Open science data analysis with style: A reproducible repoducible research report

Thomas P. Urbach

Cognitive Science Department

University of California, San Diego

September 5, 2021

Abstract

When the culmination of research is a research report, the culmination of reproducible research must be a reproducible report. To accomplish this, three problems must be solved: 1) the results of the reproducible data analysis must be incorporated into the narrative text, tables, and figures of the document; 2) the document must comply with the byzantine typographical requirements of professional publication style guides and their idiosyncratic modifications by various publishers; 3) the different parts and pieces of the report (manuscript, supplementary information, figures, tables, captions) must be reproducible digital objects in whatever specific document and image file format is required by the online platforms for submission to the journal and production by the publisher. This report describes and demonstrates a flexible and generalizable approach that combines freely available open source data analysis and document preparation software tools to solve these three problems. The report itself is reproducibly generated by the approach it describes and demonstrates for psychologists with real-world examples: the manuscript is formatted in American Psychological Association style and the digital objects are generated as required for the online submission and production platforms used by *Proceedings of the National Academy of Sciences*. The source code is publicly available and may be cloned from the GitHub repository or downloaded from the Open Science Foundation archive and freely modified or adapted for non-commercial purposes under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license. This reproducible report, together with the source code that reproduced it, comprise a complete self-contained tutorial, demonstration, and template for general use.

Open science data analysis with style: A reproducible repoducible research report

Introduction

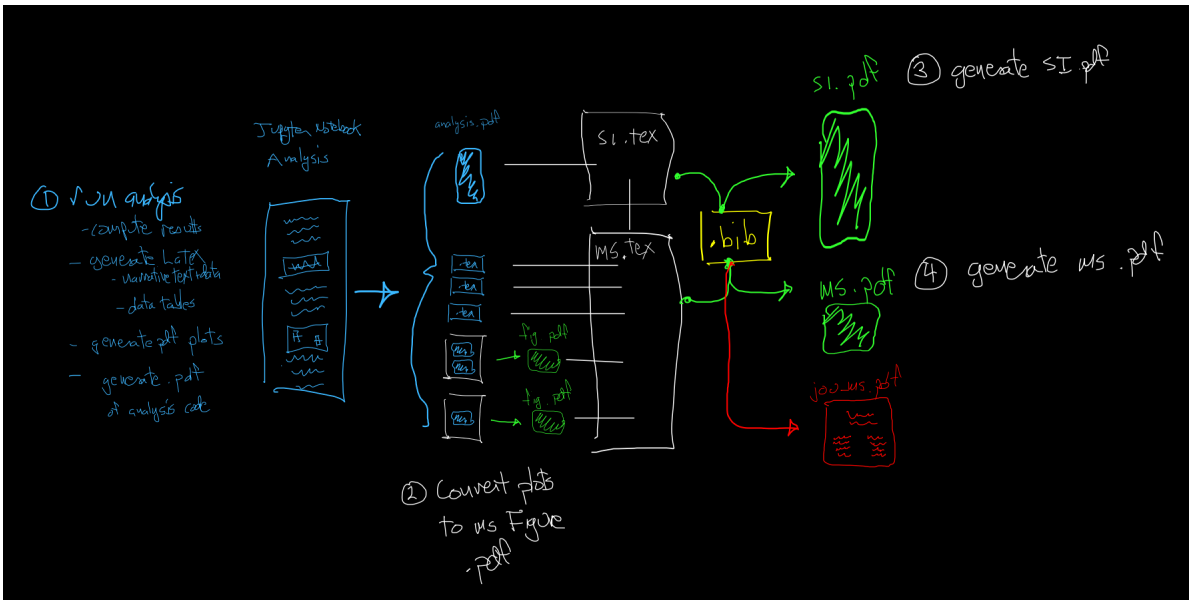
For any research project, after all the work of experimental design, implementation, and data acquisition are in place, and the data analysis is complete, there still remains the task of preparing and publishing the peer-reviewed research report with a clear and accurate presentation of the results through the text, tables, and figures. However, the "research report" is an abstraction; in practice it takes various forms on its trajectory from the authors' desks to dissemination as a journal article in print and online in digital form(s). For the authors, there all the usual chores of document preparation: Writing the narrative text with qualitative and quantitative analysis results, creating high-resolution graphics for figures, preparing tables of data and results, adding and deleting citations and bibliographic references, embedding links to URLs, and aligning cross-references to elements within or across documents, e.g., to the separate online supplementary information. During preparation and revision the report is in flux and must be editable with changes to the text tracked across versions. For pre-print archives and (re-)submission to peer-reviewed journals the text and graphics are composited into a usually un-editable but easily transmissible and viewable digital snapshot, e.g., typically Portable Document Format (PDF). Finally, for journal and book publishers, the process is unwound and the report must be comprised of separate editable text and "camera ready" high-resolution graphics suitable for production in digital form for online viewing and print form. Throughout these transformations for publication, the report must also satisfy specific style requirements and for psychologists this often means a variation of the 6th Edition of the Publication Manual of the American Psychological Association (American Psychological Association, 2010). Or maybe the 7th Edition. In short, as a research report evolves from inception to DOI, it must sometimes change and other times freeze in various highly specific forms and digital file formats as it passes through different hands with different requirements.

When the goal of reproducible research is fully embraced, the "research report" must also be reproducible throughout these stages of preparation, revision, submission, and production. This requires solving three problems: 1) the results of the reproducible data analysis must be incorporated into the narrative text, tables, and figures of the document; 2) the document must comply with the byzantine typographical requirements of professional publication style guides and their idiosyncratic modifications by various publishers; 3) the different parts and pieces of the report (manuscript, supplementary information, figures, tables, captions) must be reproducible digital objects in whatever specific document and image file format is required by the vagaries of an online journal submission platform and then subsequently by a different online production platform. Solutions to each of these problems individually abound, the

challenge is to combine them reproducibly. For instance, reproducible data analyses are becoming commonplace though the use of scientific computing platforms and open source scripting languages like Python and R encapsulated in virtual environments (conda, virtualenv) and containers (Docker, singularity). However the technology for solving the data analysis problem is decoupled from the strict typesetting requirements of different publication styles. On the other hand, mature document preparation software like Microsoft Word and LaTeX provide the fine-grained control of formatting necessary to comply with idiosyncratic style guidelines. However, typing or copy-pasting the results decouples the report from the analysis. The results of the analysis may be reproducible when the analysis is revised by co-authors or reviewers, but the results do not propagate to all the digital objects that comprise the parts and pieces of the report for (re-)submission and production.

This self-reproducing tutorial describes and demonstrates one approach to solving all three problems at once using mature freely available open-source computer software, a working knowledge of LaTeX (LaTeX developers, n.d.), and no more knowledge of computer programming than is already required to implement the reproducible data analysis it reports. The tutorial includes a sample reproducible data analysis pipeline with open-access data but focuses mainly on the reproducible report per se, i.e., solutions to the second and third problem needed to bridge the gap between the end of the reproducible data analysis and the DOI of the peer-reviewed publication in an academic journal. In addition to programmatically combining the data analysis results with the narrative text, tables, and figures of the report, the complete LaTeX source code listings in the Supplementary Materials provide working examples of some features generally useful for manuscript preparation: tracking changes across revisions, preparing camera ready graphics, automating cross-references within and between documents, formatting and masking the citations and bibliography, generating Portable Document Files, compositing documents and pieces of documents in text and PDF file formats, and preparing an author's manuscript for distribution while a published article is embargoed. The Supplementary Information provides instructions for installing the open source software required to reproduce the data analysis and this report. The complete source code for the data analysis and report generation is publicly available and may be downloaded from the Open Science Foundation archive or cloned from the GitHub repository under a Creative Commons CC BY 4.0 license "Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) [software license]," n.d. and used as a template and freely modified for other purposes with appropriate attribution.

Figure 1. Generating a reproducible APA 6th style research report: 1) Executing the reproducible data analysis code generates the complete results which appear as-is in the Supplementary Information. Selected results to be reported in the manuscript are exported to separate files as minimally styled narrative text and tables, and PDF graphics. 2. The graphics exported by the analysis are converted to camera ready APA-style figure graphic PDFs for the manuscript. 3. The Supporting information LaTeX file is typeset as a document PDF which includes the complete analysis source, results, graphics, and document source. 4. The LaTeX manuscript is typeset as a document PDF which includes the results text generated by the data analysis, the camera ready PDF figures, and bibliography.



Method

This approach to generating reproducible research reports requires the four main components, outlined schematically in Figure 1. While the approach is flexible and generalizable, the specific examples are selected for researchers in Psychology and demonstrate how to satisfy all the requirements (except word count) for submitting and publishing a research report in the journal, *Psychological Science*. Accordingly the manuscript is structured with a Cover Page, Abstract, Introduction, Method, Results, and Discussion ("Manuscript Structure, Style, and Content Guidelines," n.d.) and formatted according to the APA 6th edition style (American Psychological Association, 2010). The approach here is readily adapted to the APA 7th Edition with a change of the document class (apa7) and minor modifications to the text described in Supporting Information. The approach generalizes to other publication styles for which LaTeX style files have been defined. A conveniently inventory is collected here: Overleaf.com Templates—Academic Journal. Many styles are community contributions, for instance, arXiv, bioRxiv. A

number of journals and publishers provide official styles, such as Nature, Proceedings of the National Academy of Sciences, eLife and publishers Cambridge University Press, Oxford University Press, Springer including SAGE , the publisher of *Psychological Science*. The Supporting Information for this report provides installation instructions for the necessary software and complete source code listings for the analyses, documents, and figures which are freely available under the CC-BY-4.0 license and may serve as templates for a range of research projects in the psychological sciences.

Data analysis pipeline: `apa_analysis.ipynb`

For demonstration purposes, a toy reproducible data analysis pipeline is implemented in a Jupyter notebook running a Python kernel (Kluyver et al., 2016). The pipeline (down)loads and transforms a sample EEG dataset (Urbach, 2020), computes summary measures, and generates figures and text output. The particulars are incidental, the data may as well be response times and the analysis could be implemented in R, MATLAB, or any language that can format numerical values as strings, write string variables to a text file, and export as PDF, EPS, PNG, JPEG (or a format programmatically convertible to one of these). This PDF is used for vector graphics and PNG for raster graphics in this report since these have proved reliable and both support transparency; EPS and JPEG also work if these are required by the publisher.

Preparing camera ready figures with LaTeX and Ti*k*Z

Ideally, graphic images generated by an analysis pipeline will be in final "camera ready" form but this is not always practical or possible. A figure may require annotations, e.g., math notation, not supported by the figure generator and a multipanel figure may need to combine images from different sources. To demonstrate how this may be done programmatically for reproducibility, three of the "rough" plot graphics generated by the analysis pipeline are reconfigured, annotated and converted into two camera-ready APA-style manuscript figures (Figure 2 and Figure 3) using LaTeX and the Ti*k*Z graphic library without additional software or manual editing.

Manuscript: `apa_ms.tex`

LaTeX is a form of markup language where the document text is intermingled with short typesetting instructions. For instance, *this phrase is typeset in italics*, and the instruction looks like this: `{\it this phrase is typeset in italics}`. Mathematical symbols and more complex equations are very well-supported and set in the same way, e.g., partial eta squared (η_p^2) is set like so: `$\eta_p^2$`. Other instructions are more general. For instance, the manuscript document begins with this,

`\documentclass[man,helv,10pt,draftall,floatsintext]{apa6}`, that says to typeset the document as a manuscript, in Helvetica 10 point font with a draft watermark on all pages, formatted to the APA 6th Edition style except that tables and figures should be placed near where they appear in the text ("floatsintext") rather than collected at the end. This style, including the deviation from the APA 6th table and figure position, corresponds to the submission guidelines for Psychological Science ("Psychological Science 2020 Submission Guidelines," 2020). Like all LaTeX files, the main manuscript file is a plain text document and thus virus-free, portable, viewable, and editable with any text editor, although one that supports LaTeX syntax highlighting on-the-fly syntax error checking is strongly recommended.

Supplementary Information: `apa_si.tex`. Supplementary Information is as much a part of the report as the manuscript and must be likewise reproducible. For demonstration here, the Supplementary Information is comprised of a separate LaTeX file. It provides instructions for downloading this report from public repositories and installing the software to reproduce it. It also includes source code listings of the Makefile used to reproduce portions or all of the analysis, source code and output of the entire executed analysis Jupyter notebook and listings of all the LaTeX files used to generate the report, figures, and supporting information, which includes the self-reflexive listing of the Supporting Information listing itself.

Reproducing the report: `Makefile`

The `make` program is a widely used command line utility for managing the execution of a interdependent computer code in complex programming projects, where changes in one file may might impact some but not all other files. Reproducible data analysis and report generation is similar in that, e.g., generating the camera-ready figure PDFs depends on the rough plots generated by the analysis which in turn depends on executing the analysis. The make utility provides a useful mechanism for expressing the interdependencies and comparmentalizing the project as work progresses, e.g., `make analysis` or `make fig2` or `make ms` while `make all` ensures that all the components execute in the correct order to completely reproduce the analysis and generate all the files and documents for the figures, manuscript, supporting information. Here is a summary of the make file components for generating this report, execution times are for a high performance workstation.

`make analysis` **(45 s)** Reproduce the data analysis by executing all the computer code in the analysis notebook start to finish. This has four side effects:

1. The data analysis computations are executed and the results captured as standard output and plots in the Jupyter notebook cells.

2. Results to be included in the manuscript as narrative text and tables are embedded in text strings, minimally formatted to APA style with LaTeX, and exported as separate text files (.tex).
3. Plots to be included in the manuscript figures are exported as PDF graphics.
4. After execution is complete, a snapshot of the complete notebook—text, computer code, and results captured in the output cells—is exported to a PDF file. The PDF is included in its entirety in the Supplementary Information.

`make fig1` **(1 s)** Run `pdflatex fig1.tex` to convert two rough plot graphics as generated by the analysis pipeline into the camera-ready Figure 2 graphic shown in the manuscript.

`make fig2` **(1 s)** Run `pdflatex fig2.tex` to convert the rough plot graphic generated by the analysis pipeline into the camera-ready Figure 3 graphic shown in the manuscript.

`make figs` **(47 s)** Execute the analysis to generate the rough PDF graphic output files then make fig1 and fig2 as above.

`make ms` **(9 s)** Run `pdflatex apa_ms.tex` to generate the manuscript PDF.

`make si` **(4 s)** Run `pdflatex apa_si.tex` to generate the Supporting Information PDF.

`make all` Run make figs to execute the analysis and generated camera ready figures then make ms and si enough times to update and the cross-references between the manuscript and supplementary information.

Results

The results are this report and the Supplementary Information. Both are reproducibly reproduced using freely available open source software, a working knowledge of LaTeX and no more computer programming than the Python used for the data analysis. A few points merit further discussion.

Discussion

Linking data and arbitrary text

The essential feature of reproducible report generation is linking data from the analysis with the text of the report. Style conventions like APA 6th, 7th and others are strict and varied which means the only general solution is a mechanism for linking the analysis data and results to arbitrary text formatted arbitrarily. This is an old problem, solved long ago by string formatting functions, e.g., `sprintf()` in C,

which reappears in various forms in scripting languages like R, MATLAB, and Python where the f-string function (Python 3.6+) streamlines mixing text and variables.

To illustrate, the same Jupyter notebook that runs the analysis also generates a text file containing the entire contents of the preceding paragraph and this one, including the following sentence that describes the number of trials in each experimental condition. After screening artifacts, the proportion of target trials in the data analyzed was 0.159 (*N* = 239 trials, 201 standards, 38 targets). This narrative description formats the quantitative results in APA 6th style while the values are filled in by the same variables used to compute them. This technique can be used to generate reproducible descriptions of an entire results sections or portions thereof.

The listing below shows the minimally styled LaTeX text generated by the analysis pipeline. For illustration, it includes comments (%%), narrative text with the data values filled in progrmmatically, and `{\it N }`, which italicizes the capital N according to APA 6th style:

```
% These two paragraphs are generated when the analysis is run

The essential feature of reproducible report generation is linking
data from the analysis with the text of the report. Style conventions
like APA 6\textsuperscript{th}, 7\textsuperscript{th} and others are
strict and varied which means the only general solution is a mechanism
for linking the analysis data and results to arbitrary text formatted
arbitrarily.  This is an old problem, solved long ago by string formatting
functions, e.g., \mintinline{c}{sprintf()} in C, which reappears in
various forms in scripting languages like R, MATLAB, and Python where the
f-string function (Python 3.6+) streamlines mixing text and variables.

To illustrate, the same Jupyter notebook that runs the analysis also
generates a text file containing the entire contents of the preceding
paragraph and this one, including the following sentence that describes
the number of trials in each experimental condition.
%%
%% In the next sentence, the Python f-string formatter embeds variables
%% computed during the analysis directly into the generated text which
%% typeset to APA 6th style specifications.
%%
After screening  artifacts, the proportion of target trials in the data
analyzed was 0.159 ({\it N} = 239 trials, 201
standards, 38 targets).
%%
```

```
26  This narrative description formats the quantitative results in APA 6th style
27  while the values are filled in by the same variables used to compute them. This
28  technique can be used to generate reproducible descriptions of an
29  entire results sections or portions thereof.
```

Tables

The ability to link data with arbitrary text is nowhere more valuable than in preparing reproducible data tables styled to editorial standards. The primary challenges are the intricate requirements for laying out headings and notes as illustrated by the following exerpts, drawn from the 40 pages of APA Publication Manual 7th edition table guidelines:

> **headings** Tables may include a variety of headings depending on the nature and arrangement of the data. All tables should include column headings, including a stub heading (heading for the leftmost column). Some tables also include column spanners, decked heads, and table spanners (see Section 7.12)
>
> . . .
>
> **notes:** Thee types of notes (general, specific, and probabiity) appear below the table as needed to describe contents of the table that cannot be understood from the table title or body alone . . .

It is straightforward to reproducibly link table text to the analysis data they tabulate. It is less straightforward, but still tractable to do while also generating the three types of notes, four types of headings and column spanners, and “a border at the top and bottom of the table, beneath column headings (including decked heads), and above column spanners.” (p. 205)

The tabular exhibit labeled Table 1 illustrates a not-quite conforming tabular array of data. When the analysis runs, the table is reproducibly generated as a LaTeX .tex file with one line of code `pandas.DataFrame.to_latex()`. [1] The .tex file is imported into the manuscript the same way as the arbitrary text file above.

This approach is simple and easy and well-suited for data tables presented in supporting information where styling requirements are typically less strict. When easily generated tables will not do, the fall back is arbitrary text generation. A few lines of Python code and common string formatting methods suffice to

[1] For analyses scripted in R, the `xtable` library similarly generates LaTeX format table from dataframes https://cran.r-project.org/web/packages/xtable/index.html.

Table 1

A non-APA Style data table and note generated as LaTeX by calling `pandas.DataFrame.to_latex()`.

tone stimulus	hi	lo	All
standard	107	94	201
target	14	24	38
All	121	118	239

Note: Python variables are conventionally lower case.

generate the LaTeX required to format the table header, footer, notes and row data to APA style. The following listing shows the programmatically generated LaTeX, the result is shown as Table 2. The Python source code to is Jupyter notebook in the Supporting Information.

```
\begin{tabular}{llll}
\toprule
 &  \multicolumn{2}{c}{Tone} & \\
\cmidrule{2-3}
 & Hi & Lo & All \\
\midrule
Standard & 107 & 94 & 201 \\
Target & 14 & 24 & 38 \\
All & 121 & 118 & 239\\
\bottomrule
\end{tabular}
```

Figures

Graphics figures in PNG, PDF, and JPEG can be included in a LaTeX document with the `command`. Of these PDF seems to be the most reliable for vector graphics (plots, line drawings, charts, plots) and PNG for raster graphics. Including figures is straightforward, creating figures for a data analysis reproducibly is another matter. In some case it may be possible to generate camera-ready graphics from the data anlysis pipeline itself. Although this takes some effort to fine tune at the outset when Reviewer 2 insists on some mid-stream revision that requires re-running the analysis, the change propagates all the way through to the final figures included in the report. However this is not always possible. One recourse is to use an

Table 2

An APA style data table and note generated as LaTeX with a few lines of pure Python.

	Tone		
	Hi	Lo	All
Standard	107	94	201
Target	14	24	38
All	121	118	239

Note: APA Style capitalization.

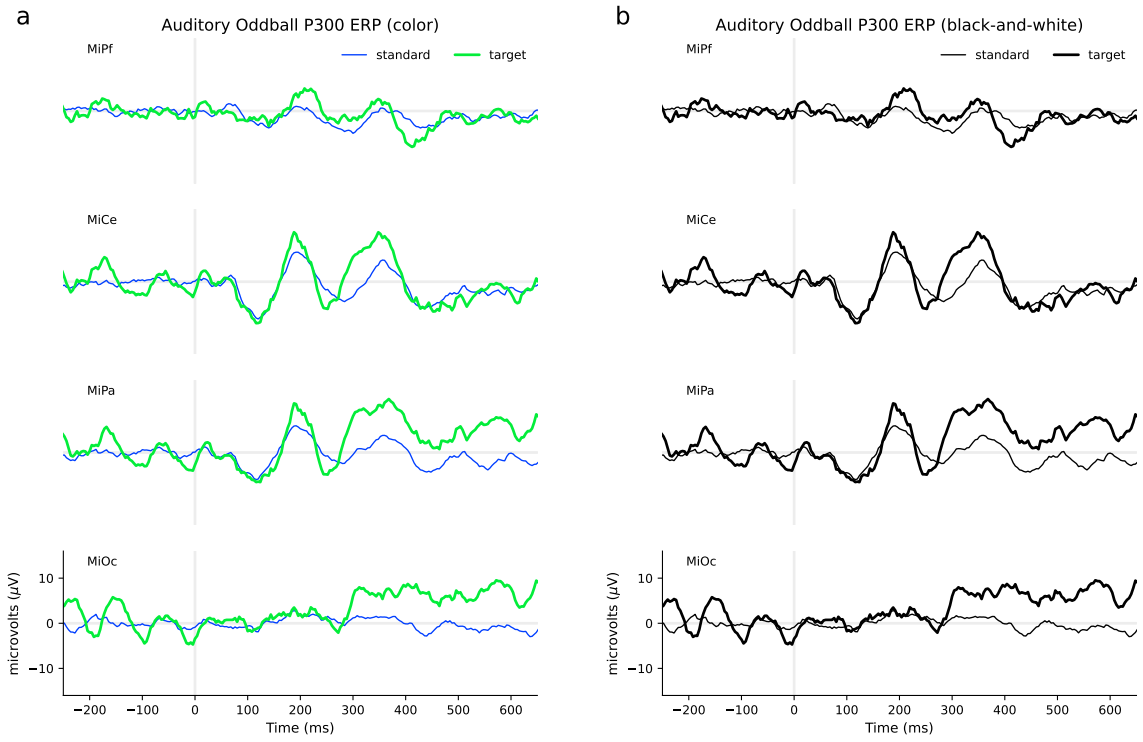
interactive vector graphics manipulation programs like Inkscape to import the graphic and edit to style but, like manually typing results into a data table, the results may change but the representation of the results does not.

Since hand editing figures amounts to using a mouse to select a sequence of drawing commands, it can be done programmatically with the right vector graphics manipulation tools. In the LaTeX ecosystem, a particularly powerful package for this is Ti*k*Z and the learning curve is correspondingly steep. However, for simple tasks like laying out and annotating the figures, it is reasonably straightforward. The tikz figure is a canvas with coordinates. Graphics can be placed and aligned, and drawing elements like lines, arrows, and shading added. Figure 2 and Figure 3 are worked examples of this approach and show how to convert graphics generated by the data analysis into "camera ready" figures to APAstyle specifications saved as separate PDF files for upload to the publisher. Figure 2 is a simple example that lays out two graphics side by side and Figure 3 illustrates a more elaborate example that selects portions of a single graphic, rearranges and resizes them and adds additional graphic and text annotations. The LaTeX and Ti*k*Z code for both figures is listed in the Supplemental Information.

Citations, masked citations, and references

In LaTeX citations in the text are indicated by typing commands like `\cite{}` with the author, name, year, parenthesis information for APA style are determined when the document is typeset. Typing the citation commands amounts to "cite-while-you-write". LaTeX automatically generates a bibliography in the APA style from the corresponding .bib file (bibliography database) according to the citations that appear in the text. There lots of options for citation format, see the `biblatex` and `apa6` docs for reference. For instance, the `\parencite` command generates a formatted citation in parentheses (Lamport, 1986). The cite command generates one without parentheses, as in Lamport, 1986. When manuscript submission

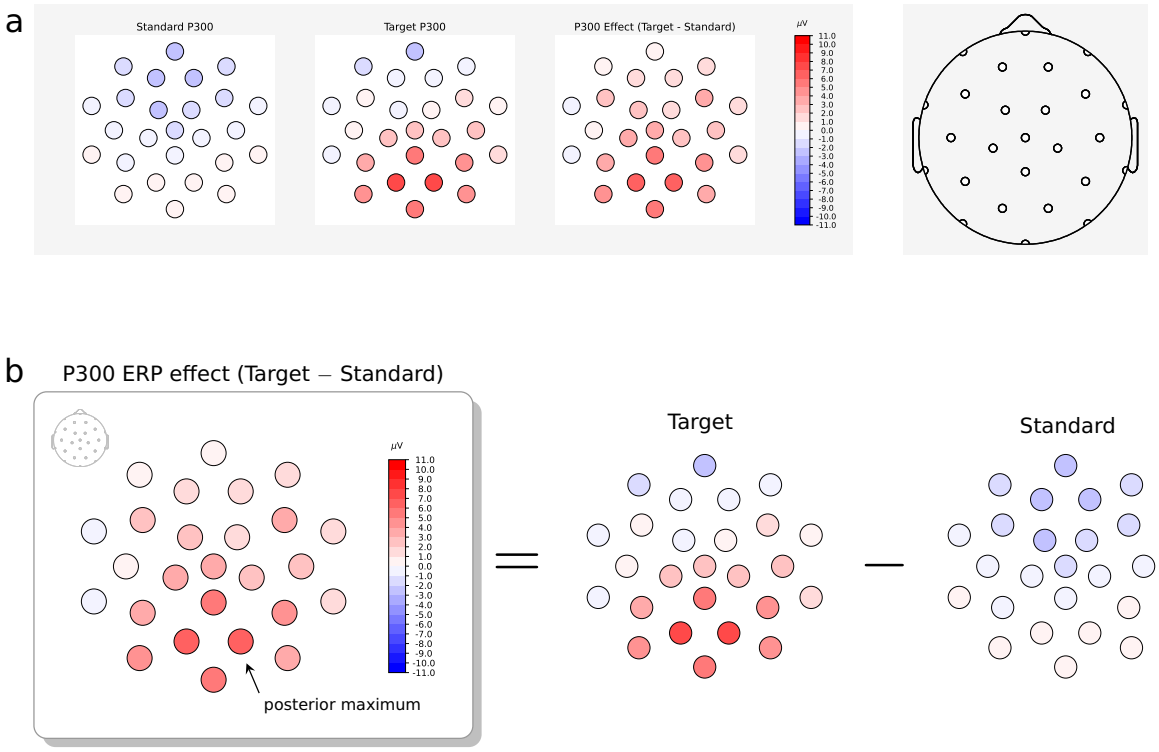
Figure 2. A complete multi-panel color figure generated reproducibly from the data to Psychological Science figure specifications. The figure is generated using the matplotlib package in Jupyter Notebook running a Python kernel. The code illustrates some useful Python idioms and matplotlib functionality including style sheets, the style context manager, how to lay out panels, add labels including with mathematical symbols, and export the figure as as a PDF graphic.



requires citation masking for blind review, the masked variants of the citation commands, e.g., `\maskparencite` can be used: (Lamport, 1986). The masked citations are indictaed in bold when the manuscript is typeset normally and replaced with *(1 citation removed for masked review)* when typeset with the mask option.

The .bib file is a text file with bibliography entries that have the usual author, title, data, publisher, fields, and a great many others, in a specific format. There are several options for where to get the .bib file. Scientific literature search engines, publisher websites routinely export citations in .bib format which can be copy-pasted instead of tediously typed. If a reference manager is already being used, it may also be able to export its references to .bib format. And there are a number of reference managers that are designed from the ground up to use .bib. As of this writing, the open-source JabRef seems to have emerged as pick of the litter, being fully featured enough to support general use and working across

Figure 3. Reproducible figure layout and annotation. Panel a shows the pdf as generated by the analysis script and a stock montage image. Panel b shows the "camera ready" figure output generated by post-processing the generated graphic with LaTeX and the Ti*k*Z drawing library as part of the documentation generation pipeline. The data are the same as in Figure 2



platforms. BibDesk is another option but only runs on OSX. If other options fail, the entry can be typed.

Cross references

To cross-reference between elements like tables, figures, and sections LaTeX links them via `\label` `\ref` pairs. However a more general approach is to use the zref package which links elements with `\zlabel` `\zref` pairs that work across documents which the built-in version does not. This is particularly useful for cross-referencing information in the Supplementary Information from the main manuscript and vice version. When there are two or more docs and a series of figures and/or tables and/or document sections in each and have to add or delete another, it is mighty handy to have the references everywhere in both documents automagically update the numbering and page locations. Here is an example cross reference a section in the Supporting Information, if that section title changes so does this reference: Source: author_analysis.ipynb. To cross-reference between .tex documents, both documents must be compiled and this may not be possible in all online submission systems, even those that accept

.tex format documents. For instance, the PNAS online submission system accepts latex for manuscripts but requires .pdf for supporting information and does not accept uploads of the auxiliary files required by zrefs in the main manuscript which means the submission system cannot correctly compile .tex manuscripts with zrefs.

Tracking changes

Revisions to a document marked and tracked in a document in the same way as other types of formatting. With the `\changes` package, authors indicate the type of change or markup, e.g., add, delete, replace, highlight, and then bracket the relevant text, like so: `\added[id=TPU]{Here is some new text}`. When the document is type typeset in draft mode: (`\usepackage[draft]{changes}`), the changes are highlighted and tagged by author. For instance This text is marked by TPU as added[TPU] and ~~this text is marked by ABC as deleted~~[ABC]. Furthermore, this text is marked by TPU as highlighted and this is XYZ's replacement text~~this text was replaced~~[XYZ].

[TPU 1] is this helpful?

In draft mode, a list of the changes can be generated by inserting the `\listofchanges` command, typically at the beginning or end, though shown here at the end of this section for illustration. Collaborators can review the changes in the pdf and add make further revisions to the .tex document. When the document is typeset for the final version (`\usepackage[final]{changes}`), the changes are applied and remaining comments, markup, and annotations stripped, similar to accepting tracked changes in a WSYSIG document. The draft and final versions may both be useful when resubmission of a document following revision requires both "clean" version with the changes made and a draft version marked up to indicate where the revisions were made. For cases where there are two versions of a .tex document and the changes are not explicitly marked up inline, the command line utility program `latexdiff` can be used to automatically generate a single pdf with the differences between the versions indicated as in changes. Both of these features are best suited to marking revisions and changes in the text of relative similar documents and are not well-suited to track massive restructuring or revisions to figures and tables. Here is the list of changes explicitly marked up in the previous paragraph.

List of changes

Compositing documents: files and file formats

Various files and formats are required go submit and publish a research report. These may include a main editable manuscript (document), supporting information (document, data), figures (vector and raster image graphics files), tables, and bibilographic info. Journals and publishers have divergent interests (readability for evaulation in review vs. production for print and digitial formats) and (thus) different requirements for document preparation. This is further complicated by open-access policies that require authors to deposit a final pre-publication manuscript if the publisher won't (but most do, eventually). For submission to Psychological Science for instance, the file formats are LaTeX(.tex) for editable text and Portable Document Format (.pdf) for graphics, a vector format that is scalable without loss of resolution. To submit the report to the journal for review the .tex and .pdf graphic files composited into a single .pdf file and all files uploaded "APS Figure Format and Style Guidelines," 2013; "Psychological Science 2020 Submission Guidelines," 2020. Whereas the journal submission portal requires the a single composited document with text and graphics all in one, the publisher's portal requires the separate editable text and graphics files, i.e., the .tex and graphics .pdfs.

Working with LaTeX simplifies some aspects of this by allowing files in different digital formats to be included in documents in various ways. As illustrated by linking results and abitrary text for narrative descriptions and tables, separate files of LaTeX can be inserted directly into the document as if typed in place. This allows the tables to be reproducibly prepared as separate files (as required by some publishers) and also incorporated in exactly the same form in the body of the manuscript (as also required by these publishers). The same holds for the camera ready graphics for Figure 2 and 3 which are also separate files included as-is in the mansucript. Additionally the `\includepdf` package, allows all or selected pages of a multi-page PDF documents to be included in a LaTeX as demonstrated in by the Supplementary Informatinon that includes the entire PDF of the fully executed data analysis Jupyter Notebook. Finally, the `\minted` package used extensively throughout this document will import the contents of separte files into the LaTeX document and also highlight the code according to the syntax of the specfic language, e.g., Python, R, LaTeX which is of great value in documenting scripted reproducible research pipelines. The Supplemental Information demonstrates this by importing and highlighting all the LaTeX files used in the producition and reproduction of this tutorial report.

Author manuscripts

Whereas journals may require submission as a double spaced manuscript, the published articles typeset single space in two columns with figures and tables where they belong are generally easier to read. Switching the `documentclass` option from man (manuscript) to jou (journal) typesets the document in

a more-nearly-journal-like format (Figure 4), which may be useful for distributing working drafts or post-publication author manuscripts during a publisher's embargo period.

Figure 4. Example of typesetting this document with the jou option

Reproducible research reports in APA Publication Manual 6th (or 7th) Edition Style with LaTeX and open-source software

Thomas P. Urbach
Cognitive Science Department
University of California, San Diego
August 23, 2020

The last step of reproducible research is the reproducible research report. However in order to publish the report, it must be prepared in various parts and pieces and digital formats, and formatted to the intricate requirements of a specific publication style such as the APA 6th or 7th Edition. Although mature document preparation software such as Microsoft Word and LaTeX streamline many aspects of manuscript preparation and provide the necessary fine-grained control of layout and formatting, typing the analysis results into the document decouples the results and the report. The results may be is reproducible but the report is not. This tutorial describes and demonstrates how the resources already in hand for a typical open-source data analysis pipeline, in combination with a simple text editor and LaTeX, can be recruited to compose, edit, and generate a fully reproducible APA 6th edition manuscript, complete with all the files in all the formats necessary for initial submission (journal), revision and re-submission (journal), and publication (publisher). In addition to reproducing its own results in APA-styled narrative text, tables, and figures, the complete LaTeX source code listings in the Supplementary Materials are a practical guide to how some features generally useful for manuscript preparation may be used in LaTeX: tracking changes across revisions, preparing camera ready graphics, automating cross-references within and between documents, formatting and masking the citations and bibliography, generating Portable Document Files, compositing documents and pieces of documents in text and PDF file formats, and preparing an author's manuscript for distribution while a published article is embargoed. The Supplementary Information provides instructions for installing the open source software necessary to reproduce this document and the Supplementary Information along with a complete listing of all the data analysis code and documents which may serve as templates adaptable for a range of research projects. All files are publicly available for download from github and OSF.

Introduction

The last step of reproducible research is a reproducible research report with a clear and accurate presentation of the results through the text, tables, and figures. However, the "research report" is an abstraction; in practice it takes various forms on its trajectory from the authors' desks to dissemination as a journal article in print and online in digital form(s). For the authors, there all the usual chores of document preparation: Writing the text, creating high-resolution graphics for figures, preparing tables of data and results, adding and deleting citations and bibliographic references, embedding links to URLs, and aligning cross-references to elements within or across documents, e.g., to the separate online supplementary information. During preparation and revision the report is in flux and must be editable with changes to the text tracked across versions. For pre-print archives and (re-)submission to peer-reviewed journals the text and graphics are composited into a usually un-editable but easily transmissible and viewable digital snapshot, e.g., typically Portable Document Format (PDF). Finally, for journal and book publishers, the process is unwound and the report must be comprised of separate editable text and "camera ready" high-resolution graphics suitable for production in digital form for online viewing and print form. Throughout these transformations for publication, the report must also satisfy specific style requirements and for psychologists this often means a version of the 6th or 7th Edition of Publication Manual of the American Psychological Association. In short, as a research report evoles from inception DOI, it must sometimes change and other times freeze in various highly specific forms, formats, and pieces and digital file formts as it passes through different hands with different requirements.

This tutorial report describes and is itself a complete fully functional worked example of one approach to meeting all of these requirements using mature freely available community supported open-source computer software, a working knowledge of LaTeX, and no more knowledge of computer programming than is already required to implement the re-

Conclusion

There are many ways to prepare a research report but far fewer to do so reproducibly while at the same time satisfying the requirements of publication styles and online journal submission and production platforms. This report illustrates one approach that does so and dovetails with best practices in open science data analysis. Once a reproducible analysis in place, the additional cost of the reproducible report is acquiring a working knowledge of LaTeX and if necessary TiZ.

References

American Psychological Association. (2010). *Publication manual of the american psychological association* (6th). American Psychological Association.

APS Figure Format and Style Guidelines. (2013). Web Page. Association for Psychological Science. Retrieved August 11, 2020, from https://www.psychologicalscience.org/publications/aps-figure-format-style-guidelines

Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) [software license]. (n.d.). Retrieved from https://creativecommons.org/licenses/by-nc-sa/4.0/legalcode

LaTeX developers. (n.d.). LaTeX — a docoument preparation system [software]. Retrieved from https://www.latex-project.org/

Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B. E., Bussonnier, M., Frederic, J., ... Corlay, S., et al. (2016). Jupyter notebooks-a publishing format for reproducible computational workflows. In F. Loizides & B. Schmidt (Eds.), *Positioning and power in academic publishing: Players, agents and agendas* (Vol. 2016). doi:https://doi.org/10.3233/10.3233/978-1-61499-649-1-87

Lamport, L. A. (1986). The gnats and gnus document preparation system. *G-Animal's Journal*, *41*(7), 73+.

Manuscript Structure, Style, and Content Guidelines. (n.d.). Web Page. Association for Psychological Science. Retrieved August 11, 2020, from https://www.psychologicalscience.org/publications/ms-structure-guidelines

Psychological science 2020 submission guidelines. (2020). Retrieved August 11, 2020, from https://www.psychologicalscience.org/publications/psychological_science/ps-submissions

Urbach, T. P. (2020). Eeg-workshops/mkpy_data_examples/data [data set]. doi:10.5281/zenodo.4099632