROI [$minF'(1,198)=0.19$, $p>0.6$]

In the whole-brain item-wise analysis [$t_2(193)=3.9$, $p<0.05$, minimum of 15 contiguous voxels], only the STG region showed a grammatical class effect. This region was significantly more active for verbs than nouns [13.2 cm^3, $t_{2\max}(193)=5.17$, $p<0.05$]. The anterior LSTG, right posterior cingulate, or left precuneus regions identified in the subject-wise analysis did not show a verb specific effect in the item-wise analysis.

In summary, for the Words–Non-words contrast, the findings of the subject-wise and item-wise analyses were generally consistent with each other. However, the subject- and item-wise analyses for the Verbs–Nouns contrast were not entirely consistent. While several regions showed grammatical-class effects in the subject-

[4] The degrees of freedom here are those for the t-statistic corresponding to a partial correlation coefficient. $df=n-q-2$, where n is the number of items, and q is the number of variables held constant. Thus the degrees of freedom for the Verbs–Nouns contrast are $199-4-2=193$.

wise analysis, only one of these effects (in the left LSTG) was significant by item.

Discussion

While item analysis is a common inferential tool used in behavioral studies of cognition, it has not been previously applied to neuroimaging data. Given the similarity between the type of inferences made by behavioral and neuroimaging studies of cognition, what could cause such a disparity? This discrepancy may stem from two implicitly held beliefs: (i) item analysis is not sufficiently powered to detect task-relevant changes in BOLD signal; and (ii) even if such changes are detectable, they are not theoretically relevant to functional neuroimaging research. The present findings rebut both of these beliefs. Using item analysis, we show differences in processing between Words and Non-words, as well as Verbs and Nouns. These findings demonstrate that item analysis can detect task-relevant changes in BOLD signal. We also find that a grammatical-class effect (in the prefrontal cortex), which has been found in some but not all studies of grammatical class, is significant across subjects, but not across items. This finding illustrates how item analysis can make theoretically relevant distinctions between findings that generalize to stimulus populations, and those that do not.

Applying item analysis to neuroimaging research: feasibility and theoretical relevance

The results of the present study demonstrate the feasibility of item analysis in neuroimaging research. The sufficient power of item analysis in fMRI can best be seen from the results of the Word–Non-word contrasts. In the Word–Non-word contrast, almost all effects that were significant in the subject-wise analysis were also significant in item-wise analysis. In some regions, the results were more robust in the item-wise analysis. Overall, in the Words–Non-words contrast, more voxels reached the threshold of significance in the item than in the subject analysis. This may be due to several factors, including the greater variability of an effect across subjects than across items. Greater variability of the Non-word effect across subjects may result from participants using different strategies for completing the task. If the Non-word effect only occurred when a particular strategy was adopted, this would introduce variability across subjects but not items.

We were able to empirically demonstrate that item analysis is feasible in neuroimaging research. However, the feasibility and power of item analysis for any given study will depend, in part, on the specifics of the experimental design. Studies that require both subject- and item-wise analysis for theoretical reasons must be designed in such a way as to make item analysis feasible. Fortunately, many current experimental designs are already amenable to item analysis.

Designs that do not permit item analysis are those where the evoked response to any one item cannot be recovered. For example, the randomized presentation of stimuli for 1.5 s each, while the BOLD response is sampled at a TR of 3 s. In such a case it would not be possible to independently estimate the BOLD responses for each of the stimuli, a necessary step in item analysis.

Stimulus randomization is another aspect of experimental design that can affect item analysis. Randomizing stimuli for each participant renders the item analysis robust to failures of response modeling at the first level of analysis. Thus, the analysis of a fixed order of stimulus presentation requires accurate specification of the shape of the hemodynamic response function. Additionally, any non-linearities that arise in the transformation of neural activity into imaging signal must also be specified. Failures of these assumptions could create improper bias. In the extreme case, consider a situation in which the true hemodynamic response across subjects persisted over 30 s, although the modeled response was complete in 15 s. In this case, a positive and similar effect would be recorded for multiple stimuli, even if there were a true neural response to only the first.

The power of item analysis (like that of subject-wise analysis) is also dependent on the specifics of experimental design. An item analysis begins with the estimation of the magnitude of neural response to individual stimuli or trials within a BOLD fMRI experiment. Thus, experimental designs that enable the best estimation of the evoked hemodynamic response will increase power in an item-wise analysis (Liu, 2004).

In considering the complementary relationship of item-wise and subject-wise analysis a question arises regarding the relative power of each. It may, at first, appear that item analysis has greater power because there are generally more items than participants. As a result, item analysis has more degrees of freedom (and thus the standard error, computed by dividing the standard deviation by the square root of the number of items rather than subjects, may be considerably lower). However, because the sample of subjects is relatively small, the effect size is estimated less accurately for each item than for each subject. A poor estimate of the effect size for each item would tend to increase the standard deviation across items relative to the standard deviation across participants, which would in turn increase the standard error. With these factors in mind, it is not obvious whether item- or subject-wise analysis has more power. What is clear is that item- and subject-wise analyses can each be used to statistically support a distinct, and important set of theoretical conclusions.

The theoretical relevance of item analysis is best illustrated by the results of our ROI analysis, which examined differences between verbs and nouns. The ROIs used in the present study were based on previous findings of greater activity for verbs than nouns in the LIFG/MFG (Perani et al., 1999) and in the LSTG regions (Davis et al., 2004). In the subject-wise analysis, the present study replicated the finding of greater activity in the LSTG for verbs than nouns, but found greater activity for nouns than verbs in the LIFG/MFG region, conflicting with the findings of Perani et al. (1999). This discrepancy was clarified by the item analysis. While the LSTG effect was significant by item, both in the ROI and whole-brain analyses, the LIFG/MFG effect was not. This suggests that the apparent grammatical class effect in the LIFG/MFG is not consistent across nouns and verbs, but rather depends on the particular set of nouns and verbs sampled in any given study.

One possible concern is that the significance of the LIFG/MFG grammatical class effect across subjects, but not items, is an artifact of threshold. That is, the effect is barely significant across subjects and just misses the $p<0.05$ threshold by item. However, this is clearly not the case in the present dataset. In the LIFG/MFG ROI the effect of grammatical class across subjects is highly reliable [$t_1(12)=-3.79$, $p<0.01$], while the t-value for the same effect across items is below 1 [$t_{2\max}(193)=-0.44$, $p>0.10$].

The present findings support the assertion that the LSTG region plays an important role in verb processing. In contrast, the present data indicate that apparent grammatical class effects in the left prefrontal cortices do not generalize to item populations. The failure of the LIFG/MFG grammatical-class effect to reach significance in the item-wise analysis indicates that the effect varies a great deal across nouns and verbs. We hypothesize that the finding of a grammatical class effect in the subject-wise analysis of any given study depends in part on which set of nouns and verbs is sampled, and in part on how the choice of stimuli interacts with task demands (Price, 2000). Grammatical class effects in the LIFG/MFG region that have been attributed to noun/verb differences may reflect unintended variation in other linguistic variables (e.g. imageability) or processing demands (selection demands) that can be confounded with grammatical class.

The present findings shed light on one source of inconsistency in the literature and provides part of the explanation for why the left prefrontal grammatical class effect seems to variably present itself across studies (Bedny and Thompson-Schill, 2006; Davis et al., 2004; Shapiro et al., 2006; Tyler et al., 2004, 2001). It may be that the left prefrontal cortex responds more to some nouns than verbs, whereas the opposite is true of a different set of nouns and verbs. Thus the effect in any given study would depend on which nouns and verbs were sampled. Nonetheless, item variability is likely not the only reason for the inconsistency in the extant literature; factors such task demands also play an important role. A full discussion of implication of the present findings for hypotheses regarding the neural bases of grammatical class effects is beyond the scope of the present paper (Bedny and Thompson-Schill, 2006). Here we simply make the argument that item analysis is necessary (although not sufficient) to determine whether any given brain region truly exhibits category-specific effects.

Our findings illustrate why a subject-wise analysis is not sufficient for many of the inferences made in neuroimaging studies. If the hypothesis of a study hinges on the generalizability of the effect to the population of stimuli then the effect of treatment needs to be evaluated relative to its variability across items. Item analysis is particularly important for studies that test hypotheses about different types of stimuli (e.g. nouns and verbs) but do not exhaustively test their hypotheses on all possible stimuli in that category.

It is critical to add that item analysis does not replace, but rather augments subject-wise analysis. Item analysis in itself is insufficient for establishing generalizability to subject populations. If an effect is found to be reliable across items but not participants, this suggests that the population of items in question shows the effect of interest. On the other hand, we have no statistical bases for the inference that an effect generalizes to the population of subjects. In such cases there may be important inter-subject differences in the sensitivity to a manipulation. In order for an effect to be generalized to the populations of participants and items, the effect must be reliable across items and participants.

Limitations of item analysis

A critical limitation of the generalizability of inference supported by an item analysis is that it remains constrained by the space of possible stimuli that are sampled for an experiment. This is exactly analogous to the limitations of a random-effects, subject-wise analysis that would be unable to generalize to all humans if the pool of possible subjects included only college-age students living in Philadelphia. Further extension of the inference would require the assumption that Philadelphia college students are identical (with reference to the experimental manipulation of interest) to people in some other group. In the same way, item analysis might indicate that an effect generalizes to the population of "animals" from which experiment stimuli were sampled. However, conclusions regarding the neural representation of animals in general continue to be limited by what other stimuli *might have been presented with equal likelihood*. If only mammals were sampled from the population of all animals, conclusions cannot be extrapolated to other classes without the further assumption that mammals and other classes are represented in the same way. This limitation is common to all instances of statistical inference, where findings are extrapolated to a population based on a sample. However, this limitation does not preclude statistical analyses across items (or subjects). In contrast, it suggests that it is important to consider the target population of items when constructing a stimulus sample.

There are several categories of studies where item analysis is not critical, or simply not applicable. Item analysis is not critical (but may still be informative) when there is no need to generalize a finding to a stimulus population. This occurs when the hypothesis is not about a stimulus type, but rather, stimuli are used as a means for engaging a process of interest. For example, to test the hypothesis that a brain region is involved in maintaining information, a researcher selects a random sample of words and has participants remember them over a delay while undergoing fMRI. Across subjects, activity in area *X* increases when words are maintained relative to baseline. The researcher concludes that area *X* maintains information in working memory. Unfortunate though this experiment might be for other reasons, if this effect were not reliable across words, this would not invalidate the hypothesis that area *X* is involved in maintenance. It would however suggest that this region is involved in maintaining some types of words and not others.

Item analysis is not applicable if a researcher is able to sample exhaustively the stimulus population of interest. For example, the hypothesis that a brain region responds more to a particular vertical bar of light than a horizontal bar of light can be tested by exhaustively sampling a population consisting of two stimuli: a vertical bar and a horizontal bar. Item analysis is not applicable in such experiments.

Item analysis is also unnecessary for certain experimental designs. One such design makes use of counterbalanced lists such that items are assigned to different conditions for different participants. For example, if a researcher were interested in the difference between the neural processing of degraded and non-degraded acoustic stimuli he might divide the set of word stimuli into lists A and B. Different subjects might be presented either the A or B list in degraded form. In such a design the effect of condition is not influenced by variability across items, and item analysis is not required. Of course, such a design is not possible in studies of the intrinsic properties of the items themselves (e.g. nouns vs. verbs) (Raaijmakers, 2003).

An alternative way to obviate item analysis is to randomly sample different items from the stimulus population for each participant. For instance, in a neuroimaging study of nouns and verbs, this would entail sampling a different set of nouns and verbs for every participant. In this case the items variable is nested within the subjects variable. The subject-wise analysis alone is then appropriate for testing the significance of treatment effect (Clark, 1973).

Statistical approaches to item analysis

As noted above, there are several experimental designs that can allow for generalizability of effects to item populations without a separate item-wise analyses. When such designs are not possible, behavioral researchers typically report the subject- and item-wise statistics separately, and may also report a *minF′* statistic. Alone, a significant subject-wise statistic indicates that an effect is reliable across participants, and conversely a significant item-wise statistic indicates than an effect is reliable across stimuli. The *minF′* indicates that the effect is reliable across both participants and stimuli, and thus will always be less than or equal to the lowest of the subject- and item-wise statistics. Thus, it is possible (although rare) for both the item- and subject-wise test statistics to reach significance, while the quasi-*F* statistic does not (Clark, 1973). Some authors have argued that rather than reporting separate statistics for subjects and items, a quasi-*F* statistic (e.g. *minF′*) should be reported based upon the item- and subject-wise analysis (Clark, 1973; Raaijmakers et al., 1999). A limitation of the *minF′*, however, is that under some circumstances it may be overly conservative (Baayen, 2004; Forster and Dickinson, 1976; Wickens and Keppel, 1983). In the setting of neuroimaging data, *minF′* can be readily calculated for voxels or brain regions. As noted earlier, however, the calculation of a map-wise threshold for *minF′* is complicated by the non-stationary distribution of the statistic across voxels.

It seems, therefore, that a reasonable approach is to report subject- and item-wise significance in neuroimaging studies, both for ROI and whole-brain analysis. This will allow the reader to know whether an effect is reliable across participants, items, or both. Computing and reporting an item-wise test statistic when appropriate (in addition to a conventional subject-wise statistic) will substantially reduce the inflation of the Type I error that occurs as a result of failing to treat items as a random variable (Forster and Dickinson, 1976). Because the probability of a spuriously significant subject- and item-wise test statistic is somewhat larger than either of the subject or item-wise probability, when possible (i.e. for ROI analysis), a min*F′* statistic should also be reported.

Mixed-effects (or multilevel models) are a different statistical approach to establishing joint item- and subject-wise significance. Using multilevel modes does not require performing separate subject- and item-wise analyses as we described in the current paper. Instead, a single model is fit to the individual data points where the independent variables subject, item, and treatment are arranged in a hierarchical structure (Baayen, 2004). Multilevel models have been used in functional neuroimaging research to combine data across multiple runs or subjects in a hierarchical random effects analysis (Worsley et al., 2002). These models may offer an alternative approach to the problem of generalizing findings to item populations within neuroimaging data analysis. The goal of the present paper is not to advocate a particular approach to the problem, but rather to empirically demonstrate the feasibility and theoretical importance of item analysis in neuroimaging research and offer one effective and accessible method for generalizing findings to item populations.

The present findings demonstrate that item analysis is both feasible and theoretically relevant in functional neuroimaging research. Incorporating item analysis into fMRI studies will clarify discrepancies in the literature and weed out spurious effects. Item analysis is particularly critical for neuroimaging studies that test hypotheses about stimulus categories. If an effect is significant in a subject-wise analysis, but fails to reach significance in an item-wise analysis, there is no evidence that it generalizes beyond the stimulus sample of that study to the population of stimuli. Conducting an item-wise analysis may resolve conflicting results that arise from studies that attempt to address the same questions using different stimuli. In addition to weeding out spurious effects, item analysis can point to the presence of confounding variables and suggest directions for future research.

Acknowledgments

This research was supported by the National Institutes of Health (R01-MH67008, R01-MH60414) and the Searle Scholars Program. Additional funding was provided by a Career Award in the Biomedical Sciences from the Burroughs Wellcome Fund (GKA). We thank Robert Lorch and Saul Sternberg for their advice regarding the use of item analysis with continuous predictors, and the anonymous reviewers for their helpful comments on this manuscript. We would also like to acknowledge the impetus Herb Clark provided one of the authors (STS) for pursuing these issues.

References

Aguirre, G.K., Zarahn, E., D'Esposito, M., 1998. The variability of human. BOLD hemodynamic responses. NeuroImage 8 (4), 360–369.

Baayen, R.H. (2004). Statistics in Psycholinguistics: A critique of some current gold standards. In *Mental Lexicon Working Papers 1* (pp. 1–47). Edmonton.

Bedny, M., Thompson-Schill, S.L., 2006. Neuroanatomically separable effects of imageability and grammatical class during single-word comprehension. Brain Lang. 98 (2), 127–139.

Clark, H., 1973. The language-as-a-fixed-effect fallacy: critique of language statistics in psychological research. J. Verbal Learn. Verbal Behav. 12, 335–359.

Coleman, E.B., 1964. Generalizing to a language population. Psychol. Rep. 14, 219–226.

Davis, M.H., Meunier, F., Marslen-Wilson, W.D., 2004. Neural responses to morphological, syntactic, and semantic properties of single words: an fMRI study. Brain Lang. 89 (3), 439–449.

Forster, K.I., Dickinson, R.G., 1976. More on the language-as-fixed-effect fallacy: Monte Carlo estimates of error rates for F_1, F_2, F', and minF'. J. Verbal Learn. Verbal Behav. 15, 135–142.

Francis, W., Kucera, H., 1982. Frequency Analysis of English Usage. Houghton Mifflin Co, New York.

Friston, K.J., Holmes, A.P., Worsley, K.J., 1999. How many subjects constitute a study? NeuroImage 10 (1), 1–5.

Holmes, A.P., Friston, K.J., 1998. Generalizability, random effects, and population inference. NeuroImage 7, S754.

Kleinbaum, D.G., Kupper, L.L., Muller, K.E., Nizam, A., 1998. Applied Regression Analysis and Other Multivariable Methods, 3rd edition. Duxbury Press, Pacific Grove, CA.

Liu, T.T., 2004. Efficiency, power, and entropy in event-related fMRI with multiple trial types: Part II. Design of experiments. NeuroImage 21 (1), 401–413.

Nichols, T.E., Holmes, A.P., 2002. Nonparametric permutation tests for functional neuroimaging: a primer with examples. Hum. Brain Mapp. 15 (1), 1–25.

Perani, D., Cappa, S.F., Schnur, T., Tettamanti, M., Collina, S., Rosa, M.M., et al., 1999. The neural correlates of verb and noun processing: a PET study. Brain 122 (12), 2337–2344.

Price, C.J., 2000. The anatomy of language: contributions from functional neuroimaging. J. Anat. 197 (Pt 3), 335–359.

Raaijmakers, J.G., 2003. A further look at the "language-as-fixed-effect fallacy". Can. J. Exp. Psychol. 57 (3), 141–151.

Raaijmakers, J.G., Schrijnemakers, J.M.C., Gremmen, F., 1999. How to deal with "The Language-as-Fixed-Effect Fallacy": common misconceptions and alternative solutions. J. Mem. Lang. 41, 416–426.

Shapiro, K.A., Moo, L.R., Caramazza, A., 2006. Cortical signatures of noun and verb production. Proc. Natl. Acad. Sci. U. S. A. 103 (5), 1644–1649.

Tyler, L.K., Bright, P., Fletcher, P., Stamatakis, E., 2004. Neural processing of nouns and verbs: the role of inflectional morphology. Neuropsychologia 42 (4), 512–523.

Tyler, L.K., Russell, R., Fadili, J., Moss, H.E., 2001. The neural representation of nouns and verbs: PET studies. Brain 124 (8), 1619–1634.

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, 296–309.

Wise, R.J., Howard, D., Mummery, C.J., Fletcher, P., Leff, A., Buchel, C., et al., 2000. Noun imageability and the temporal lobes. Neuropsychologia 38 (7), 985–994.

Worsley, K.J., Friston, K.J., 1995. Analysis of fMRI time-series revisited—Again. NeuroImage 2 (3), 173–181.

Worsley, K.J., Liao, C.H., Aston, J., Petre, V., Duncan, G.H., Morales, F., et al., 2002. A general statistical analysis for fMRI data. NeuroImage 15 (1), 1–15.

Zarahn, E., Aguirre, G., D'Esposito, M., 1997a. A trial-based experimental design for fMRI. NeuroImage 6 (2), 122–138.

Zarahn, E., Aguirre, G.K., D'Esposito, M., 1997b. Empirical analyses of BOLD fMRI statistics: I. Spatially unsmoothed data collected under null-hypothesis conditions. NeuroImage 5 (3), 179–197.