

Reproducible (Open) Science

Introducing the tools

Marc A.T. Teunis, PhD

2020-06-10 13:56:43

Contents

This is part 1 of a series of three webinars

- ▶ Part 1; Introducing Reproducible (Open) Science (June 11th, 2020)
- ▶ Part 2; Managing your project files and data with 'Guerilla Analytics' (~June 23rd, 2020)
- ▶ Part 3; Reproducible (Open) Science @HU - Tools (July 6th, 2020)

The complete source code for this presentation and all dependent data and files can be found on [Github.com/uashogeschoolutrecht](https://github.com/uashogeschoolutrecht)

Part 1; Introducing Reproducible (Open) Science

1. When things go wrong
2. Why Reproducible (Open) Science?
3. The need for learning programming
4. Reproducible (Open) Science tools overview

Is chloroquine really an option?

As you probably know, hydro-chloroquine was touted as a promising cure for COVID-19 by US President Donald Trump

POLITICS • Analysis

Trump's promotion of hydroxychloroquine is certainly about politics, not profits

At least, not his profits



But how are we really doing with hydro-chloroquine and the treatment for COVID-19?

THE LANCET

ARTICLES | ONLINE FIRST

RETRACTED: Hydroxychloroquine and macrolide for treatment of COVID-19

Prof Mandeep R Mehra, MD   • Sapan S Desai, MD

Published: 1 May 2020 • DOI: [https://doi.org/10.1016/S0140-6736\(20\)30617-8](https://doi.org/10.1016/S0140-6736(20)30617-8)

What was the reason for retracting this paper?

<https://www.sciencemag.org/news/2020/06/two-elite-medical-journals-retract-coronavirus-papers-over-data-integrity-questions>

“Our independent peer reviewers informed us that Surgisphere would not transfer the full dataset, client contracts, and the full ISO audit report to their servers for analysis as such transfer would violate client agreements and confidentiality requirements”

- ▶ Company Surgisphere ('data owner') did not share raw data
- ▶ At time of publication (raw) data and analysis (code) was not included in the manuscript
- ▶ The authors initiated the retract

Why is this a problem?

- ▶ Scientific conclusions get picked up by the media, reversing statements is difficult

Why is this a problem?

- ▶ Scientific conclusions get picked up by the media, reversing statements is difficult
- ▶ The credibility of the Journal, the researchers and the affiliated institutions are at stake

Why is this a problem?

- ▶ Scientific conclusions get picked up by the media, reversing statements is difficult
- ▶ The credibility of the Journal, the researchers and the affiliated institutions are at stake
- ▶ Clinical studies to hydrochloroquine were halted because of this paper

Why is this a problem?

- ▶ Scientific conclusions get picked up by the media, reversing statements is difficult
- ▶ The credibility of the Journal, the researchers and the affiliated institutions are at stake
- ▶ Clinical studies to hydrochloroquine were halted because of this paper
- ▶ The credibility of the company Surgisphere is at stake (they should have seen this. . .)

Why is this a problem?

- ▶ Scientific conclusions get picked up by the media, reversing statements is difficult
- ▶ The credibility of the Journal, the researchers and the affiliated institutions are at stake
- ▶ Clinical studies to hydrochloroquine were halted because of this paper
- ▶ The credibility of the company Surgisphere is at stake (they should have seen this. . .)
- ▶ The credibility of Science as a whole is at stake

The Lancet does not adhere to Reproducible (Open) Science

Would the Lancet have adopted the Reproducible (Open) Science framework:

- ▶ There would have been no publication, so no retraction necessary

The Lancet does not adhere to Reproducible (Open) Science

Would the Lancet have adopted the Reproducible (Open) Science framework:

- ▶ There would have been no publication, so no retraction necessary
- ▶ The manuscript of this paper would not even have made it through the first check round

The Lancet does not adhere to Reproducible (Open) Science

Would the Lancet have adopted the Reproducible (Open) Science framework:

- ▶ There would have been no publication, so no retraction necessary
- ▶ The manuscript of this paper would not even have made it through the first check round
- ▶ All data, code, methods and conclusions would have been submitted

The Lancet does not adhere to Reproducible (Open) Science

Would the Lancet have adopted the Reproducible (Open) Science framework:

- ▶ There would have been no publication, so no retraction necessary
- ▶ The manuscript of this paper would not even have made it through the first check round
- ▶ All data, code, methods and conclusions would have been submitted
- ▶ This would have enabled a complete and thorough peer-review process that includes replication of (part of) the data analysis of the study

The Lancet does not adhere to Reproducible (Open) Science

Would the Lancet have adopted the Reproducible (Open) Science framework:

- ▶ There would have been no publication, so no retraction necessary
- ▶ The manuscript of this paper would not even have made it through the first check round
- ▶ All data, code, methods and conclusions would have been submitted
- ▶ This would have enabled a complete and thorough peer-review process that includes replication of (part of) the data analysis of the study
- ▶ Focus should be on the data and methods, not on the academic narratives and results . . .

The Lancet does not adhere to Reproducible (Open) Science

Would the Lancet have adopted the Reproducible (Open) Science framework:

- ▶ There would have been no publication, so no retraction necessary
- ▶ The manuscript of this paper would not even have made it through the first check round
- ▶ All data, code, methods and conclusions would have been submitted
- ▶ This would have enabled a complete and thorough peer-review process that includes replication of (part of) the data analysis of the study
- ▶ Focus should be on the data and methods, not on the academic narratives and results . . .
- ▶ In physics and bioinformatics this is already common practice

Data, methods and logic

"...in science, three things matter:

1. the data,

everything else is a distraction."

Brown, Kaiser & Allison, PNAS, 2018

Data, methods and logic

"...in science, three things matter:

1. the data,
2. the methods used to collect the data [...], and

everything else is a distraction."

Brown, Kaiser & Allison, PNAS, 2018

Data, methods and logic

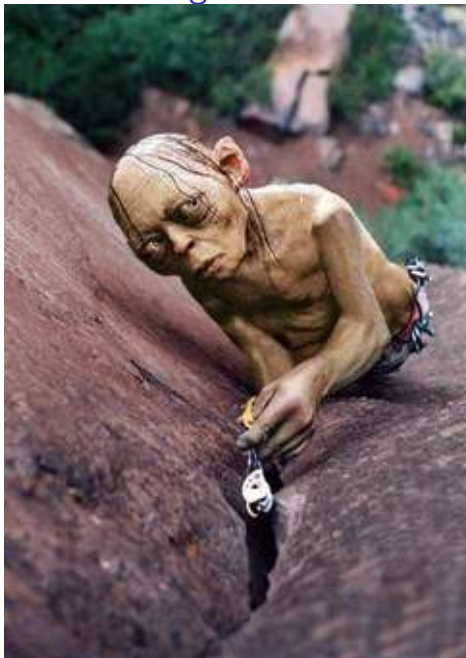
"...in science, three things matter:

1. the data,
2. the methods used to collect the data [...], and
3. the logic connecting the data and methods to conclusions,

everything else is a distraction."

Brown, Kaiser & Allison, PNAS, 2018

Gollums lurking about



Why Reproducible (Open) Science?

- ▶ To assess validity of science and methods we need access to data, methods and conclusions

Why Reproducible (Open) Science?

- ▶ To assess validity of science and methods we need access to data, methods and conclusions
- ▶ To learn from choices other researchers made

Why Reproducible (Open) Science?

- ▶ To assess validity of science and methods we need access to data, methods and conclusions
- ▶ To learn from choices other researchers made
- ▶ To learn from omissions, mistakes or errors

Why Reproducible (Open) Science?


- ▶ To assess validity of science and methods we need access to data, methods and conclusions
- ▶ To learn from choices other researchers made
- ▶ To learn from omissions, mistakes or errors
- ▶ To prevent publication bias (also negative results will be available in reproducible research)

Why Reproducible (Open) Science?

- ▶ To assess validity of science and methods we need access to data, methods and conclusions
- ▶ To learn from choices other researchers made
- ▶ To learn from omissions, mistakes or errors
- ▶ To prevent publication bias (also negative results will be available in reproducible research)
- ▶ To be able to re-use and/or synthesize data (from many and diverse sources)

Example of impossibility

How would you 'describe' the steps of an analysis or creation of a graph when you use GUI* based software?



| | A | B | C | D | E | F | G | H |
|----|---------------------------------------|-----------------|--------------------------|--------|-------|------------|---------|-------|
| 1 | Correlation LLNA-EC ₃ data | | | | | | | |
| 2 | | | | | | | | |
| 3 | Compound | EC ₃ | EC ₅₀ (mg/ml) | | | | | |
| 4 | | | HU | DisFeB | VUmc | BASF | Average | |
| 5 | Benzocaine | 6.57 | 21.45 | 37.02 | 17.45 | 17.42 | 23.34 | 9.32 |
| 6 | Citral | 13.00 | 5.08 | 4.63 | 4.45 | 5.82 | 5.00 | 0.61 |
| 7 | Eugenol | 13.00 | 16.30 | 16.33 | 8.76 | 9.31 | 12.67 | 4.21 |
| 8 | x-Hexylcinnamaldehyde | 11.00 | 130.25 | 77.22 | 78.60 | 72.78 | 89.71 | 27.14 |
| 9 | 2-Mercaptobenzothiazol | 1.70 | 10.78 | 15.23 | 16.95 | not tested | 14.32 | 3.18 |
| 10 | Cinnamaldehyde | 1.40 | 5.56 | 2.72 | 1.80 | 5.38 | 3.86 | 1.89 |
| 11 | Isoeugenol | 1.20 | 13.72 | 18.92 | 14.21 | 12.14 | 14.75 | 2.92 |

Programming is essential for Reproducible (Open) Science

- ▶ Only programming an analysis (or creation of a graph) records every step

(Literate) programming is a way to connect narratives to data, methods and results

```
78  
79 - only programming can really  
80 - Learning to use a programming  
graphs takes time but pays off in the long run  
81 - Programming could also solve  
82  
83 _Literate programming is a way  
84  
85
```

Programming is essential for Reproducible (Open) Science

- ▶ Only programming an analysis (or creation of a graph) records every step
- ▶ The script(s) function as a (data) analysis journal

(Literate) programming is a way to connect narratives to data, methods and results

```
79 - only programming can really  
80 - Learning to use a programming  
graphs takes time but pays off in the long run  
81 - Programming could also solve  
82  
83 _Literate programming is a way
```

Programming is essential for Reproducible (Open) Science

- ▶ Only programming an analysis (or creation of a graph) records every step
- ▶ The script(s) function as a (data) analysis journal
- ▶ Learning to use a programming language takes time but pays off at the long run (for all of science)

(Literate) programming is a way to connect narratives to data, methods and results

```
78  
79 - only programming can really  
80 - Learning to use a programming  
graphs takes time but pays off at  
81 - Programming could also solve  
82  
83 _Literate programming is a way  
84  
85
```

To replicate a scientific study we need at least:

- ▶ Scientific context, research questions and state of the art [P]

*P = Publication, D = Data, C = Code, OAcc = Open Access,
OSrc = Open Source*

To replicate a scientific study we need at least:

- ▶ Scientific context, research questions and state of the art [P]
- ▶ (Experimental) model or characteristics of population or matter studied [P]

*P = Publication, D = Data, C = Code, OAcc = Open Access,
OSrc = Open Source*

To replicate a scientific study we need at least:

- ▶ Scientific context, research questions and state of the art [P]
- ▶ (Experimental) model or characteristics of population or matter studied [P]
- ▶ Data that was generated and corresponding meta data [D, C]

*P = Publication, D = Data, C = Code, OAcc = Open Access,
OSrc = Open Source*

To replicate a scientific study we need at least:

- ▶ Scientific context, research questions and state of the art [P]
- ▶ (Experimental) model or characteristics of population or matter studied [P]
- ▶ Data that was generated and corresponding meta data [D, C]
- ▶ **Exact** (experimental) design of the study [P, D, C]

*P = Publication, D = Data, C = Code, OAcc = Open Access,
OSrc = Open Source*

To replicate a scientific study we need at least:

- ▶ Scientific context, research questions and state of the art [P]
- ▶ (Experimental) model or characteristics of population or matter studied [P]
- ▶ Data that was generated and corresponding meta data [D, C]
- ▶ **Exact** (experimental) design of the study [P, D, C]
- ▶ Exploratory data analysis of the data [P, C]

*P = Publication, D = Data, C = Code, OAcc = Open Access,
OSrc = Open Source*

To replicate a scientific study we need at least:

- ▶ Scientific context, research questions and state of the art [P]
- ▶ (Experimental) model or characteristics of population or matter studied [P]
- ▶ Data that was generated and corresponding meta data [D , C]
- ▶ **Exact** (experimental) design of the study [P , D , C]
- ▶ Exploratory data analysis of the data [P , C]
- ▶ **Exact** methods that were used to conduct any formal inference [P , C]

P = *Publication*, D = *Data*, C = *Code*, $OAcc$ = *Open Access*,
 $OSrc$ = *Open Source*

To replicate a scientific study we need at least:

- ▶ Scientific context, research questions and state of the art [P]
- ▶ (Experimental) model or characteristics of population or matter studied [P]
- ▶ Data that was generated and corresponding meta data [D, C]
- ▶ **Exact** (experimental) design of the study [P, D, C]
- ▶ Exploratory data analysis of the data [P, C]
- ▶ **Exact** methods that were used to conduct any formal inference [P, C]
- ▶ Model diagnostics [C]

*P = Publication, D = Data, C = Code, OAcc = Open Access,
OSrc = Open Source*

To replicate a scientific study we need at least:

- ▶ Scientific context, research questions and state of the art [P]
- ▶ (Experimental) model or characteristics of population or matter studied [P]
- ▶ Data that was generated and corresponding meta data [D, C]
- ▶ **Exact** (experimental) design of the study [P, D, C]
- ▶ Exploratory data analysis of the data [P, C]
- ▶ **Exact** methods that were used to conduct any formal inference [P, C]
- ▶ Model diagnostics [C]
- ▶ Interpretations of the (statistical) model results/model fitting process [P, C]

*P = Publication, D = Data, C = Code, OAcc = Open Access,
OSrc = Open Source*

To replicate a scientific study we need at least:

- ▶ Scientific context, research questions and state of the art [P]
- ▶ (Experimental) model or characteristics of population or matter studied [P]
- ▶ Data that was generated and corresponding meta data [D, C]
- ▶ **Exact** (experimental) design of the study [P, D, C]
- ▶ Exploratory data analysis of the data [P, C]
- ▶ **Exact** methods that were used to conduct any formal inference [P, C]
- ▶ Model diagnostics [C]
- ▶ Interpretations of the (statistical) model results/model fitting process [P, C]
- ▶ Conclusions and academic scoping of the results [P, C]

*P = Publication, D = Data, C = Code, OAcc = Open Access,
OSrc = Open Source*

To replicate a scientific study we need at least:

- ▶ Scientific context, research questions and state of the art [P]
- ▶ (Experimental) model or characteristics of population or matter studied [P]
- ▶ Data that was generated and corresponding meta data [D, C]
- ▶ **Exact** (experimental) design of the study [P, D, C]
- ▶ Exploratory data analysis of the data [P, C]
- ▶ **Exact** methods that were used to conduct any formal inference [P, C]
- ▶ Model diagnostics [C]
- ▶ Interpretations of the (statistical) model results/model fitting process [P, C]
- ▶ Conclusions and academic scoping of the results [P, C]
- ▶ **Access to all of the above** [OAcc, OSrc]

P = Publication, D = Data, C = Code, OAcc = Open Access, OSrc = Open Source

Example; The Open Science Framework OSF



OSF - Reproducible Project: Psychology

- ▶ 100 publications in Psychology journals
- ▶ Results from half of these publications could be reproduced
- ▶ Full access to P, D and C in OSF
- ▶ The publication is not published in an OAcc journal but:
- ▶ The submitted manuscript is available in OSF
- ▶ The R code used is available in OSF

RP : Psychology = P + D + C + OSrc (+OAcc)

A short, but complete example of reproducible science

Assume we have the following question:

“Which of 4 types of chairs takes the least effort to arise from when seated in?”

We have the following setup:

- ▶ 4 different types of chairs
- ▶ 9 different subjects (probably somewhat aged)
- ▶ Each subject is required to provide a score (from 6 to 20, 6 being very lightly strenuous, 20 being very very extremely strenuous) when arising from each of the 4 chairs.

To analyze this experiment statistically,
the model would need to include: the rating score as the **measured (or dependent) variable**, the type of chair as the **experimental factor** and the subject as the **blocking factor**

Mixed effects models

A typical analysis method for this type of randomized block design is a so-called 'multi-level' or also called 'mixed-effects' or 'hierarchical' models. An analysis method much used in clinical or biological scientific practice.

You could also use one-way ANOVA but I will illustrate why this is not a good idea

What do we need to replicate the science of this experiment?

I will illustrate:

- ▶ the data and an exploratory graph,
- ▶ the statistical model,
- ▶ the statistical model and the results
- ▶ the conclusions

in the next four slides. Hopefully this will illustrate the power of using literate programming to communicate such an analysis.

Example reproduced from: Pinheiro and Bates, 2000,
Mixed-Effects Models in S and S-PLUS, Springer, New York.

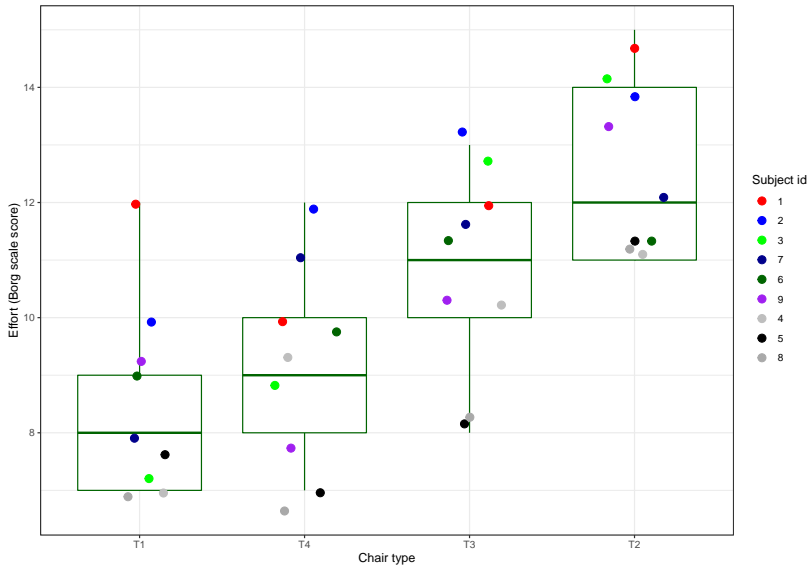
The data of the experiment

Wretenberg, Arborelius & Lindberg, 1993

```
library(nlme)
ergoStool %>% as_tibble()
```

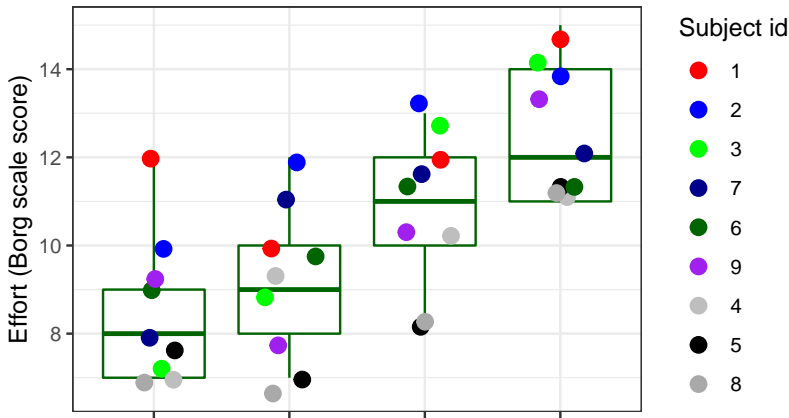
```
## # A tibble: 36 x 3
##   effort Type  Subject
##   <dbl> <fct> <ord>
## 1     12 T1      1
## 2     15 T2      1
## 3     12 T3      1
## 4     10 T4      1
## 5     10 T1      2
## 6     14 T2      2
## 7     13 T3      2
## 8     12 T4      2
## 9      7 T1      3
## 10    14 T2      3
## # with 26 more rows
```

An exploratory graph



Mind the variability per subject, what do you see?

- ▶ Can you say something about within-subject variability (note 'Minster Blue')?
- ▶ Can you say something about between-subject variability (note 'Mister Green', vs 'Mister Black')?
- ▶ Which chair type takes, on average the biggest effort to arise from?



The statistical questions

1. Which chair type takes, on average the biggest effort to arise from? (ANOVA / MEM, fixed effects)
 - ▶ Do individual (within subject) differences play a role in appointing a average score to a chair type? (MEM, random effects)
 - ▶ Does variability between subjects play a role in determining the 'best' chair type (ANOVA / MEM, confidence intervals)

The statistical model

Statistical models (in R) can be specified by a model formula. The left side of the formula is the dependent variable, the right side are the 'predictors'. Here we include a fixed and a random term to the model (as is common for mixed-effects models)

```
library(nlme)
```

```
ergo_model <- lme(  
  data = ergoStool, # the data to be used for the model  
  fixed = effort ~ Type, # the dependent and fixed effects  
  random = ~1 | Subject # random intercepts for Subject variables  
)
```

The `lme()` function is part of the `{nlme}` package for mixed effects modelling in R

Example reproduced from: Pinheiro and Bates, 2000, *Mixed-Effects Models in S and S-PLUS*, Springer, New York.

The statistical results

| | Value | Std.Error | DF | t-value | p-value |
|-------------|-----------|-----------|----|-----------|-----------|
| (Intercept) | 8.5555556 | 0.5760123 | 24 | 14.853079 | 0.0000000 |
| TypeT2 | 3.8888889 | 0.5186838 | 24 | 7.497610 | 0.0000001 |
| TypeT3 | 2.2222222 | 0.5186838 | 24 | 4.284348 | 0.0002563 |
| TypeT4 | 0.6666667 | 0.5186838 | 24 | 1.285305 | 0.2109512 |

Model diagnostics

- ▶ Diagnostics of a fitted model is the most important step in a statistical analysis

A residual plot shows the 'residual' error ('unexplained variance') after fitting the model. Under the Normality assumption standardized residuals should:

Model diagnostics

- ▶ Diagnostics of a fitted model is the most important step in a statistical analysis
- ▶ In most scientific papers the details are lacking

A residual plot shows the 'residual' error ('unexplained variance') after fitting the model. Under the Normality assumption standardized residuals should:

Model diagnostics

- ▶ Diagnostics of a fitted model is the most important step in a statistical analysis
- ▶ In most scientific papers the details are lacking
- ▶ Did the authors omit to perform this step? Or did they not report it?

A residual plot shows the 'residual' error ('unexplained variance') after fitting the model. Under the Normality assumption standardized residuals should:

Model diagnostics

- ▶ Diagnostics of a fitted model is the most important step in a statistical analysis
- ▶ In most scientific papers the details are lacking
- ▶ Did the authors omit to perform this step? Or did they not report it?
- ▶ If you do not want to include it in your paper, put it in an appendix!

A residual plot shows the 'residual' error ('unexplained variance') after fitting the model. Under the Normality assumption standardized residuals should:

Model diagnostics

- ▶ Diagnostics of a fitted model is the most important step in a statistical analysis
- ▶ In most scientific papers the details are lacking
- ▶ Did the authors omit to perform this step? Or did they not report it?
- ▶ If you do not want to include it in your paper, put it in an appendix!

A residual plot shows the 'residual' error ('unexplained variance') after fitting the model. Under the Normality assumption standardized residuals should:

1. Be normally distributed around 0

Model diagnostics

- ▶ Diagnostics of a fitted model is the most important step in a statistical analysis
- ▶ In most scientific papers the details are lacking
- ▶ Did the authors omit to perform this step? Or did they not report it?
- ▶ If you do not want to include it in your paper, put it in an appendix!

A residual plot shows the 'residual' error ('unexplained variance') after fitting the model. Under the Normality assumption standardized residuals should:

1. Be normally distributed around 0
2. Display no obvious 'patterns'

Model diagnostics

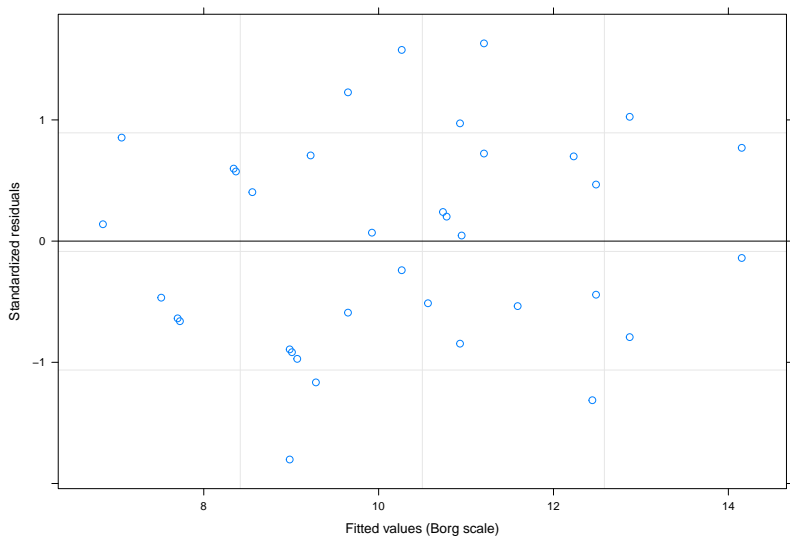
- ▶ Diagnostics of a fitted model is the most important step in a statistical analysis
- ▶ In most scientific papers the details are lacking
- ▶ Did the authors omit to perform this step? Or did they not report it?
- ▶ If you do not want to include it in your paper, put it in an appendix!

A residual plot shows the 'residual' error ('unexplained variance') after fitting the model. Under the Normality assumption standardized residuals should:

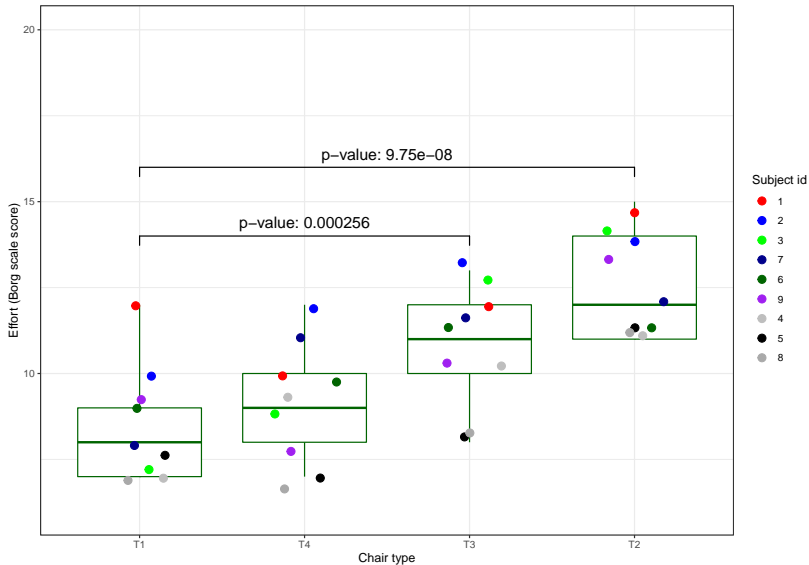
1. Be normally distributed around 0
2. Display no obvious 'patterns'
3. Should display overall equal 'spread' above and below 0 ('assumption of equal variance')

Residual plot

```
plot(ergo_model) ## type = 'pearson' (standardized residuals)
```



The conclusions in a plot



And the most important part...

odz: *Practice what you preach*

If you want to reproduce, add-on, falsify or apply your own ideas to this example, you can find the code (and data) in Github.com

In webinar 3, I will show you how to actually run, use and organize code like this!



Thank you for your attention!

