R Markdown Lesson 04: Using Papaja

Alexander Strobel[1] & Christoph Scheffel[1]

[1] 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-stye manuscript (6th edition) {papaja}* upon which a document similar to this one ist created that contains a header where you can enter the tritle 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

This is the place where you need to provide a statement on how you determined your sample size. This can but need not be based on a power calculation, but in any case you should add a power calculation. In our case, we wanted to have as many participants of our survey as we could get in a given time, but at least $N = 250$ (cf. Schönbrodt & Perugini, 2013). We therefore would write:

We aimed at a sample size of at least $N = 250$. With this sample size, we wanted to make sure to arrive at stable estimates of the correlations obtained in the present study (cf. Schönbrodt & Perugini, 2013) and to have adequate power to detect medium correlations according to the classification of Gignac and Szodorai (2016).

Without knowing the actual sample size eventually achieved, 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). With an $N = 256$, we were able to detect correlations of $r \geq .21$ with a power of $1 - \beta = .80$ at a Bonferroni-corrected significance level of $\alpha' = .05/5 = .01$.

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 aprocedure—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 another *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)].

In the present example, we have five continuous variables X1 to X5 to correlate to each other, will regress X2 on X1 and will run an analysis of variance with Gender as the independent variable and X1 as the dependent variable.

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 make a difference ...

Data analysis

We used R (Version 4.1.1; R Core Team, 2021) and the R-packages *car* (Version 3.0.12; Fox & Weisberg, 2019; Fox, Weisberg, & Price, 2020), *carData* (Version 3.0.4; Fox et al., 2020), *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 variables of interest")
```

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 that, as you may notice, *papaja* not only saves as PDF, but also as PNG (which still may not be sufficient for submission to a journal):

```
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 the `apa.print` function. As an example, your code could look like:

```
model = lm(X2 ~ X1, data = df)
papaja_model = apa_print(model)
apa_table(papaja_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, you do so earlier in your R Markdown file, and your manuscript in an instant will give you the correct results for this analysis. Another great feature of *papaja* is that it automatically formats your results tables according to APA conventions (see Table 5) that was created using the command ‘apa_table(papaja_model)“

Let us try another analysis type: an analysis of variance (ANOVA) using gender as independent variable and X1 as dependent variable. To do so, we must first convert gender into a factor via:

```r
# generate variables for ANOVA
iv = factor(df$gender, levels = c(0, 1), labels = c("female", "male"))
dv = df$X1
# set linear orthogonal contrasts
options(contrasts = c("contr.sum", "contr.poly"))
# run Anova function of the car package to have Type III sum of squares
ANOVA = Anova(lm(dv ~ iv), type = "3")
papaja_ANOVA = apa_print(ANOVA)
if (ANOVA$`Pr(>F)`[2] >= .05) {
  ANOVA_significance = "insignificant"
} else {
  ANOVA_significance = "significant"
}
```

Here, you would write:

An ANOVA with Gender as independent variable and X1 as dependent variable was insignificant, $F(1, 254) = 0.07$, $p = .797$, $\hat{\eta}^2_G = .000$, 90% CI [.000, .001].

As already said: *papaja* formats your analysis output in a way that you can report it fully formatted in your manuscript. It is recommended to read *papaja's* help files to get accustomed with its capabilities for various output formats.

Let us close with a figure and some recommendation on how not only have it saved automatically, but in a submittable format. To do so, we create a figure using R code, but then save it as Encapsulated PostScript which is the format most if not all journals accept. To have your figure appear as you want it to appear, you will need to perform a couple of trials, reduce the character expansion (aka `cex` or `cex.axis`), adjust the `mgp` parameter settings and so on. Also, to have the exact width your journal does accept for, say, a one-column figure of 90 mm width (equaling 90 * 0.0393701 = 3.54 inch), you need to make

some further adjustments. Here is an example (code not visible in this document, but see Figure 2 at the end of the generated file and the figure-latex folder for the code's output):

Discussion

This will be the hard part most of the time, and *papaja* will not help you in writing your discussion. So, for now, let us recapitulate:

Interim Summary

In this lesson, you should have learned how to

- use the the *papaja* template to write dynamic and reproducible scientific manuscripts in APA style
- use placeholders for reporting appropriately formatted statistical results without resorting to data at all
- create tables and figures in APA style in an instance

Exercises

To exercise what you have learned in this lesson:

1. Report the results of a correlation analysis involving variables X1 and X2 in data.frame `df` using `apa_print` with the `cor.test` function.
2. Format a correlation table in a way that prints significant correlations bold-faced.
3. Save Figure 2 in Tagged Image File Format (tiff, another figure format commonly accepted at scientific journals) with 300 dpi resolution.

Outlook

In the next and final lesson, you will learn how to

- collaborate on R Markdown files using *GitHub*
- use a portable and reproducible R environment for your collaborative work using the R package *renv*
- never have to edit paths in your R scripts via the R package *here*

209 **References**

210 American Psychological Association. (1994). *Diagnostic and statistical manual of*
211 *mental disorders* (4th ed.). Washington, D.C.: American Psychological
212 Association.

213 Aust, F., & Barth, M. (2020). *papaja: Prepare reproducible APA journal articles*
214 *with R Markdown*. Retrieved from https://github.com/crsh/papaja

215 Barth, M. (2021). *tinylabels: Lightweight variable labels*. Retrieved from
216 https://github.com/mariusbarth/tinylabels

217 Bertrams, A., & Dickhäuser, O. (2009). Measuring dispositional self-control
218 capacity. A German adaptation of the short form of the Self-Control Scale
219 (SCS-K-D). *Diagnostica*, *55*(1), 2–10. https://doi.org/10.1026/0012-1924.55.1.2

220 Bless, H., Wänke, M., Bohner, G., Fellhauer, R. L., & Schwarz, N. (1994). Need for
221 Cognition: Eine Skala zur Erfassung von Engagement und Freude bei
222 Denkaufgaben [Need for Cognition: A scale measuring engagement and
223 happiness in cognitive tasks]. *Zeitschrift für Sozialpsychologie*, *25*, 147–154.

224 Champely, S. (2020). *Pwr: Basic functions for power analysis*. Retrieved from
225 https://CRAN.R-project.org/package=pwr

226 Fleischhauer, M., Enge, S., Brocke, B., Ullrich, J., Strobel, A., & Strobel, A. (2010).
227 Same or different? Clarifying the relationship of Need for Cognition to
228 personality and intelligence. *Personality & Social Psychology Bulletin*, *36*(1),
229 82–96. https://doi.org/10.1177/0146167209351886

230 Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (Third).
231 Thousand Oaks CA: Sage. Retrieved from
232 https://socialsciences.mcmaster.ca/jfox/Books/Companion/

233 Fox, J., Weisberg, S., & Price, B. (2020). *carData: Companion to applied regression*
234 *data sets*. Retrieved from https://CRAN.R-project.org/package=carData

235 Gignac, G. E., & Szodorai, E. T. (2016). Effect size guidelines for individual

differences researchers. *Personality and Individual Differences*, *102*, 74–78. https://doi.org/10.1016/j.paid.2016.06.069

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

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, *48*(2), 1–36. Retrieved from https://www.jstatsoft.org/v48/i02/

Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*, 609–612. https://doi.org/10.1016/j.jrp.2013.05.009

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2012). *A 21 word solution*. 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. $N = 256$

Table 5

Predictor	b	95% CI	$t(254)$	p
Intercept	0.07	[-0.05, 0.18]	1.15	.250
X1	0.31	[0.19, 0.43]	5.08	$< .001$

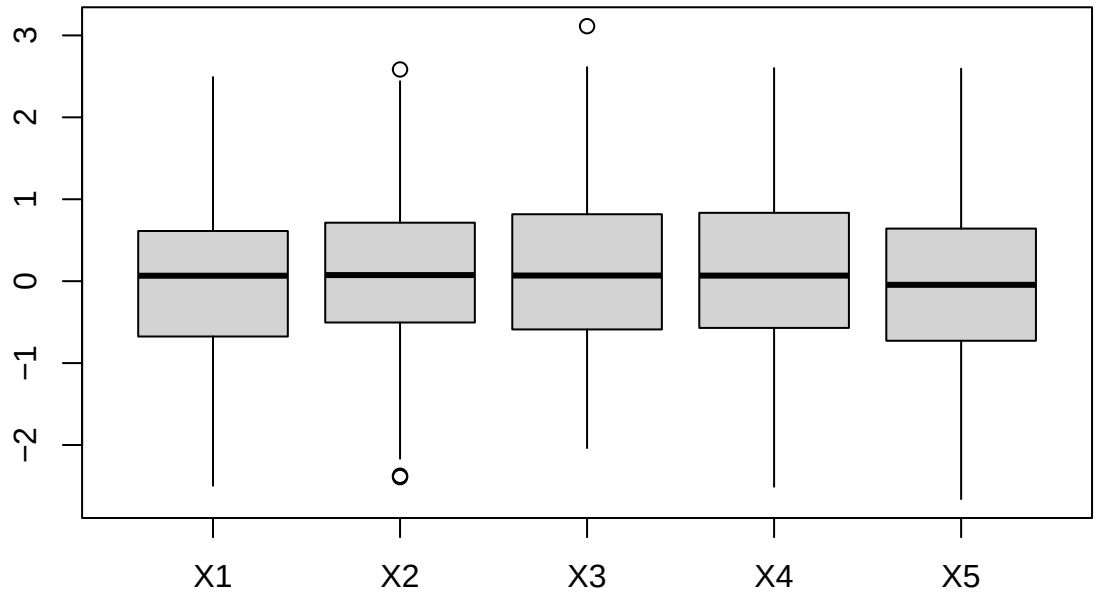

Figure 1. Boxplots of the variables of interest

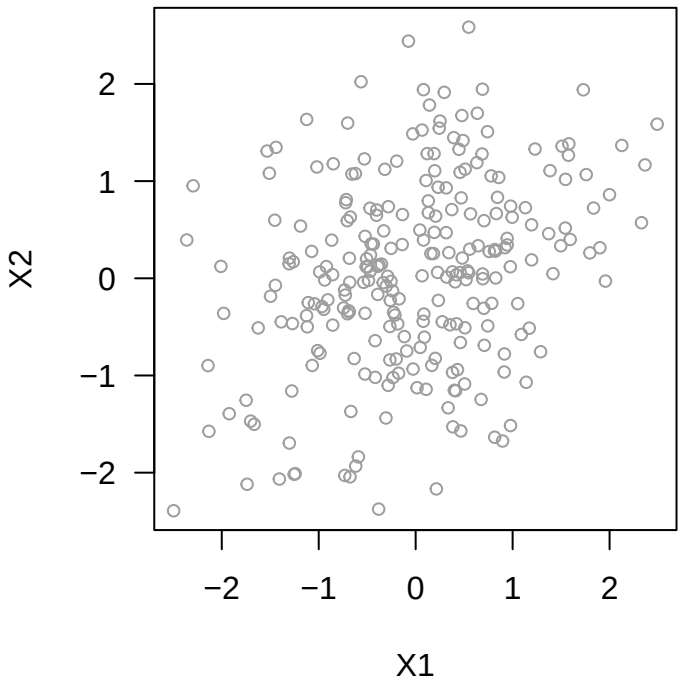

Figure 2. Scatterplot of X1 and X2