

1 Re-analysing the data from Moffatt et al. (2020): A textbook illustration of the absence of
2 evidence fallacy

3 Ladislav Nalborczyk¹

4 ¹ Univ. Grenoble Alpes, CNRS, Grenoble INP, GIPSA-lab, 38000 Grenoble, France

5 Author Note

6 Correspondence concerning this article should be addressed to Ladislav Nalborczyk,
7 GIPSA-lab, CNRS, Univ. Grenoble Alpes, 11 Rue des Mathématiques, 38400
8 Saint-Martin-d'Hères, France. E-mail: ladislav.nalborczyk@gipsa-lab.fr

Abstract

Moffatt et al. (2020) reported the results of an experiment ($N = 26$ in the final sample) comparing the facial (surface) electromyographic correlates of mental rumination and distraction, following an experimentally induced stressor. Based on the absence of significant difference in the perioral muscular activity between the rumination and distraction conditions, Moffatt et al. (2020) concluded that *self-reported* inner experience was unrelated to peripheral muscular activity as assessed using surface electromyography. We suggest this conclusion is hasty and based on waggly evidence. Indeed, concluding on the absence of an effect based on a low-powered non-significant p-value is strongly problematic/uninformative. Moreover, the relation between self-reports and physiological measures was not *directly* assessed, but only indirectly inferred from differences (or absence thereof) in group means. Given the ample inter-individual variability in these measures (as suggested by our reanalysis), we think inferring the individual-level relation between self-reports and physiological measures from group means is inappropriate. Given these limitations, we conclude that it is unclear whether the target article adds to the current/extent knowledge and we suggest ways forward, both from a theoretical and from a methodological perspective. Complete source code, reproducible analyses, and figures are available at https://github.com/lnalborczyk/inner_experience_EMG.

Keywords: NHST, Bayesian, fallacy, reanalysis, inner speech, rumination, electromyography

29 Re-analysing the data from Moffatt et al. (2020): A textbook illustration of the absence of
30 evidence fallacy

31 Wordcount (excluding abstract, references, tables, and figures): 1727

Introduction

The activity of silently talking to oneself or “inner speech” is a foundational ability... despite its multiple adaptive functions in everyday life, inner speech can go awry and leads to sustained negative... These inner speech “dysfunctions” (for reviews, see Alderson-Day & Fernyhough, 2015; Loevenbruck et al., 2018; Perrone-Bertolotti et al., 2014)...

Given the predominantly verbal nature of rumination [], we previously proposed to study rumination as other forms of inner speech have been studied in the past, namely using surface electromyography and motor interference protocols (e.g., Nalborczyk et al., 2017; Nalborczyk, 2019; Nalborczyk, Perrone-Bertolotti, et al., 2020; Nalborczyk, Grandchamp, et al., 2020)...

We have previously shown that... However, it was unclear... therefore, the extension of our study by Moffatt et al. (2020), which consisted by including a distraction control group, is more than welcome... However...

The main conclusion from Moffatt et al. (2020) is that inner experience between induced rumination and distraction differs “without a change in electromyographic correlates of inner speech”. In other words, their conclusion is that inner experience is unrelated (or loosely related) to the electromyographic correlates of inner speech, which are thought to be represented mostly by the EMG amplitude recorded over the OOI and OOS muscles. However, for this in-sample observation to be of interest in an out-of-sample context (i.e., to be informative of other non-observed individuals, or said otherwise, to bring information about the population), this absence of difference has to be based on sufficiently powered sample size (given the target effect size) as well as on reliable measures. Moreover, a simple visual exploration of the data reveals important variability between individuals in the main effect of interest. That is, some participants had higher perioral (OOS and OOI) muscular activity in the rumination condition than in the distraction

condition, and some other participants showed the reverse pattern. This suggests unexplored variation in the determinants of this effects (e.g., the content of the inner experience). Indeed, the relation between the inner experience and the physiological correlates of inner speech production was only inferred from group means. However, given the previous point, this appears highly problematic. We explore each of these limitations and suggests ways forward in the following section.

Exploring the data

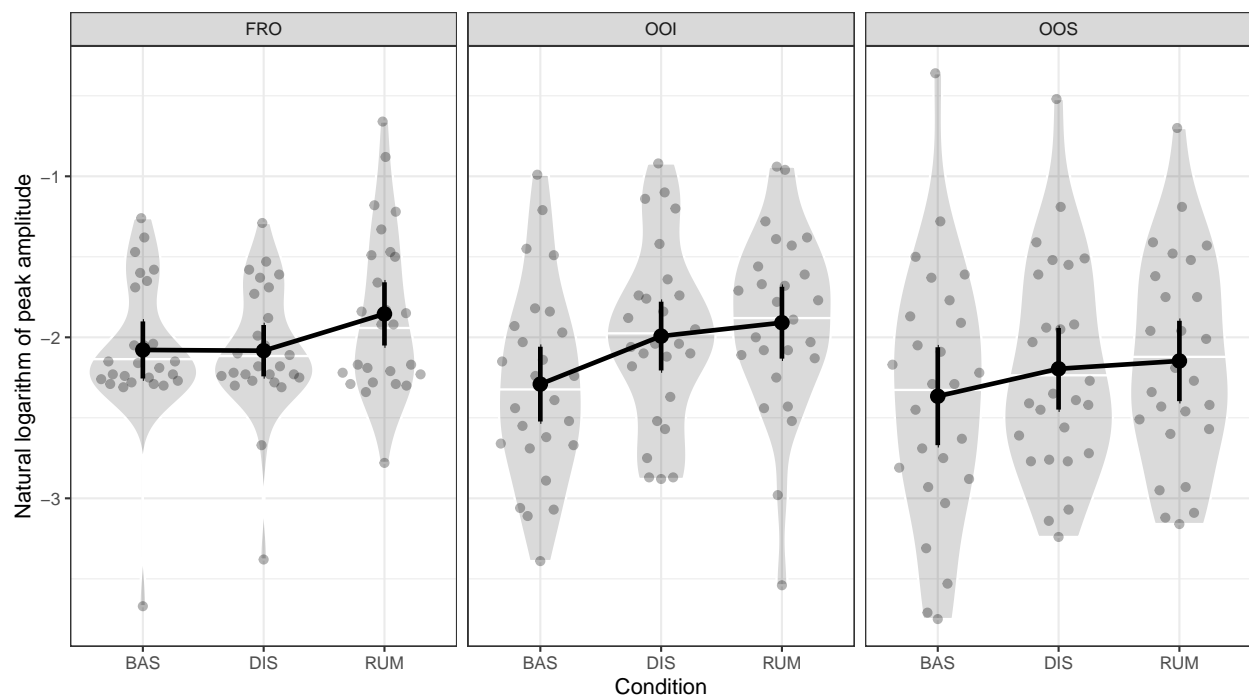


Figure 1. Average log-EMG amplitude by muscle and condition. The black dots and intervals represent the by-group average and 95% confidence interval ($N = 26$). The horizontal white line in the violin plot represents the median. The grey dots represent the individual-level average natural logarithm of the EMG amplitude by muscle and condition.

...

Posterior distribution of the difference between the distraction and rumination

conditions...

Concluding on the null from low-powered studies is uninformative

There is an infamous tradition of running uninformative null-hypothesis significance tests in Psychology (e.g., Meehl, 1997, 1978, 1990a, 1990b, 1967). By “uninformative”, we mean that some null-hypothesis significance tests are often *not* diagnostic with regards to the substantive question of interest...

As highlighted by many authors (e.g., Pollard & Richardson, 1987; Rouder et al., 2016), concluding on an absence of difference based on not obtaining evidence for the difference is the continuous extension of the logical fallacy of the... The argument from ignorance, such as “Science has found no proof of intelligent life nearby us in space, therefore intelligent life does not exist nearby us in space.”... the absence of evidence fallacy or fallacy of acceptance...

This problem is tackled in modern usages of null-hypothesis significance test by ensuring that the test has good *severity* (e.g., Mayo & Spanos, 2006; Mayo, 2018). In general terms, we have evidence for a claim to the extent that it survives a stringent scrutiny, that is if it survives *severe tests*. In other words, some claim (e.g., $\theta = 0$) is said to be *severely tested* if it had great chances of being falsified, was the claim false. More formally, we can define $\text{SEV}(T, x_0, H)$, the severity with which claim H passes test T with outcome x_0 , and $\text{SEV}(\mu > \mu_1) = \Pr(d(X) \leq d(x_0); \mu = \mu_1)$ (Mayo, 2018; Mayo & Spanos, 2006)...To put it simply... <https://www.analytics-toolkit.com/glossary/severity/>...

Anticipating the critics on the power of their study (a critic that was probably raised during peer review), Moffatt et al. (2020) report the results of a (possibly ran a posteriori) power analysis using the effect size reported in Nalborczyk et al. (2017) of $d = 0.72$, which is highly optimistic estimate of the substantive effect of interest in the target article (i.e.,

the difference in EMG amplitude between the rumination and distraction conditions) as this effects represents the standardised mean difference *between a rest period and a rumination one* (Nalborczyk et al., 2017)...

```
# How many participants do we need for a target statistical power of 0.8?
library(pwr)
pwr.t.test(
  d = 0.72, sig.level = 0.05, power = 0.8,
  type = "one.sample", alternative = "two.sided"
)
```

```
##
##      One-sample t test power calculation
##
##              n = 17.16004
##              d = 0.72
##      sig.level = 0.05
##              power = 0.8
##      alternative = two.sided
```

We suggest the (a priori) power of the study ran by Moffatt et al. (2020) was much lower than suggested by the authors. Indeed, we may speculate that the effect (i.e., the standardised mean difference in EMG amplitude) between the rumination and distraction conditions may be much weaker than the effect (i.e., the standardised mean difference in EMG amplitude) between the rumination and the rest conditions. If we assume that the former is half the size of the latter (which seems reasonable given the high inter-individual variability in such effects, cf. the next section but also Nalborczyk, Grandchamp, et al., 2020), therefore the a priori power of the main statistical test from Moffatt et al. (2020) was around 0.44, meaning that they had less than 1 chance out of 2 to find a significant

effect, given that the effect in the population was actually 0.36. Because this is less than the chance of obtaining a head in a coin flip, we feel these resources may have been better invested.

```
# A priori power for n = 26 (per condition) and d = 0.36
pwr.t.test(
  n = 26, d = 0.72 / 2, sig.level = 0.05,
  type = "one.sample", alternative = "two.sided"
)
```

```
##
##      One-sample t test power calculation
##
##              n = 26
##              d = 0.36
##      sig.level = 0.05
##      power = 0.4228455
##      alternative = two.sided
```

Anticipating the legitimate critique that the absence of a significant difference is not *necessarily* “significant” evidence for the absence of an effect, Moffatt et al. (2020) reported the following Bayes factor (BF) analysis:

“[...] therefore it is possible that the sample size of the present study lacked sufficient power to detect the effect of rumination on muscle activity. In order to test this, a Bayesian paired samples t-test was conducted for the peak log values of muscle activity between the rumination and distraction conditions. This revealed strong evidence in favour of the alternative hypothesis for the FRO muscle ($B_{10} = 18.79$), and moderate evidence in favour of the null

hypothesis for the OOS ($B_{10} = 0.232$) and OOI ($B_{10} = 0.278$) muscles,
according to current guidelines for interpreting Bayes factors [43].”

While we appreciate the effort, the current approach poses new problems. First,
contrary to what the authors suggest, computing a BF (i.e., comparing two models) does
not solve *at all* the problem of low power...

We first fitted a multivariate Bayesian regression model with varying-intercepts (by
participant) and weakly informative priors on these data using the **brms** package (Bürkner,
2017). From there, we i) generated new datasets from the posterior predictive distribution
and ii) we computed the BF in favour of the alternative hypothesis (BF_{10}) using the
BayesFactor package (Morey & Rouder, 2018). We used a “medium” prior (i.e., a scale of
1) on the scale of the Cauchy prior for the alternative hypothesis. We repeated this
procedure for varying sample sizes from 20 to 200 participants (by increments of 10
participants) with 1000 simulations (i.e., 1000 simulated datasets) for each sample size.

As shown in Figure 2, the natural logarithm of the BF in favour of the alternative
hypothesis is growing proportionally with the sample size. More precisely, whereas low
sample sizes (i.e., below 80) support the null hypothesis, adequately-powered sample sizes
support the alternative hypothesis for all three facial muscles for sample sizes larger than 80
participants. For instance, the average BF_{10} computed for the OOI muscle with a sample
size of 160 participants is of $\exp(2.18) \approx 8.85$, indicating that these data are approximately
8.85 times more likely under the alternative hypothesis than under the null hypothesis.¹

We should keep in mind some limitations of this analysis, which uses simulated
datasets from the posterior predictive distribution estimated on the data collected by
Moffatt et al. (2020). This analysis is the loose Bayesian analogue of the frequentist
post-hoc power analysis, which has been much criticised (e.g., Lakens, 2014). Most

¹ Alternatively, the BF can be interpreted as an updating factor, from prior odds to posterior odds.

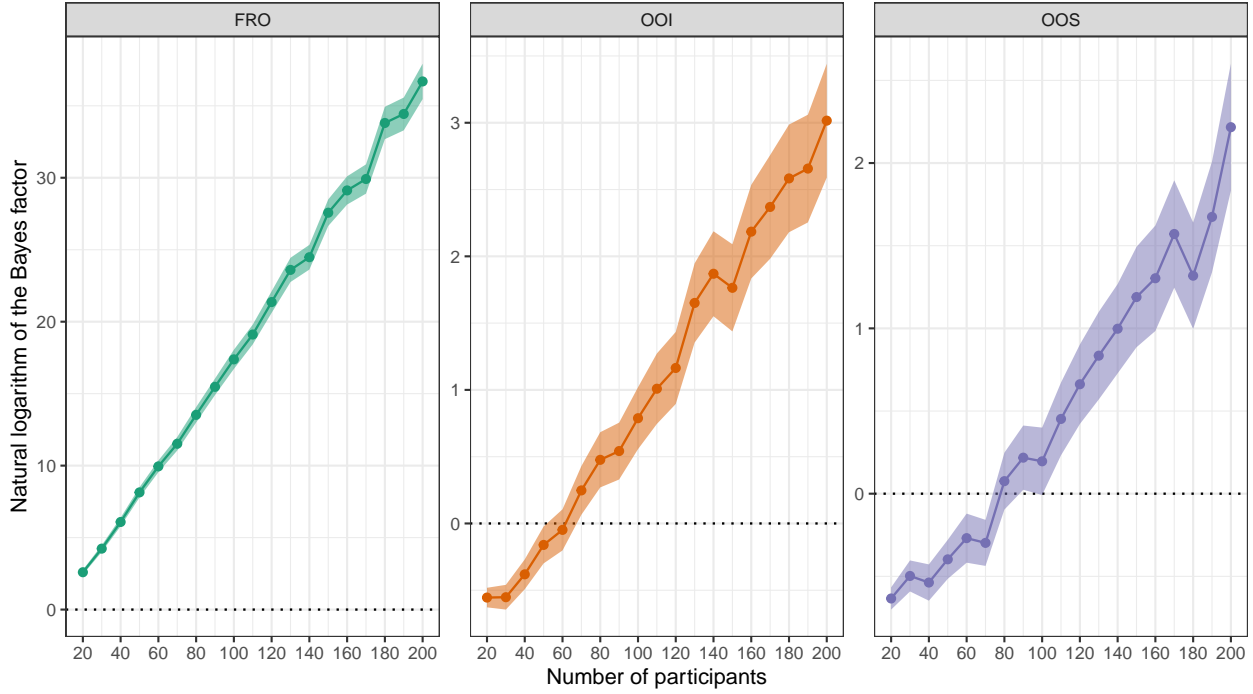


Figure 2. Average natural logarithm of the Bayes factor in favour of the alternative hypothesis, along with its standard error, computed over 1000 datasets of increasing size simulated from the posterior predictive distribution of the varying-intercept multivariate Bayesian regression model. A log-BF belows 0 represents evidence for the null hypothesis (relative to the alternative) and a log-BF above 0 represents evidence for the alternative hypothesis (relative to the null).

importantly, an assumption of the present analysis is that the data from Moffatt et al. (2020) is our best source of information regarding the main effect of interest (in addition to the prior we specified earlier). However, the present analysis also differs from the frequentist post-hoc power analysis on several grounds. First, with the present analysis, we do not aim to assess the ability of our statistical test to pass some dichotomic threshold (e.g., accept/reject). Instead, we aim to assess how the BF_{10} (i.e., the evidence for the alternative hypothesis, relative to the null hypothesis) behaves with varying sample sizes. Second, the present analysis relies on the posterior predictive distribution of the model fitted on the data from Moffatt et al. (2020), which naturally incorporates uncertainty

about the effect of interest. By simulating datasets of varying sample sizes from the posterior predictive distribution (and by relying on a large number of simulations), uncertainty about the effect size is naturally incorporated into the simulation.

Manipulating rumination within-subject

In Nalborczyk, Banjac, et al. (2020), we manipulated the modality of rumination (whether it is verbal or non-verbal) in a between-subject manner to avoid order effects... In contrast to this approach, Moffatt et al. (2020) asked participants to ruminate and then distract themselves (or reciprocally), after an induced stressor (an induced failure)...

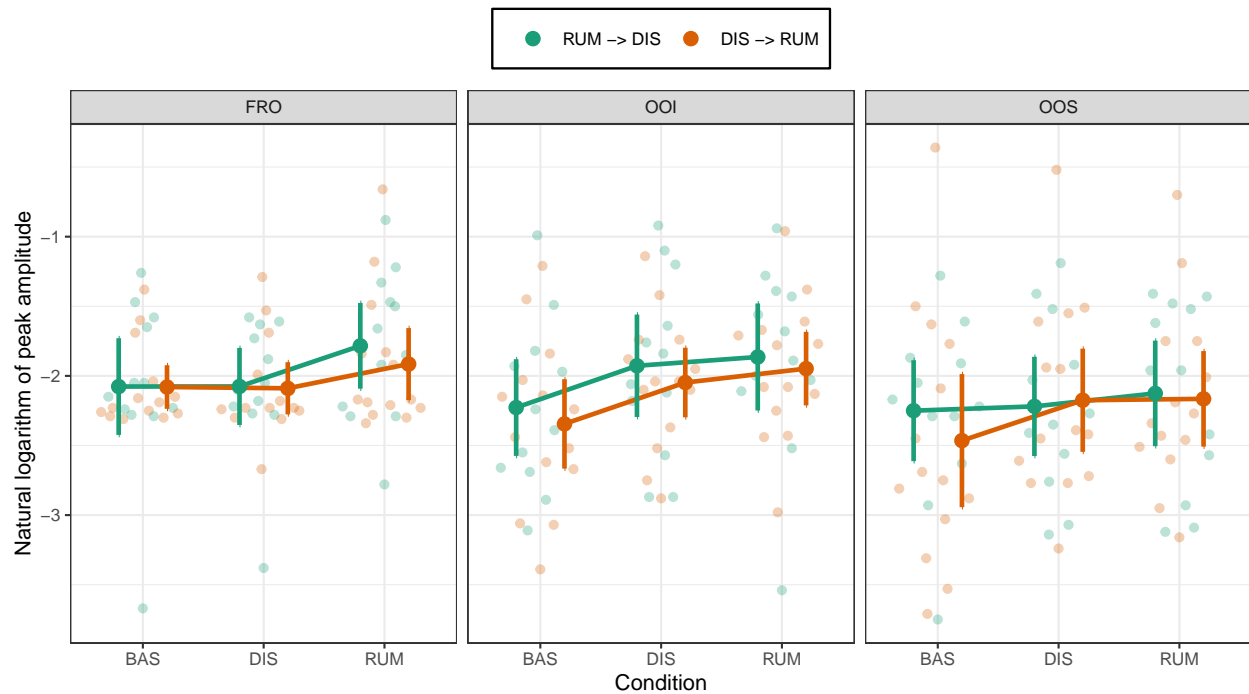


Figure 3. Average log-EMG amplitude by muscle and condition. The black dots and intervals represent the by-group average and 95% confidence interval ($N = 26$). The horizontal white line in the violin plot represents the median. The grey dots represent the individual-level average natural logarithm of the EMG amplitude by muscle and condition.

About the order effects, Moffatt et al. (2020) say:

“Unless otherwise reported, the inclusion of order in which the conditions were completed as a between-subjects variable as part of a mixed-design ANOVA produced no significant main effects or interactions involving order.”

Unfortunately, the same line of reasoning applies for testing the effect of the order, which is even less powered than the test of the main effect of interest, rendering it practically uninformative...

Does everyone?

Haaf and Rouder (2017)...

Huge inter-individual variability... which leads to the next point, what is the relation between self-reports and EMG?

Relation between self-report and EMG correlates

...

Discussion and conclusions

...

Supplementary materials

Reproducible code and figures are available at
https://github.com/linalborczyk/inner_experience_EMG.

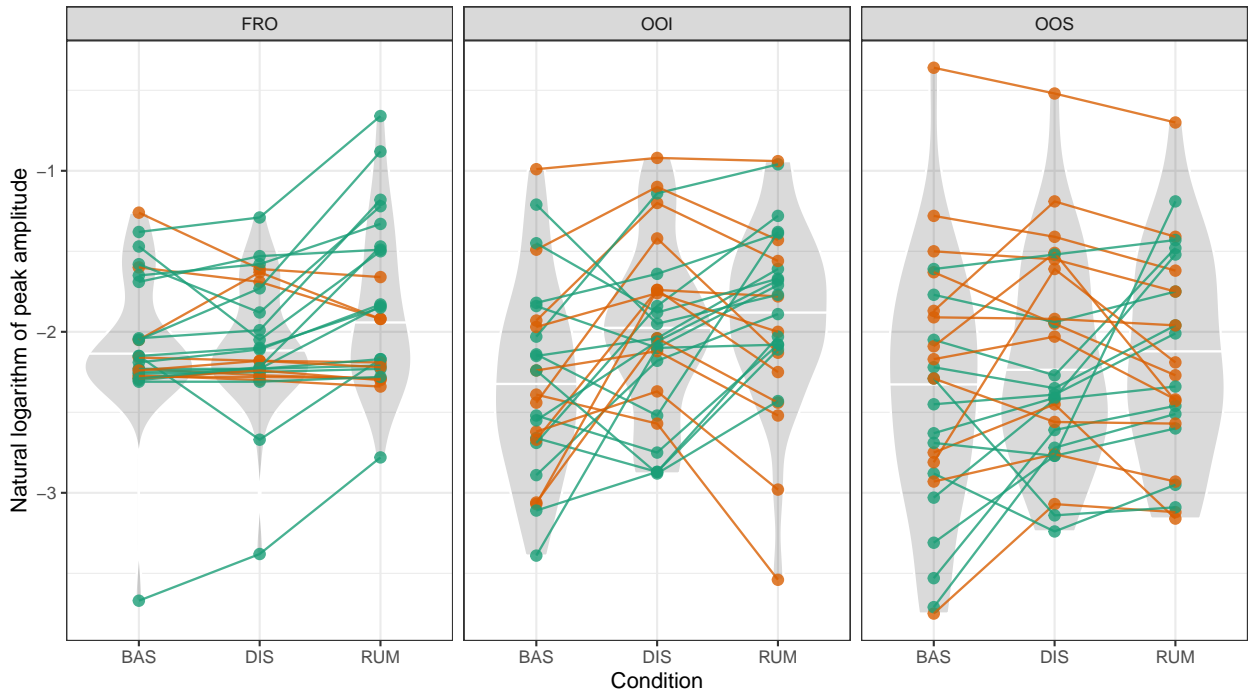


Figure 4. Inter-individual variability in the main effect of interest (i.e., the difference between the rumination and distraction conditions). Green dots and lines represent the average natural logarithm of the EMG amplitude of participants that showed a higher EMG amplitude in the rumination condition than in the distraction condition, whereas orange dots and lines represent the average natural logarithm of the EMG amplitude of participants that showed a higher EMG amplitude in the distraction condition than in the rumination condition.

Acknowledgements

...

References

Alderson-Day, B., & Fernyhough, C. (2015). Inner speech: Development, cognitive functions, phenomenology, and neurobiology. *Psychological Bulletin*, 141(5), 931–965. <https://doi.org/10.1037/bul0000021>

Aust, F., & Barth, M. (2017). *papaja: Create APA manuscripts with R Markdown*.

<https://github.com/crsh/papaja>

Bürkner, P.-C. (2017). brms: An R package for bayesian multilevel models using Stan.

Journal of Statistical Software, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>

Haaf, J. M., & Rouder, J. N. (2017). Developing constraint in Bayesian mixed models.

Psychological Methods, 22(4), 779–798. <https://doi.org/10.1037/met0000156>

Lakens, D. (2014). The 20% Statistician: Observed power, and what to do if your editor

asks for post-hoc power analyses. In *The 20% Statistician*.

Loevenbruck, H., Grandchamp, R., Rapin, L., Nalborczyk, L., Dohen, M., Perrier, P.,

Baciu, M., & Perrone-Bertolotti, M. (2018). A cognitive neuroscience view of inner

language: To predict and to hear, see, feel. In P. Langland-Hassan & A. Vicente (Eds.),

Inner speech: New voices (p. 37). Oxford University Press.

Marwick, B. (2019). *Wordcountaddin: Word counts and readability statistics in r*

markdown documents.

Mayo, D. G. (2018). *Statistical Inference as Severe Testing: How to Get Beyond the*

Statistics Wars. Cambridge University Press. <https://doi.org/10.1017/9781107286184>

Mayo, D. G., & Spanos, A. (2006). Severe Testing as a Basic Concept in a NeymanPearson

Philosophy of Induction. *The British Journal for the Philosophy of Science*, 57(2),

323–357. <https://doi.org/10.1093/bjps/axl003>

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the

slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46(4),

806–834. <https://doi.org/10.1037/0022-006X.46.4.806>

Meehl, P. E. (1990a). Why Summaries of Research on Psychological Theories are Often Uninterpretable. *Psychological Reports*. <https://doi.org/10.2466/pr0.1990.66.1.195>

Meehl, P. E. (1997). The problem is epistemology, not statistics: Replace significance tests by confidence intervals and quantify accuracy of risky numerical predictions. *What If There Were No Significance Tests?*, 393–425.

Meehl, P. E. (1990b). Appraising and Amending Theories: The Strategy of Lakatosian Defense and Two Principles that Warrant It. *Psychological Inquiry*, 1(2), 108–141. https://doi.org/10.1207/s15327965pli0102_1

Meehl, P. E. (1967). Theory-testing in Psychology and Physics: A methodological paradox. *Philosophy of Science*, 34(2), 103–115. <https://doi.org/10.1086/288135>

Moffatt, J., Mitrenga, K. J., Alderson-Day, B., Moseley, P., & Fernyhough, C. (2020). Inner experience differs in rumination and distraction without a change in electromyographical correlates of inner speech. *PLOS ONE*, 15(9), e0238920. <https://doi.org/10.1371/journal.pone.0238920>

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of bayes factors for common designs*. <https://CRAN.R-project.org/package=BayesFactor>

Müller, K. (2017). *Here: A simpler way to find your files*. <https://CRAN.R-project.org/package=here>

Nalborczyk, L. (2019). *Understanding rumination as a form of inner speech: Probing the role of motor processes* [PhD Thesis]. Univ. Grenoble Alpes & Ghent University.

Nalborczyk, L., Banjac, S., Celine, B., Grandchamp, R., Koster, E. H. W., Marcela, P.-B., & Loevenbruck, H. (2020). *Dissociating facial electromyographic correlates of visual and verbal induced rumination*. <https://doi.org/10.31234/osf.io/vfjn2>

- Nalborczyk, L., Grandchamp, R., Koster, E. H. W., Perrone-Bertolotti, M., & Løevenbruck, H. (2020). Can we decode phonetic features in inner speech using surface electromyography? *PLOS ONE*, 15(5), e0233282.
<https://doi.org/10.1371/journal.pone.0233282>
- Nalborczyk, L., Perrone-Bertolotti, M., Baeyens, C., Grandchamp, R., Polosan, M., Spinelli, E., Koster, E. H. W., & Løevenbruck, H. (2017). Orofacial electromyographic correlates of induced verbal rumination. *Biological Psychology*, 127, 53–63.
<https://doi.org/10.1016/j.biopsycho.2017.04.013>
- Nalborczyk, L., Perrone-Bertolotti, M., Baeyens, C., Grandchamp, R., Spinelli, E., Koster, E. H. W., & Løevenbruck, H. (2020). *Articulatory suppression effects on induced rumination* [Under Review].
- Perrone-Bertolotti, M., Rapin, L., Lachaux, J. P., Baciú, M., & Løevenbruck, H. (2014). What is that little voice inside my head? Inner speech phenomenology, its role in cognitive performance, and its relation to self-monitoring. *Behavioural Brain Research*, 261, 220–239. <https://doi.org/10.1016/j.bbr.2013.12.034>
- Pollard, P., & Richardson, J. T. (1987). On the probability of making Type I errors. *Psychological Bulletin*, 102(1), 159–163. <https://doi.org/10.1037/0033-2909.102.1.159>
- R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rouder, J. N., Morey, R. D., Verhagen, J., Province, J. M., & Wagenmakers, E.-J. (2016). Is There a Free Lunch in Inference? *Topics in Cognitive Science*, 8(3), 520–547.
<https://doi.org/10.1111/tops.12214>
- Wickham, H. (2017). *Tidyverse: Easily install and load the 'tidyverse'*.
<https://CRAN.R-project.org/package=tidyverse>

- 264 Xie, Y. (2015). *Dynamic documents with R and knitr* (2nd ed.). Chapman; Hall/CRC.
265 <https://yihui.org/knitr/>
- 266 Xie, Y., Allaire, J. J., & Golemund, G. (2018). *R markdown: The definitive guide*.
267 Chapman; Hall/CRC. <https://bookdown.org/yihui/rmarkdown>