```
course = "Improving your statistical inferences through simulation studies in R"


lesson_iteration = 6
lesson_title = "different ways to analyse RCTs (2X2 within-between experiments)“


auth = "Ian Hussey"
dept = "Psychology of Digitalisation"
```

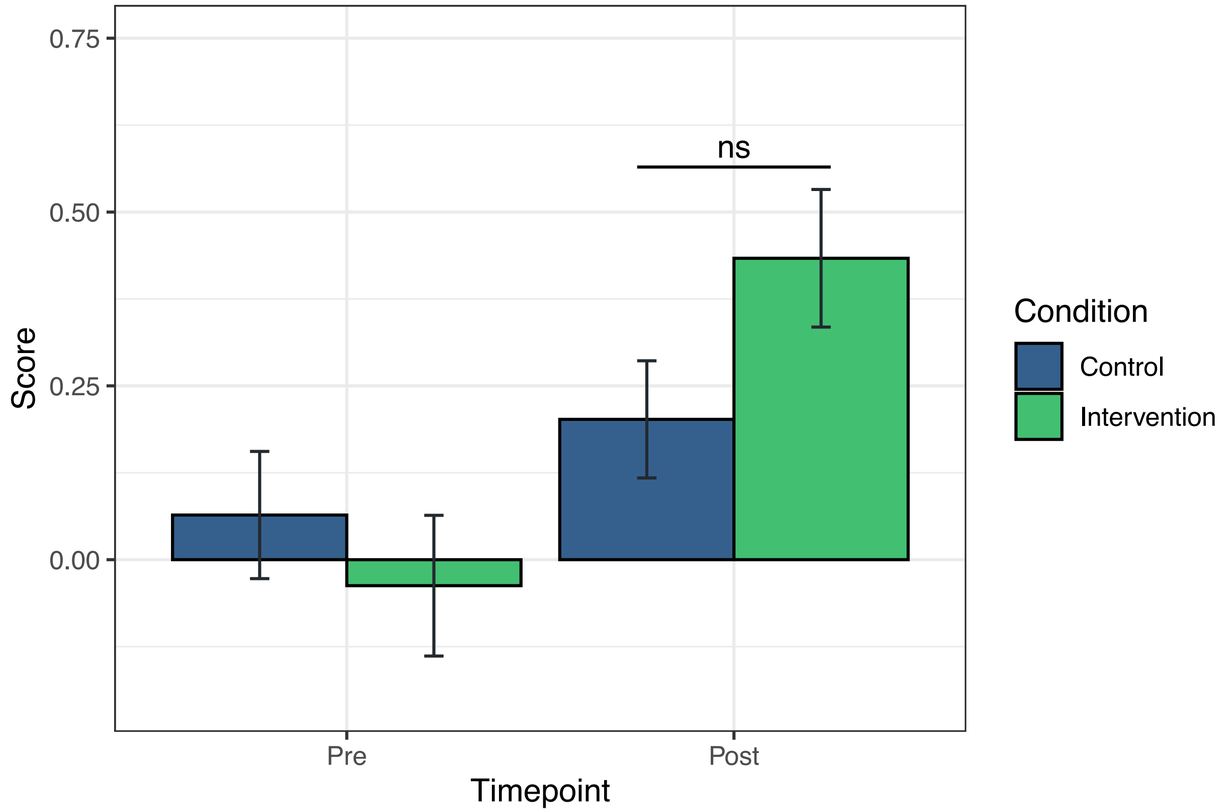
Assignment Qs

- In analytic strategy 4 where all four tests are run, are any of the tests redundant to answering the primary research question of ‘is the intervention effective’?
- What are the inappropriate ways of analyzing these scores? Why?

Assignment Qs

Only the independent t-test comparing scores at post is relevant (i.e., Scenario 2).

All the others are redundant at best, or misleading at worst.



aut = "Ian Hussey";

Assignment Qs

Only the independent t-test comparing scores at post is relevant (i.e., Scenario 2).

All the others are redundant at best, or misleading at worst.

Why?

> Because any differences at baseline must be random, as group allocation was randomised.

  > No statistical test can tell you if you randomised: this is procedural. You either did or didn’t randomise.

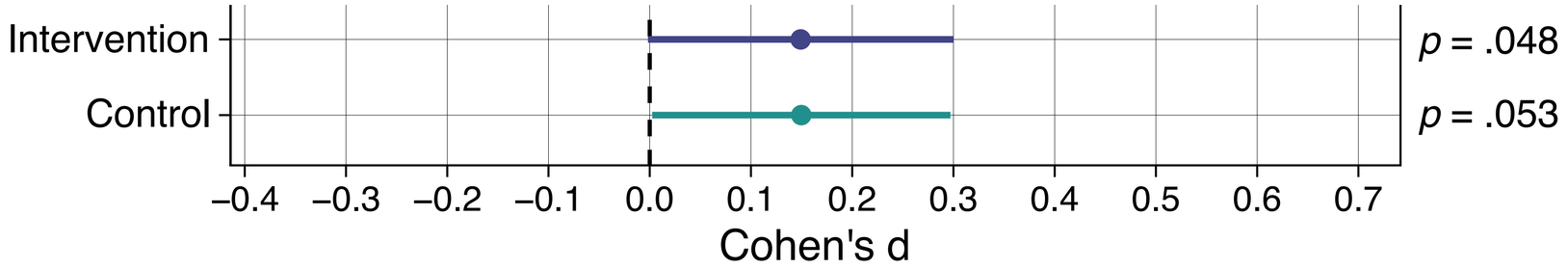
> Differences at post test the key counterfactual:

  > If you had not intervened, how would participants be doing at post?

Assignment Qs

Why are the dependent t-tests within each group misleading?

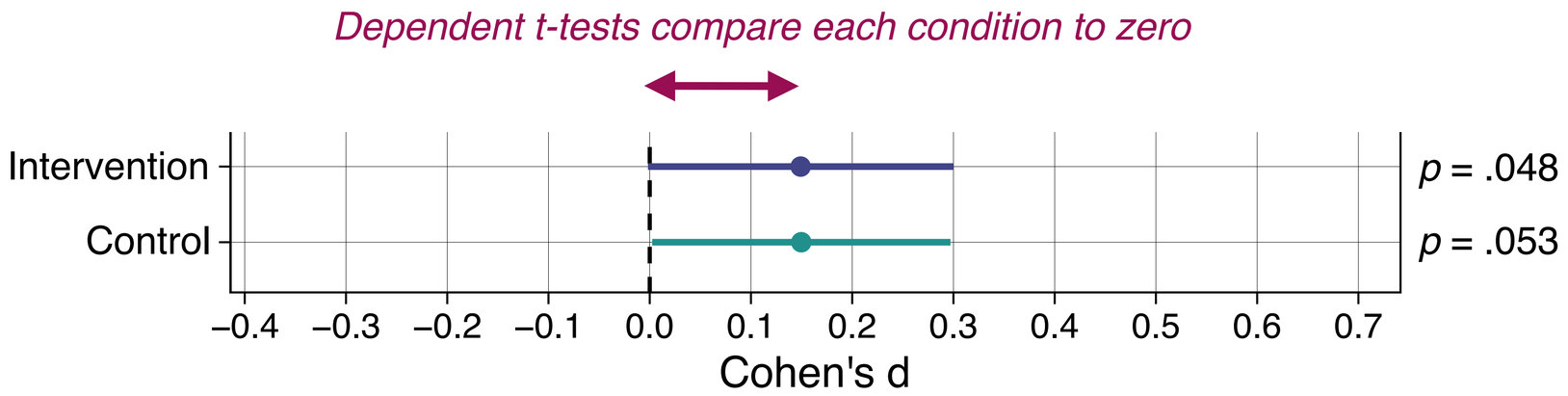
> Because “the difference between significant and non-significant is not itself significant” (Gelman & Stern, 2006)



Assignment Qs

Why are the dependent t-tests within each group misleading?

> Because “the difference between significant and non-significant is not itself significant” (Gelman & Stern, 2006)

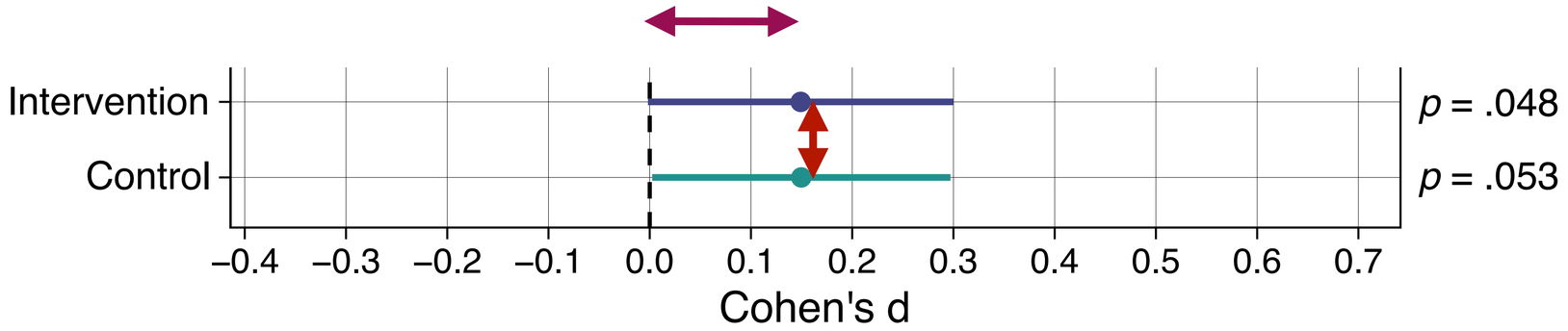


Assignment Qs

Why are the dependent t-tests within each group misleading?

> Because “the difference between significant and non-significant is not itself significant” (Gelman & Stern, 2006)

Dependent t-tests compare each condition to zero



Independent t-test at post compares the conditions against one another

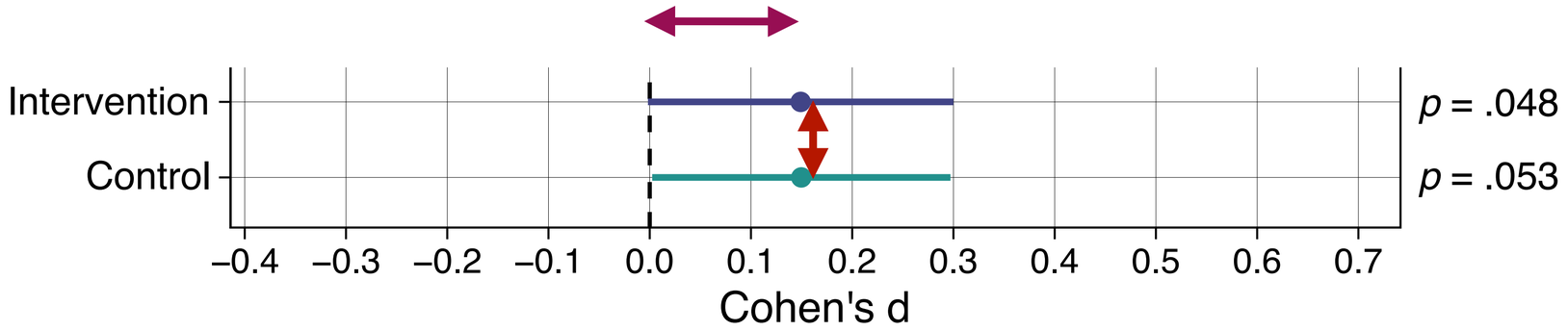
Assignment Qs

Why are the dependent t-tests within each group misleading?

> Because “the difference between significant and non-significant is not itself significant” (Gelman & Stern, 2006)

> If you want to compare whether intervention is more effective than control, you have to *compare intervention to control*

Dependent t-tests compare each condition to zero



Independent t-test at post compares the conditions against one another

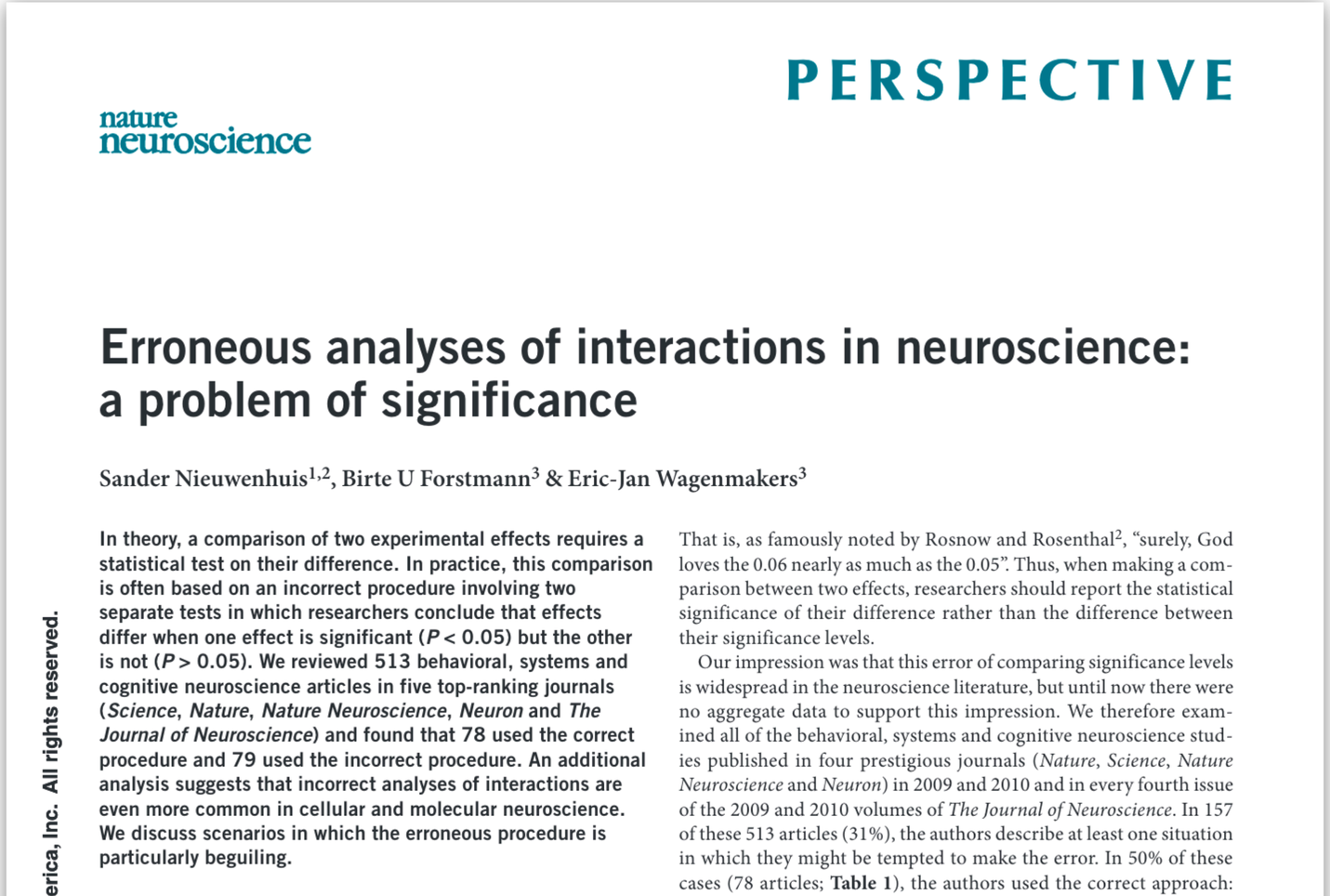
Assignment Qs

aut = "Ian Hussey";

> If you want to compare whether intervention is more effective than control, you have to *compare intervention to control*

Around 50% of published articles don't do this,

undermined their conclusions



PERSPECTIVE

nature neuroscience

Erroneous analyses of interactions in neuroscience: a problem of significance

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

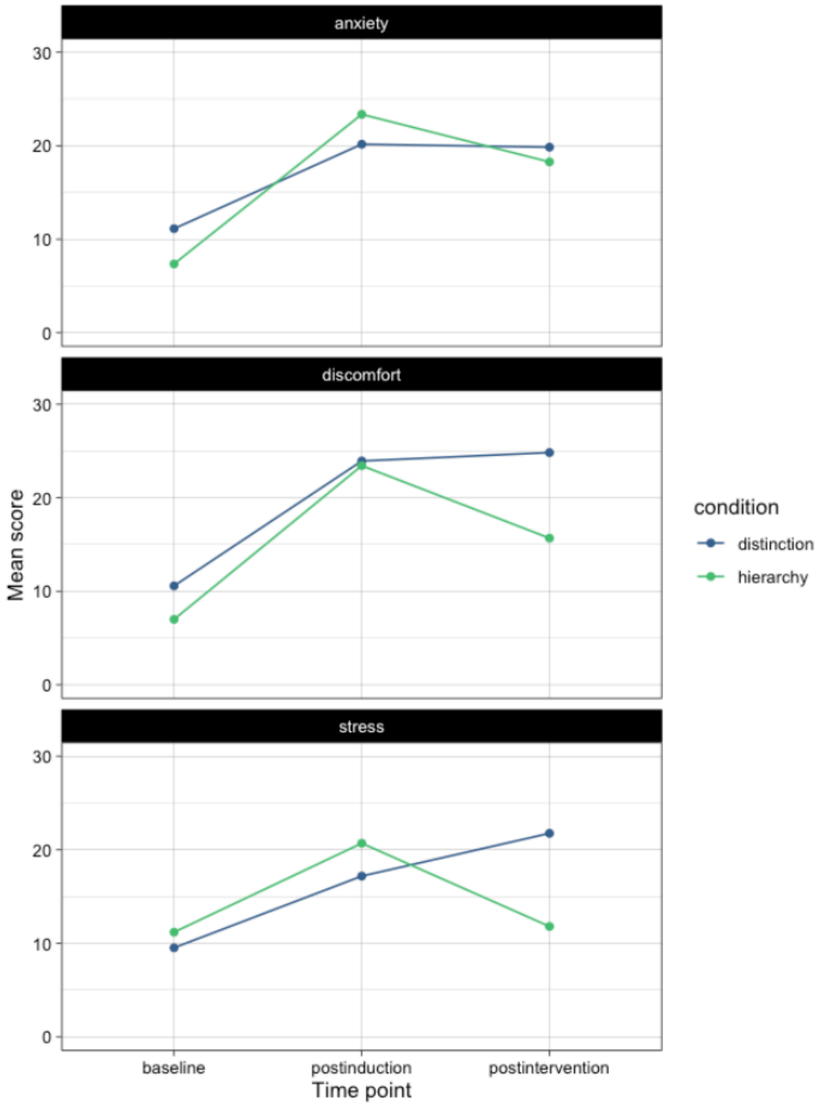
In theory, a comparison of two experimental effects requires a statistical test on their difference. In practice, this comparison is often based on an incorrect procedure involving two separate tests in which researchers conclude that effects differ when one effect is significant ($P < 0.05$) but the 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.

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:

Assignment Qs

Foody et al. (2013) induced distress & compared two different Acceptance and Commitment Therapy interventions to reduce it



Distress measured at baseline

Distress induction procedure to increase it

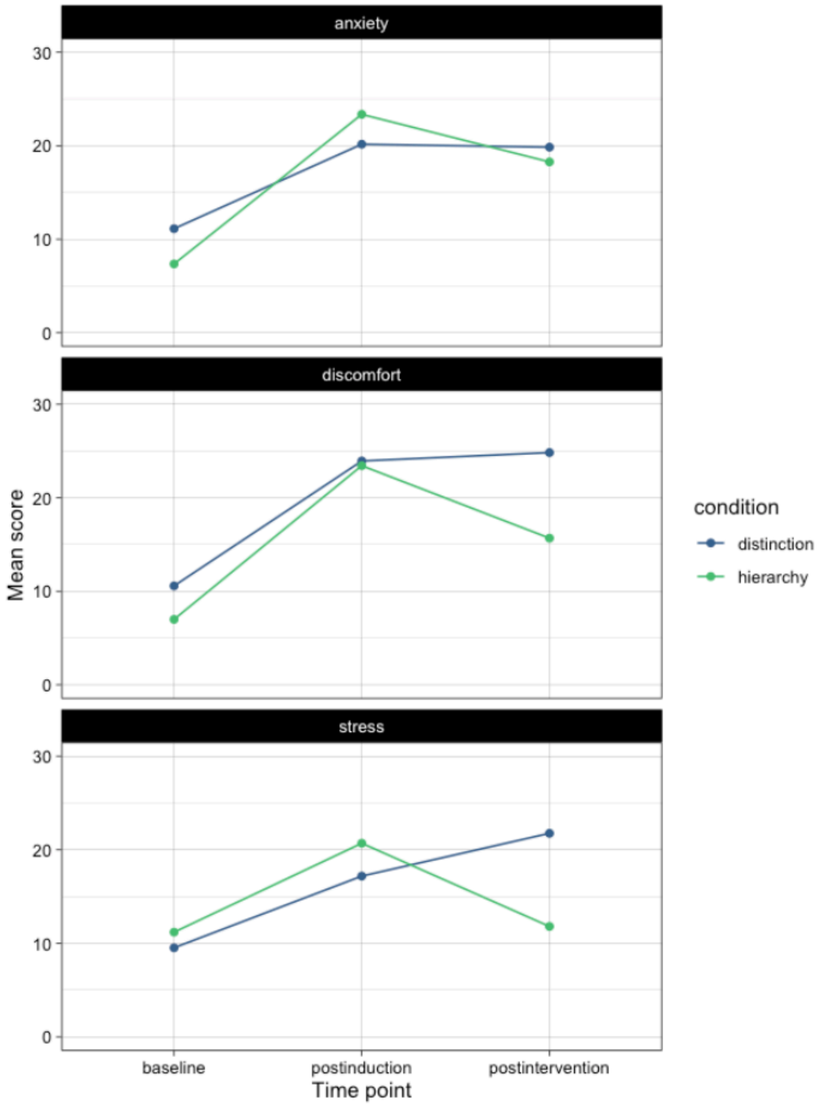
Distress measured post-induction

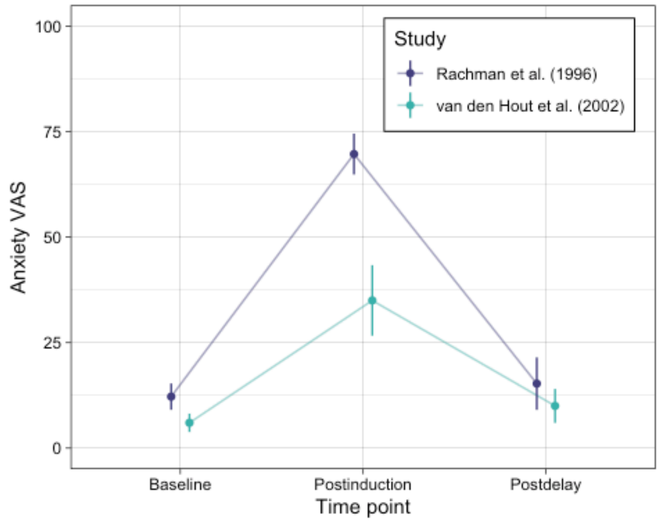
Intervention (‘distinction’ vs. ‘hierarchy’)

Distress measured post-intervention

Assignment Qs

Foody et al. (2013) induced distress & compared two different Acceptance and Commitment Therapy interventions to reduce it





No control condition hides that distress reduces in the absence of intervention

> ‘hierarchy’ intervention was likely less effective than doing nothing
> ‘distinction’ intervention was likely harmful

Assignment Qs

- What other ways are there of analyzing these data?
  - What are the pros and cons of each of them?

Assignment Qs

- What other ways are there of analyzing these data?
  - What are the pros and cons of each of them?

Default / good: Independent t-test applied to scores at post

`score_post ~ condition`

1. Calculate pre-post difference scores, use difference scores into an independent t-test

   `score_pre_post_diff ~ condition`

2. Mixed within-between RM-ANOVA

   `score ~ condition * timepoint (p value for interaction)`

3. ANCOVA with baseline scores entered as a covariate

   `score_post ~ condition + score_pre (p value for condition)`

Assignment Qs

- What other ways are there of analyzing these data?
  - What are the pros and cons of each of them?

Default / good: Independent t-test applied to scores at post

`score_post ~ condition`

1. Calculate pre-post difference scores, use difference scores into an independent t-test

   `score_pre_post_diff ~ condition`

   intuitive but hides some assumptions about baseline

2. Mixed within-between RM-ANOVA

   `score ~ condition * timepoint` (p value for interaction)

   interaction effect can hide nature of effect; no Cohen's d

3. ANCOVA with baseline scores entered as a covariate

   `score_post ~ condition + score_pre` (p value for condition)

   can increases statistical power; but no Cohen's d

Assignment Qs

Does the choice of analysis ever make a difference?

Yes!

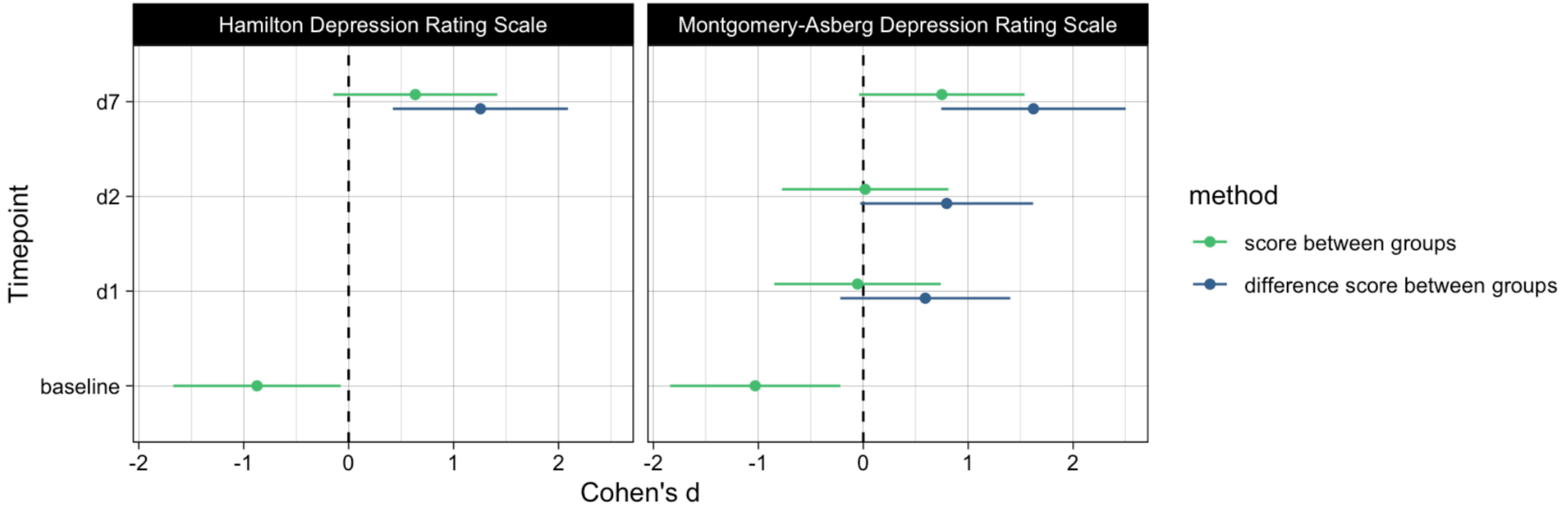
Eg Palhano-Fontes et al. (2019) “Rapid antidepressant effects of the psychedelic ayahuasca in treatment-resistant depression”

`score_pre_post_diff ~ condition`

- Significant results on both outcome measures at follow-up (day 7)

`score_post ~ condition`

- Null results on both outcome measures at follow-up (day 7)



aut = "Ian Hussey";

Readings

> German & Stern (2006) The Difference Between “Significant” and “Not Significant” is not Itself Statistically Significant. The American Statistician. https://doi.org/10.1198/000313006X152649

> Nieuwenhuis et al. (2011) Erroneous analyses of interactions in neuroscience: a problem of significance. Nature Neuroscience. https://doi.org/10.1038/nn.2886

> Van Breukelen (2006) ANCOVA versus change from baseline: more power in randomized studies, more bias in nonrandomized studies. Journal of Clinical Epidemiology. https://doi.org/10.1016/j.jclinepi.2006.02.007

Readings