nature neuroscience

PERSPECTIVE

Erroneous analyses of interactions in neuroscience: a problem of significance

Sander Nieuwenhuis[1,2], Birte U Forstmann[3] & Eric-Jan Wagenmakers[3]

In theory, a comparison of two experimental effects requires a statistical test on their difference. In practice, this comparison is often based on an incorrect procedure involving two separate tests in which researchers conclude that effects differ when one effect is significant ($P < 0.05$) but the other is not ($P > 0.05$). We reviewed 513 behavioral, systems and cognitive neuroscience articles in five top-ranking journals (*Science*, *Nature*, *Nature Neuroscience*, *Neuron* and *The Journal of Neuroscience*) and found that 78 used the correct procedure and 79 used the incorrect procedure. An additional analysis suggests that incorrect analyses of interactions are even more common in cellular and molecular neuroscience. We discuss scenarios in which the erroneous procedure is particularly beguiling.

"The percentage of neurons showing cue-related activity increased with training in the mutant mice ($P < 0.05$), but not in the control mice ($P > 0.05$)." "Animals receiving vehicle (control) infusions into the amygdala showed increased freezing to the conditioned stimulus compared with a control stimulus ($P < 0.01$); in animals receiving muscimol infusions into the amygdala, this difference was abolished ($F < 1$)."

These two fictive, but representative, statements illustrate a statistical error that is common in the neuroscience literature. The researchers who made these statements wanted to claim that one effect (for example, the training effect on neuronal activity in mutant mice) was larger or smaller than the other effect (the training effect in control mice). To support this claim, they needed to report a statistically significant interaction (between amount of training and type of mice), but instead they reported that one effect was statistically significant, whereas the other effect was not. Although superficially compelling, the latter type of statistical reasoning is erroneous because the difference between significant and not significant need not itself be statistically significant[1]. Consider an extreme scenario in which training-induced activity barely reaches significance in mutant mice (for example, $P = 0.049$) and barely fails to reach significance for control mice (for example, $P = 0.051$). Despite the fact that these two P values lie on opposite sides of 0.05, one cannot conclude that the training effect for mutant mice differs statistically from that for control mice. That is, as famously noted by Rosnow and Rosenthal[2], "surely, God loves the 0.06 nearly as much as the 0.05". Thus, when making a comparison between two effects, researchers should report the statistical significance of their difference rather than the difference between their significance levels.

Our impression was that this error of comparing significance levels is widespread in the neuroscience literature, but until now there were no aggregate data to support this impression. We therefore examined all of the behavioral, systems and cognitive neuroscience studies published in four prestigious journals (*Nature*, *Science*, *Nature Neuroscience* and *Neuron*) in 2009 and 2010 and in every fourth issue of the 2009 and 2010 volumes of *The Journal of Neuroscience*. In 157 of these 513 articles (31%), the authors describe at least one situation in which they might be tempted to make the error. In 50% of these cases (78 articles; **Table 1**), the authors used the correct approach: they reported a significant interaction. This may be followed by the report of the simple main effects (that is, separate analyses for the main effect of training in the mutant mice and control mice). In the other 50% of the cases (79 articles), the authors made at least one error of the type discussed here: they reported no interaction effect, but only the simple main effects, pointing out the qualitative difference between their significance values (for example, vehicle infusions were associated with a statistically significant increase in freezing behavior; muscimol infusions were not associated with a reliable increase in freezing behavior).

Are all these articles wrong about their main conclusions? We do not think so. First, we counted any paper containing at least one erroneous analysis of an interaction. For a given paper, the main conclusions may not depend on the erroneous analysis. Second, in roughly one third of the error cases, we were convinced that the critical, but missing, interaction effect would have been statistically significant (consistent with the researchers' claim), either because there was an enormous difference between the two effect sizes or because the reported methodological information allowed us to determine the approximate significance level. Nonetheless, in roughly two thirds of the error cases, the error may have had serious consequences. In all of these cases, the nonsignificant difference, although smaller in size, was in the same direction as the significant difference. In addition, the methodological information did not allow us to determine the significance level of the missing interaction test. We have no way of assessing the severity of these cases. Most of the errors may not have severe implications. In some cases, however, the error may contribute substantially to the article's main conclusions.

Because of our background expertise, our main analysis focused on behavioral, systems and cognitive neuroscience. However, it is

[1]Department of Psychology, Leiden University, Leiden, The Netherlands. [2]Leiden Institute for Brain and Cognition, Leiden, The Netherlands. [3]Cognitive Science Center Amsterdam, University of Amsterdam, Amsterdam, The Netherlands. Correspondence should be addressed to S.N. (s.nieuwenhuis@fsw.leidenuiv.nl).

Published online 26 August 2011; doi:10.1038/nn.2886

Table 1 Outcome of the main literature analysis

	Nature	*Science*	*Nature Neuroscience*	*Neuron*	*Journal of Neuroscience*	Summed
Total reviewed	34	45	117	106	211	513
Correct count	3	9	17	13	36	78
Error count	7	11	16	15	30	79

For this analysis, we included every article of which the abstract referred to behavior, cognitive function or brain imaging.

likely that the incorrect analysis of interactions is not just limited to these disciplines. To confirm this intuition, we reviewed an additional 120 cellular and molecular neuroscience articles published in *Nature Neuroscience* in 2009 and 2010 (the first five Articles in each issue). We did not find a single study that used the correct statistical procedure to compare effect sizes. In contrast, we found at least 25 studies that used the erroneous procedure and explicitly or implicitly compared significance levels. In general, data collected in these cellular and molecular neuroscience studies were analyzed mostly with *t* tests (possibly corrected for multiple comparisons or unequal variances) and occasionally with one-way ANOVAs, even when the experimental design was multifactorial and required a more sophisticated statistical analysis.

Our literature analyses showed that the error occurs in many different situations: when researchers compared the effects of a pharmacological agent versus placebo; patients versus controls; one versus another task condition, brain area or time point; genetically modified versus wild-type animals; younger versus older participants; etc. We describe three general types of situations in which the error occurs and illustrate each with a prototypical (fictive) example.

First, most of the errors that we encountered in our analysis occurred when comparing effect sizes in an experimental group/condition and a control group/condition (for example, sham-TMS, vehicle infusion, placebo pill, wild-type mice). The two examples at the start of this article belong to this type. Another example would be "Optogenetic photoinhibition of the locus coeruleus decreased the amplitude of the target-evoked P3 potential in virally transduced animals ($P = 0.012$), but not in control animals ($P = 0.3$)" (**Fig. 1a**). The researchers contrast the significance levels of the two effect sizes instead of reporting the significance level of a direct statistical comparison between the effect sizes. The claim that the effect of the optogenetic manipulation on P3 amplitude is larger in the virally transduced animals than in the control animals requires a significant interaction between the manipulation (photoinhibition versus baseline) and group (virally transduced versus control mice). Because the plotted results reflect the group averages of individual averages that we generated ourselves (for ten mice in each group), we know that the interaction in this example is not significant ($P > 0.05$). Thus, the claim that the researchers intend to make is not statistically valid.

Second, comparing effect sizes during a pre-test and a post-test can be seen as a special case of the situation described above, in which the pre-test (before the experimental manipulation) is the control condition and the post-test (after the manipulation) is the experimental condition. An example is "Acute fluoxetine treatment increased social approach behavior (as indexed by sniff time) in our mouse model of depression ($P < 0.01$)" (**Fig. 1b**). Errors of this type are less common and often less explicit. In this example, the researchers contrast only the post-test scores of the two groups, on the tacit assumption that they need not take into account the corresponding pre-test scores, perhaps because the pre-test scores do not reliably differ between groups. Thus, the researchers implicitly base their claim on the difference between the significant post-test difference and the nonsignificant pre-test difference, when instead they should have directly compared the effect sizes, for example, by examining the time × group interaction in a repeated-measures analysis of variance.

The third type of error occurs when comparing several brain areas and claiming that a particular effect (property) is specific for one of these brain areas. In this type of situation, researchers do not compare a designated region of interest with a control area, but instead compare a number of brain areas with more or less equal 'a priori status'. An example would be "Escape latency in the Morris water maze was affected by lesions of the entorhinal cortex ($P < 0.05$), but was spared by lesions of the perirhinal and postrhinal cortices (both P values > 0.1), pointing to a specific role for the enthorinal cortex in spatial memory" (**Fig. 1c**). Although this type of conclusion is less salient than the explicit claim of a difference between brain areas, the specificity claim nevertheless requires a direct statistical comparison. That is, at the very least, spatial memory should be more impaired in animals with enthorinal lesions than in animals with lesions in other areas. Thus, the specificity claim requires that the researchers report a significant time × lesion type interaction, followed by significant pair-wise comparisons between the specific brain area and the other brain areas.

These three examples involve errors that we would classify as being potentially serious, as the nonsignificant effect is in the same direction as the significant effect (except for the perirhinal cortex), and because the information in **Figure 1a–c** is not sufficient to estimate the significance of the missing interaction test. The reason is that each of these three graphs contains repeated measurements (for example, before and after treatment). In the case of repeated measurements on the same group(s) of subjects, the standard-error bars do not give the information needed to assess the significance of the differences between the repeated measurements, as they are not sensitive to the correlations between these measurements[3]. Standard-error bars can

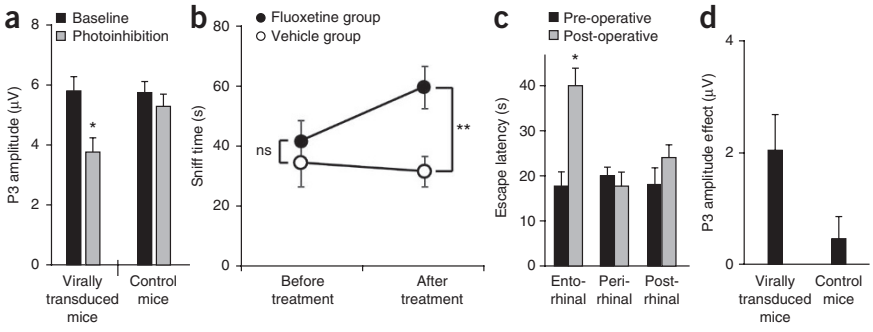

Figure 1 Graphs illustrating the various types of situations in which the error of comparing significance levels occurs. (**a**) Comparing effect sizes in an experimental group/condition and a control group/condition. (**b**) Comparing effect sizes during a pre-test and a post-test. (**c**) Comparing several brain areas and claiming that a particular effect (property) is specific for one of these brain areas. (**d**) Data presented in **a**, after taking the difference of the two repeated-measures (photoinhibition and baseline). Error bars indicate s.e.m.; ns, nonsignificant ($P > 0.05$), $*P < 0.05$, $**P < 0.01$.

only be used to assess the significance of between-group differences. Thus, the reader can only judge whether an interaction would be significant if the means and standard errors reflect the difference between repeated measurements (as in **Fig. 1d**, which is based on the same data as **Fig. 1a**). Thus, unlike **Figure 1a**, we can use **Figure 1d** to estimate the significance of the interaction by comparing the size of the gap (or in other situations the degree of overlap) between the two error bars[4].

We have discussed errors that occur when researchers compare experimental effects. However, in our analysis, we found that the error also occurs when researchers compare correlations. A fictive example would be "Hippocampal firing synchrony correlated with memory performance in the placebo condition ($r = 0.43$, $P = 0.01$), but not in the drug condition ($r = 0.19$, $P = 0.21$)". When making a comparison between two correlations, researchers should directly contrast the two correlations using an appropriate statistical method.

As noted by others[5,6], the error of comparing significance levels is especially common in the neuroimaging literature, in which results are typically presented in color-coded statistical maps indicating the significance level of a particular contrast for each (visible) voxel. A visual comparison between maps for two groups might tempt the researcher to state, for example, that "the hippocampus was significantly activated in younger adults, but not in older adults". However, the implied claim is that the hippocampus is activated more strongly in younger adults than in older adults, and such a claim requires a direct statistical comparison of the effects. Similarly, claims about differences in activation across brain regions must be supported by a significant interaction between brain region and the factor underlying the contrast of interest. For example, "Compared to non-moral dilemmas, the moral dilemmas activated only the insular cortex, suggesting that this area is uniquely involved in making moral judgments". Identification of the significant response in the insular cortex does not imply that this region is uniquely or more strongly involved in making moral judgments than other regions. It merely implies that, although the null hypothesis has been rejected in this region, it has not been rejected elsewhere.

It is interesting that this statistical error occurs so often, even in journals of the highest standard. Space constraints and the need for simplicity may be the reasons why the error occurs in journals such as *Nature* and *Science*. Reporting interactions in an analysis of variance design may seem overly complex when one is writing for a general readership. Perhaps, in some cases, researchers choose to report the difference between significance levels because the corresponding interaction effect is not significant. Peer reviewers should help authors avoid such mistakes. The statistical error may also be a manifestation of the cliff effect[7], the phenomenon that many people's confidence in a result drops abruptly when a P value increases just beyond the 0.05 level. Indeed, people are generally tempted to attribute too much meaning to the difference between significant and not significant. For this reason, the use of confidence intervals may help prevent researchers from making this statistical error. Whatever the reasons for the error, its ubiquity and potential effect suggest that researchers and reviewers should be more aware that the difference between significant and not significant is not itself necessarily significant.

AUTHOR CONTRIBUTIONS

S.N. conceived the project and made the figure. S.N., B.U.F. and E.-J.W. conducted the literature analyses and wrote the paper.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

Published online at http://www.nature.com/natureneuroscience/.
Reprints and permissions information is available online at http://www.nature.com/reprints/index.html.

1. Gelman, A. & Stern, H. The difference between "significant" and "not significant" is not itself statistically significant. *Am. Stat.* **60**, 328–331 (2006).
2. Rosnow, R.J. & Rosenthal, R. Statistical procedures and the justification of knowledge in psychological science. *Am. Psychol.* **44**, 1276–1284 (1989).
3. Loftus, G.R. & Masson, M.E.J. Using confidence intervals in within-subject designs. *Psychon. Bull. Rev.* **1**, 476–490 (1994).
4. Cumming, G., Fidler, F. & Vaux, D.L. Error bars in experimental biology. *J. Cell Biol.* **177**, 7–11 (2007).
5. Henson, R. What can functional neuroimaging tell the experimental psychologist? *Q. J. Exp. Psychol. A* **58**, 193–233 (2005).
6. Poldrack, R.A. *et al.* Guidelines for reporting an fMRI study. *Neuroimage* **40**, 409–414 (2008).
7. Rosenthal, R. & Gaito, J. The interpretation of levels of significance by psychological researchers. *J. Psychol.* **55**, 33–38 (1963).