

R Markdown Lesson 04: Using Papaja

Alexander Strobel<sup>1</sup> & Christoph Scheffel<sup>1</sup>

<sup>1</sup> Technische Universität Dresden

Author Note

Alexander Strobel, Christoph Scheffel, Faculty of Psychology, Technische Universität Dresden, 01062 Dresden, Germany; This work was supported by SFB 940/2.

Correspondence concerning this article should be addressed to Alexander Strobel, Faculty of Psychology, Technische Universität Dresden, 01062 Dresden, Germany. E-mail: alexander.strobel@tu-dresden.de

## Abstract

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

One sentence clearly stating the **general problem** being addressed by this particular study.

One sentence summarizing the main result (with the words “**here we show**” or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more **general context**.

Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

*Keywords:* keywords

Word count: X

## R Markdown Lesson 04: Using Papaja

In this lesson, we will now come to the really helpful R *papaja* written by Aust and Barth (2020) that enables you to format your R Markdown document according to APA style, although currently, only the 6th edition (American Psychological Association, 1994) is supported, to easily report results of common statistical procedures and to appropriately format and place tables and figures. To create an R Markdown manuscript with *papaja*, you first need to install the *papaja* package from GitHub via

```
devtools::install_github("crsh/papaja")
devtools::install_github("crsh/papaja@devel")
```

Then choose **File > New File > R Markdown... > From Template** and select the template *APA-style manuscript (6th edition) {papaja}* upon which a document similar to this one is created that contains a header where you can enter the title and the running head (entry **shorttitle**), all the authors with their affiliations, the author note and so on. Under **bibliography**, per default *papaja* states **r-references.bib**. This file is automatically generated and includes references to the R version you are using as well as to all the packages you source in your manuscript. You will of course need additional references, so you need to create a separate BibTeX file with all the further references needed and list it along with **r-references.bib** under **bibliography** (see the header of this file). You now first write your Introduction section as usual using R Markdown language, citing your references in your BibTeX library either in the text like Aust and Barth (2020) or in parentheses like (Aust & Barth, 2020).

## Methods

We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study.

Another fine issue of the *papaja* template is that it per default puts the 21-word solution (cf. Simmons, Nelson, & Simonsohn, 2012) at the beginning of your Methods section. This is also the place where you should note whether your study was preregistered and comes with open data and code (and if so, link to the respective repository).

Now you are ready to write your Methods section. It is recommended that your analyses should not be run within one big R code chunk, but in separate ones as your analysis pipeline progresses, i.e., one code chunk for demographics ahead of the *Participants* subsection, one code chunk for descriptive statistics, another one for analysis 1 yet another one for analysis 2 and so on ... But we need some data, so we generate five random variables along with some age and gender variable.

## Participants

Without knowing the actual sample size, gender composition and age statistics, you can use placeholders that refer to the data and write:

The sample consisted of 256 participants (125 women, age  $M = 21.99$ ,  $SD = 1.03$ , range 20-25 years).

This enables you to write large parts of your Results section without resorting to the actual data or to data at all, as long as you know the respective variable names in your final data set.

## Material

In this subsection, you describe your material, e.g., questionnaires or stimuli used. Often enough, this will be material that you already described in earlier manuscript, and more likely than not it took you some time to arrive the perfect description of your material back then. While it is generally accepted that authors should omit self-plagiarism, an optimal description of a material or a procedure—at least in our opinion—can and

indeed *should* be reused. Would Pythagoras have paraphrased his theorem  $a^2 + b^2 = c^2$  in later writings as “if you square  $a$  and add the square of  $b$ , this equals the square of  $c$ ” or “the sum of squares of  $x$  and  $y$  gives the square of  $z$ ?” So, you may save your standard material description in a separate R Markdown file and include it in your main manuscript as a so-called *child*.

*Need for Cognition* was assessed with the 16-item short version of the German NFC scale (Bless, Wänke, Bohner, Fellhauer, & Schwarz, 1994). Responses to each item (e.g., “Thinking is not my idea of fun,” recoded) were recorded on a 7-point Likert scale ranging from -3 (completely disagree) to +3 (completely agree). The scale shows comparably high internal consistency [Cronbach’s  $\alpha > .80$ ; Bless et al. (1994); Fleischhauer et al. (2010)].

Of course, you need to add any references in your separate R Markdown file to your main BibTeX file by hand. Another option would be to save all the entries of your separate R Markdown file as a separate BibTeX file too and to attach this file to your manuscript in the `bibliography` entry in the document header. As a demonstration: the file “SCS.bib” is added to the `bibliography` entry and the instrument description is added to the present manuscript as a *child*.

*Self-Control* was measured using the short form of the German Self-Control Scale [SCS-K-D; Bertrams and Dickhäuser (2009)] that comprises 13 items (e.g., “I am able to work effectively toward long-term goals”) with a 5-point Likert scale ranging from -2 (completely disagree) to +2 (completely agree). The scale shows comparably high reliability [Cronbach’s  $\alpha \sim .80$ , 7-week retest reliability  $r_{tt} = .82$ ; Bertrams and Dickhäuser (2009)].

## Procedure

Here, you describe in all necessary detail how your study was conducted. This is one of the most important parts of your manuscript, because ideally, everyone reading your

manuscript should be able to directly replicate your study. You may assume that writing something like “Participants were seated in a dimly lit room . . . stimuli were presented on a 21 inch monitor . . . participants responded via the keyboard,” should do. Actually, you should provide some measure of what *dimly lit* actually meant in your lab, should give the point size of your stimuli and the angle of separation together with the refresh rate of your monitor (ideally also the monitor’s brand and model number), and the same holds true for your keyboard. Gamers will tell you that keyboards really can make a difference . . .

## Data analysis

We used R (Version 4.1.1; R Core Team, 2021) and the R-packages *lavaan* (Version 0.6.9; Rosseel, 2012), *papaja* (Version 0.1.0.9997; Aust & Barth, 2020), *psych* (Version 2.1.9; Revelle, 2021), *pwr* (Version 1.3.0; Champely, 2020), and *tinylabels* (Version 0.2.1; Barth, 2021) for all our analyses.

The above sentence demonstrates one great feature of *papaja*: via the `cite_r` function, it automatically reports the R version and all the packages used at the beginning of the section *Data analysis* in the Methods section. As this can be somewhat annoying because even automatically loaded packages will be listed—making the packages list quite long—and because if using RStudio, this is not automatically reported, you may want to refer to the main packages used by hand and give all the supporting packages in the supplement.

## Results

### Descriptives

Here, you run another R code chunk and provide a table with all the relevant descriptive statistics. As above for the demographics, you can write all the code already without having actual data. Now, we use the `apa_table` function to do so:

```
apa_table(describe(df)[1:5 , c(3:5, 11:12)], caption = "Descriptive statistics of the va
```

Note that the table is automatically placed at the end of the manuscript, just as APA 6th edition style requires. The same is true for figures:

```
boxplot(df[, 1:5], lty = 1, staplewex = NA)
```

## Correlation analysis

Now, we want to run a correlation analysis of our five variables of interest. Again, we use the `apa_table` function, this time along with the `printnum` function that lets you control how your numeric values are presented. APA 6th requires you to provide correlations without a zero before the dot, because correlations cannot be greater than one.

```
correlations = cor(df[, 1:5])
apa_table(printnum(correlations, gt1 = F),
          caption = "Intercorrelations of the variables of interest")
```

This already looks fine, but the ones in the diagonal of the correlation matrix are now presented as “> .99.” We would rather have “–” in the diagonal (and some note to the table) and therefore write:

```
correlations = sub("> .99", "--", printnum(cor(df[, 1:5]), gt1 = F))
apa_table(correlations,
          caption = "Intercorrelations of the variables of interest",
          note = "\\textit{N} = 256",
          escape = F)
```

The `escape = F` argument enables you to use LaTeX syntax in your caption or note. You could also want to present the reliabilities of the variables in question in the diagonal, and you might want to present them in italics. To do so, write:

```
diag(correlations) = paste0("\\textit{(", printnum(runif(5, .7, .9), gt = F), ")}")

apa_table(correlations,
          caption = "Intercorrelations of the variables of interest",
          note = "\\textit{N} = 256",
          escape = F)
```

The latter example shows how you can combine *papaja* functions with LaTeX commands to format your output *ad libitum*.

The *papaja* package also comes with a number of analysis-specific output formats. Consider, you want to perform a regression analysis and to regress X1 on X2. In this case, *papaja* allows you to provide the relevant statistics by referring to the slots generated by `apa.print`. As an example, your code could look like:

```
model = lm(X2 ~ X1, data = df)
papaja_model = apa_print(model)
```

In the text, you would write: X1 predicted X2,  $b = 0.31$ , 95% CI [0.19, 0.43],  $t(254) = 5.08$ ,  $p < .001$ , with  $R^2 = .09$ ,  $F(1, 254) = 25.85$ ,  $p < .001$ . This way, you do not need to extract all the statistics from your analyses and type them in, but simply refer to a statistical model and set a placeholder for the—fully formatted—statistical results. When you later on decide (or being forced by reviewers) to drop some outliers earlier on in your R Markdown file, your manuscript still will give you the correct results for this analysis.



## Discussion

149

150

This is the hard part, and *papaja* will not help you in writing your discussion.

## References

- American Psychological Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, D.C.: American Psychological Association.
- Aust, F., & Barth, M. (2020). *papaja: Prepare reproducible APA journal articles with R Markdown*. Retrieved from <https://github.com/crsh/papaja>
- Barth, M. (2021). *tinylabels: Lightweight variable labels*. Retrieved from <https://github.com/mariusbarth/tinylabels>
- Bertrams, A., & Dickhäuser, O. (2009). Measuring dispositional self-control capacity. A German adaptation of the short form of the Self-Control Scale (SCS-K-D). *Diagnostica*, 55(1), 2–10. <https://doi.org/10.1026/0012-1924.55.1.2>
- Bless, H., Wänke, M., Bohner, G., Fellhauer, R. L., & Schwarz, N. (1994). Need for Cognition: Eine Skala zur Erfassung von Engagement und Freude bei Denkaufgaben [Need for Cognition: A scale measuring engagement and happiness in cognitive tasks]. *Zeitschrift für Sozialpsychologie*, 25, 147–154.
- Champely, S. (2020). *Pwr: Basic functions for power analysis*. Retrieved from <https://CRAN.R-project.org/package=pwr>
- Fleischhauer, M., Enge, S., Brocke, B., Ullrich, J., Strobel, A., & Strobel, A. (2010). Same or different? Clarifying the relationship of Need for Cognition to personality and intelligence. *Personality & Social Psychology Bulletin*, 36(1), 82–96. <https://doi.org/10.1177/0146167209351886>
- R Core Team. (2021). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>
- Revelle, W. (2021). *Psych: Procedures for psychological, psychometric, and personality research*. Evanston, Illinois: Northwestern University. Retrieved from <https://CRAN.R-project.org/package=psych>

178 Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal*  
179 *of Statistical Software*, 48(2), 1–36. Retrieved from

180 <https://www.jstatsoft.org/v48/i02/>

181 Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). *A 21 word solution*.

182 <https://doi.org/10.2139/ssrn.2160588>

Table 1

*Descriptive statistics of the variables of  
interest*

	mean	sd	median	skew	kurtosis
X1	-0.02	0.95	0.07	-0.01	-0.15
X2	0.06	0.97	0.07	-0.17	-0.25
X3	0.10	0.98	0.07	0.13	-0.32
X4	0.12	1.02	0.07	0.00	-0.26
X5	-0.03	0.95	-0.04	-0.04	-0.28

Table 2

*Intercorrelations of the variables of interest*

	X1	X2	X3	X4	X5
X1	> .99	.30	.39	.30	.34
X2	.30	> .99	.31	.23	.34
X3	.39	.31	> .99	.26	.36
X4	.30	.23	.26	> .99	.30
X5	.34	.34	.36	.30	> .99

Table 3

*Intercorrelations of the variables of interest*

	X1	X2	X3	X4	X5
X1	–	.30	.39	.30	.34
X2	.30	–	.31	.23	.34
X3	.39	.31	–	.26	.36
X4	.30	.23	.26	–	.30
X5	.34	.34	.36	.30	–

*Note.*  $N = 256$

Table 4

*Intercorrelations of the variables of  
interest*

	X1	X2	X3	X4	X5
X1	(.75)	.30	.39	.30	.34
X2	.30	(.83)	.31	.23	.34
X3	.39	.31	(.86)	.26	.36
X4	.30	.23	.26	(.83)	.30
X5	.34	.34	.36	.30	(.81)

*Note.*  $N = 256$

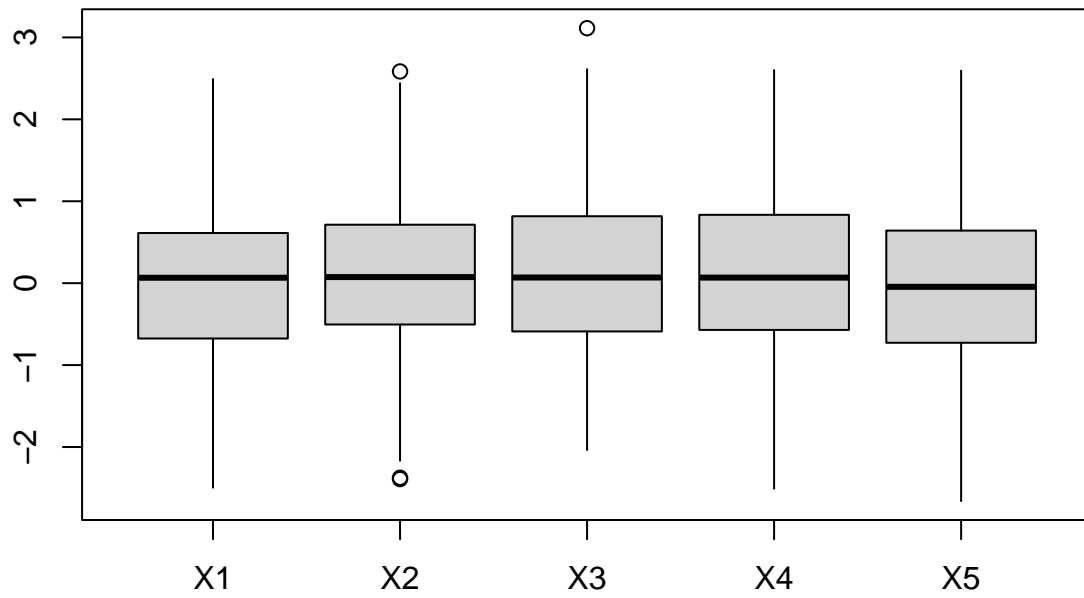


Figure 1. Boxplots of the variables of interest