The problem of illusory power for imaginary interactions

Dario Paape & Shravan Vasishth
University of Potsdam
paape@uni-potsdam.de

Introduction

- Suppose we run an experiment with a 2×2 design, with factors **factor1** and **factor2**, and predict a **statistical interaction** between the factors, e.g.

A.	...	**constraining context**	...	**low-frequency word** ...
B.	...	**constraining context**	...	**high-frequency word** ...
C.	...	**non-constraining context**	...	**low-frequency word** ...
D.	...	**non-constraining context**	...	**high-frequency word** ...

- The correct way of testing for an interaction is to fit the following model and check if the interaction term (**factor1:factor2**) comes out significant:

```
lmer(rt~factor1*factor2+..., data)                    (1)
```

- An alternative, incorrect approach to "interaction" testing is **splitting** the dataset according to **factor1** and then testing for the effect of **factor2** in both of the resulting subsets:

```
subset1<-subset(data,factor1==1)
lmer(rt~factor2+..., subset1)
subset2<-subset(data,factor1==0)
lmer(rt~factor2+..., subset2)                         (2)
```

- Yet another approach is to apply **nested contrasts**, that is, to code comparisons for **factor2** within the levels of **factor1**:

```
data$c1 <- ifelse(data$factor1==1,ifelse(data$factor2==1,1,0),0)
data$c2 <- ifelse(data$factor1==0,ifelse(data$factor2==1,1,0),0)
lmer(rt~c1+c2+factor1+..., data)                      (3)
```

- Under the incorrect approaches (2) and (3), authors argue for an interaction if either of the differences comes out as significant while the other does not – but the interaction term in model (1) tests whether **the difference of the differences** between conditions is significant; this is different from asking whether **one difference is significant and the other is not**

> Note: The difference between significant and not significant is not necessarily statistically significant!

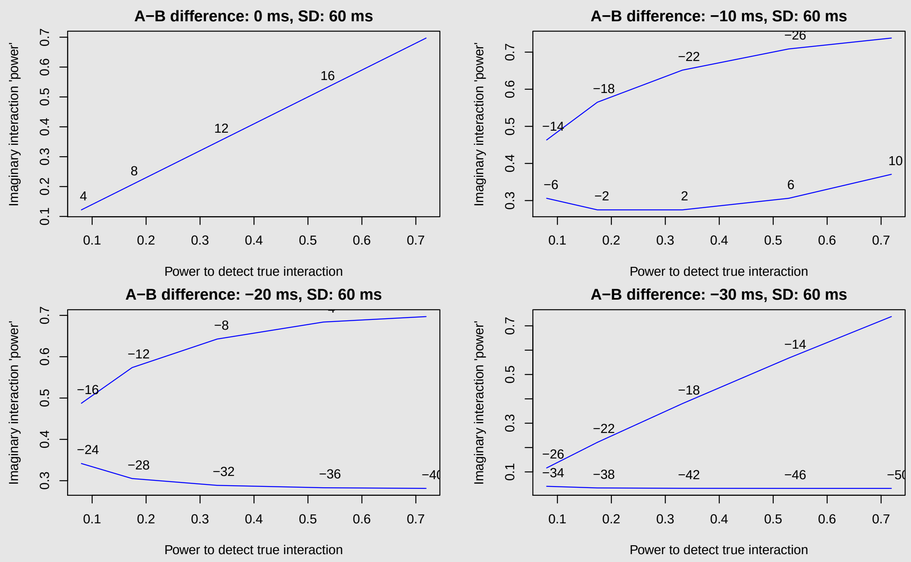
- Significance thresholds are arbitrary, and it's a matter of chance if an effect ends up slighly above or slightly below the criterion (e.g. Gelman & Stern, 2006)

The problem

- The incorrect approaches can systematically lead to **potentially gross overestimates of statistical power**: Depending on the relative sizes of the true nested differences, "imaginary interactions" can lead to **unwarranted discovery claims**
- Illusory "power" inflation is due to detecting one difference but failing to detect the other (**Type II error**)
- The relationship between real and illusory power can be visualized, assuming Hypothesizing After Results Are Known (HARKing; Kerr, 1998) in addition to incorrect analysis
- Formula for illusory "power" with HARKing:

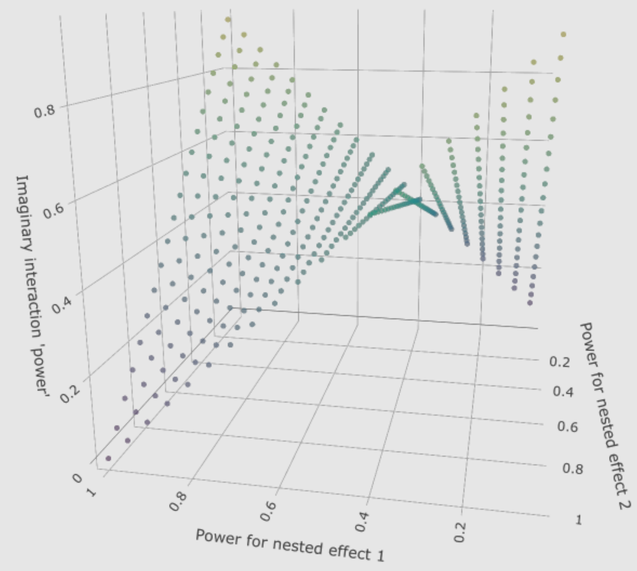
$$P_i = P_{AB} * (1 - P_{CD}) + P_{CD} * (1 - P_{AB})$$

Numbers on lines show C-D difference (in ms)



Further implications

- Again assuming HARKing, plotting illusory "power" for the imaginary interaction against the (actual) power for the nested effects yields a **saddle** shape (independently of the statistical test used):



- Illusory "power" for the imaginary interaction is highest when **actual statistical power is high for one nested difference but low for the other**
- This relationship can be **exploited**: Testing for the effect of a manipulation in two groups or conditions with different variances (e.g. high versus low constraint, native versus non-native speakers, impaired versus unimpaired individuals) will likely produce the required imbalance in statistical power, **even if the true effect sizes are the same**
  - The nested contrasts approach remedies the problem somewhat due to pooling of variances (but heteroskedasticity remains an issue!)

The Shiny app

- Our Shiny app is available at

https://dpaape.shinyapps.io/ipower/

- Assuming a 2×2 design, it allows the user to interactively calculate power for the actual interaction and "power" for the imaginary interaction, along with the power for the nested effects, using `power.t.test`

Take-home message

- The problem of discovery claims based on imaginary interactions is **widespread** in neuroscience (Nieuwenhuis, Forstmann & Wagenmakers, 2011), and probably in psycholinguistics and psychology as well, though a systematic review has not been conducted so far
- As 2×2 factorial designs with predictions for a statistical interaction are the most commonly encountered designs in psycholinguistics, it is imperative that **claimed interactions are actually substantiated by the data**

References. **Gelman & Stern** (2006). Am Stat 60(4), 328–331. **Kerr** (1998). Pers Soc Psychol Rev 2(3), 196–217. **Nieuwenhuis, Forstmann & Wagenmakers** (2011). Nat Neurosci 14(9), 1105–1107.

Created with LaTeX `beamerposter` http://www-i6.informatik.rwth-aachen.de/~dreuw/latexbeamerposter.php