Reproducible Research

Christopher Gandrud

18 March 2016

SG1022, City University London

Table of contents

- **What** is research?
- **What** is reproducible research?
- **Why** do we need reproducible research
- **How** to do really reproducible research

What is research?

What is research?

Is this research?

It's (Change in) the (Future) Economy, Stupid: Economic Indicators, the Media, and Public Opinion

Stuart N. Soroka University of Michigan
Dominik A. Stecula University of British Columbia
Christopher Wlezien University of Texas at Austin

Economic perceptions affect policy preferences and government support. It thus matters that these perceptions are driven by factors other than the economy, including media coverage. We nevertheless know little about how media reflect economic trends, and whether they influence (or are influenced by) public economic perceptions. This article explores the economy, media, and public opinion, focusing in particular on whether media coverage and the public react to changes in or levels of economic activity, and the past, present, or future economy. Analyses rely on content-analytic data drawn from 30,000 news stories over 30 years in the United States. Results indicate that coverage reflects change in the future economy, and that this both influences and is influenced by public evaluations. These patterns make more understandable the somewhat surprising finding of positive coverage and public assessments in the midst of the Great Recession. They also may help explain previous findings in political behavior.

A growing body of work demonstrates a link between economic conditions and both attitudes about government policy and preferences for spending (e.g., Durr 1993; Erikson, MacKuen, and Stimson 2002; Soroka and Wlezien 2010; Stevenson 2001; Wlezien 1995). There also are vast literatures exploring the degree to which support for governments and leaders follows economic trends. Some work focuses on economic conditions and assessments of presidential/government performance and voting (e.g., Bartels and Zaller 2001; Campbell 1996; Clarke and Stewart 1995; Erikson 1989; Happy 1986; Hibbs 1987; MacKuen, Erikson, and Stimson 1992; Nadeau, Niemi, and Amato 1994, 1996; Nadeau et al. 1999; Price and Sanders 1993; Sanders 1996, 1999; Sanders, Marsh, and Ward 1993; for reviews, see Lewis-Beck and Stegmaier 2000, 2007); a related body of research uses key economic variables to predict—with a good degree of success—the outcome of elections, both in the United States and elsewhere (e.g., Abramowitz 1988; Erikson and Wlezien 2012; Lewis-Beck 1988b; Wlezien and Erikson 1996; for recent reviews, see Kayser and Wlezien 2011; Linn, Nagler, and Morales 2010). There is, in sum, a considerable body of evidence highlighting the political importance of economic conditions.

Public *perceptions* of the economy matter as well. In fact, past work suggests that economic perceptions influence vote intentions and government evaluations above and beyond the impact of the actual economy (e.g., Nadeau, Niemi, and Amato 1994). The sources of

Stuart N. Soroka is Professor of Communication Studies and Political Science, and Faculty Associate at the Center for Political Studies, Institute for Social Research, University of Michigan, 5370 North Quad, 105 South State Street, Ann Arbor, MI 48109-1285 (ssoroka@umich.edu). Dominik A. Stecula is a Ph.D. student, Department of Political Science, University of British Columbia, C425 – 1866 Main Mall, Vancouver, British Columbia V6T 1Z1 (d.stecula@alumni.ubc.ca). Christopher Wlezien is Hogg Professor of Government, Department of Government, University of Texas at Austin, Batts Hall, Austin, TX 78712-1704 (wlezien@austin.utexas.edu).

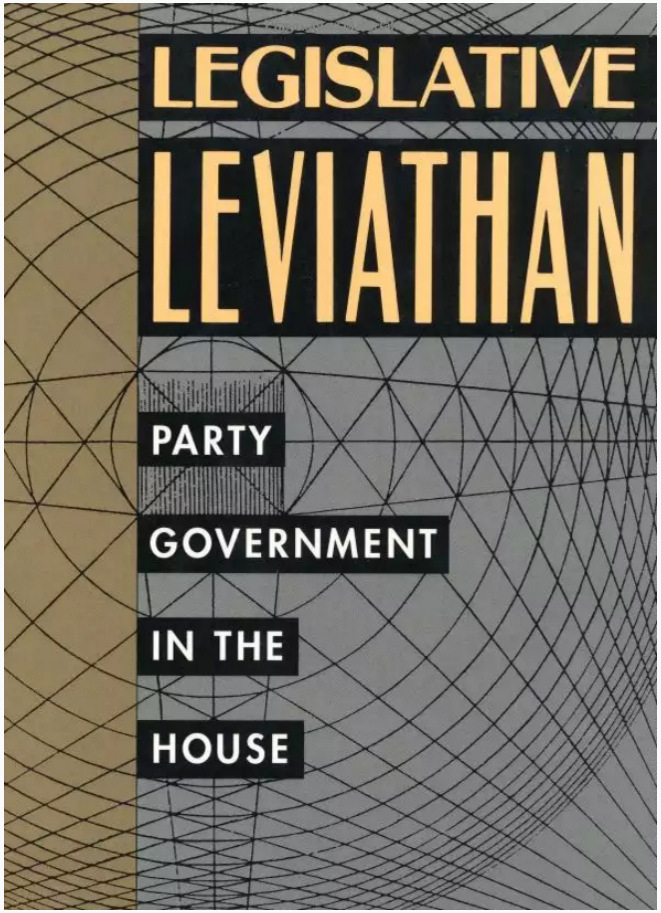
Previous versions of this article were presented at the 68th Annual Conference of the American Association for Public Opinion Research, Boston, May 2013, and the 71st Annual Conference of the Midwest Political Science Association, Chicago, April 2013. We are grateful to Stephen Farnsworth and J. Scott Matthews for helpful advice along the way, to Brian J. Fogarty for comments on a previous draft, to Ataman Ozyildirim for help with Conference Board indicators, and to Lori Young for her work on the Lexicoder Sentiment Dictionary. Partial support for this research was provided by the Fonds de recherche sur la société et la culture, Quebec, Canada. The data used in this study, as well as a do-file replicating all the results, are available directly from the authors at snsoroka.com, as well as in the AJPS Data Archive on Dataverse (http://dvn.iq.harvard.edu/dvn/dv/ajps).

American Journal of Political Science, Vol. 59, No. 2, April 2015, Pp. 457–474

©2014, Midwest Political Science Association
DOI: 10.1111/ajps.12145
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

What is research?

Is this research?



Are they research?

No

Papers, articles, slideshows, talks, books are **advertising, not research**.

Are they research?

What are they?

Presentation documents *announce* select findings and *try* to convince an audience that they are correct (Mesirov 2010).

What is research?

Quantitative social science research involves the **procedures** and **choices** researchers make to gather data, process it, and analysis it in order to address their research questions.

For computational research, this includes "the full software environment, code, and data that produced the results" (Donoho 2010, 3015).

We need to make our research available, not just the advertising!

What is reproducible research?

Replicability

If we make the research available, then it will be more likely that other researchers can **replicate** our work.

Replicability

Replicable Research

When there is *sufficient information* available for *independent researchers* to make the *same findings*, using the *same procedures* with *new data*.

Replicability

No one study proves anything.

Instead, a individual study **contributes** to a cumulative **body of evidence**.

Replication is an important part of **strengthening** this body of evidence.

However...

Sometimes full replications **are not feasible** because:

- *limited resources* for gathering new data (e.g. very expensive to build another Large Hadron Collider),

- the original research already *sampled the universe* of cases.

So...

Reproducibility

Reproducible Research

When there is sufficient information available for independent researchers to make the same findings, using the same procedures with the *same data*.

Reproducibility

Really Reproducible Computational Research

"... the **data and code** used to make a finding are available and they are sufficient for an independent researcher to recreate the finding" (Peng 2011, 1226)

Reproducible and Replicable

Reproducible research **enhances** replicability.

- Reproducible research is a precondition for replicable research.
- Reproducibility is a 'second best' if attempting a replication is not possible.
- If it is **easy** to reproduce your work, then it is more likely that someone else will be able to **replicate** it.

Why reproducible research?

For you

Better work habits

- If you are making your research reproducible from the start you are more clearly **documenting** and **organising** your work.

- So, you will be more likely to **remember** what you did in the future! Bowers (2011, 2) describes this as making a "*better relationship with your future self*".
- So, you are **less likely to make errors** and you are more likely to **find** and **fix** the errors you do make.

We **all make mistakes** during all stages of our research!

Really reproducible research

Instead of pretending like mistakes don't happen, we should **have procedures** that help us **minimise** our errors and allow us (and others) to **find** and **correct** the errors we do make.

Really reproducible research is an **important part** of this process.

If your data gathering and analysis process–your data pipeline–is really reproducible, then it is possible to actually **find your mistakes and correct them**.

For you

Better teamwork

- If you are making your work reproducible for independent researchers, then your work will be **easier for your teammates to understand** and **collaborate** on with you.

For science

A core tenant of science: Scientific conclusions that are **not replicable** should be **abandoned or modified** "when confronted with more complete or reliable . . . evidence"

APS `http://www.aps.org/policy/statements/99_6.cfm`

Reproducible and Replicable

Important!

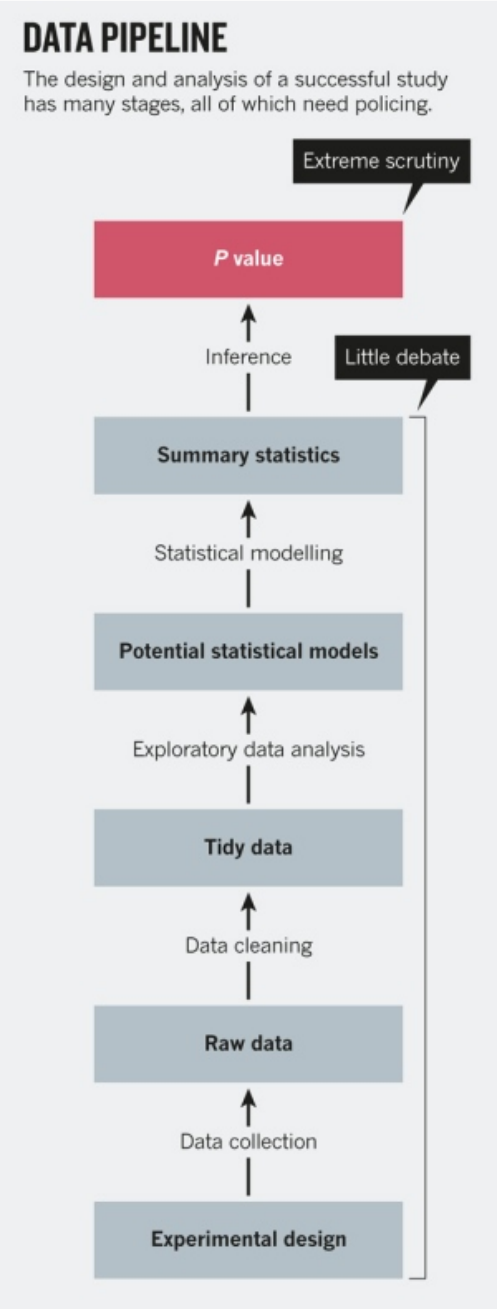
"**A study can be reproducible and still be wrong**" (Peng 2014)

E.g. a finding that is statistically significant in one study may remain statistically significant when reproduced using the original data/code, but replication studies are unable to find a similar result.

The original finding could just have been **noise** or caused by some **data processing error** and so on.

Scientific Scrutiny and the Data Pipeline

Largely **until recently**, most of the evaluation of social science research focused on the **final statistical** presentation of the results. Less attention was paid to the **data pipeline**.


DATA PIPELINE
The design and analysis of a successful study has many stages, all of which need policing.
Extreme scrutiny
P value
Inference
Little debate
Summary statistics
Statistical modelling
Potential statistical models
Exploratory data analysis
Tidy data
Data cleaning
Raw data
Data collection
Experimental design

Peng (2015)

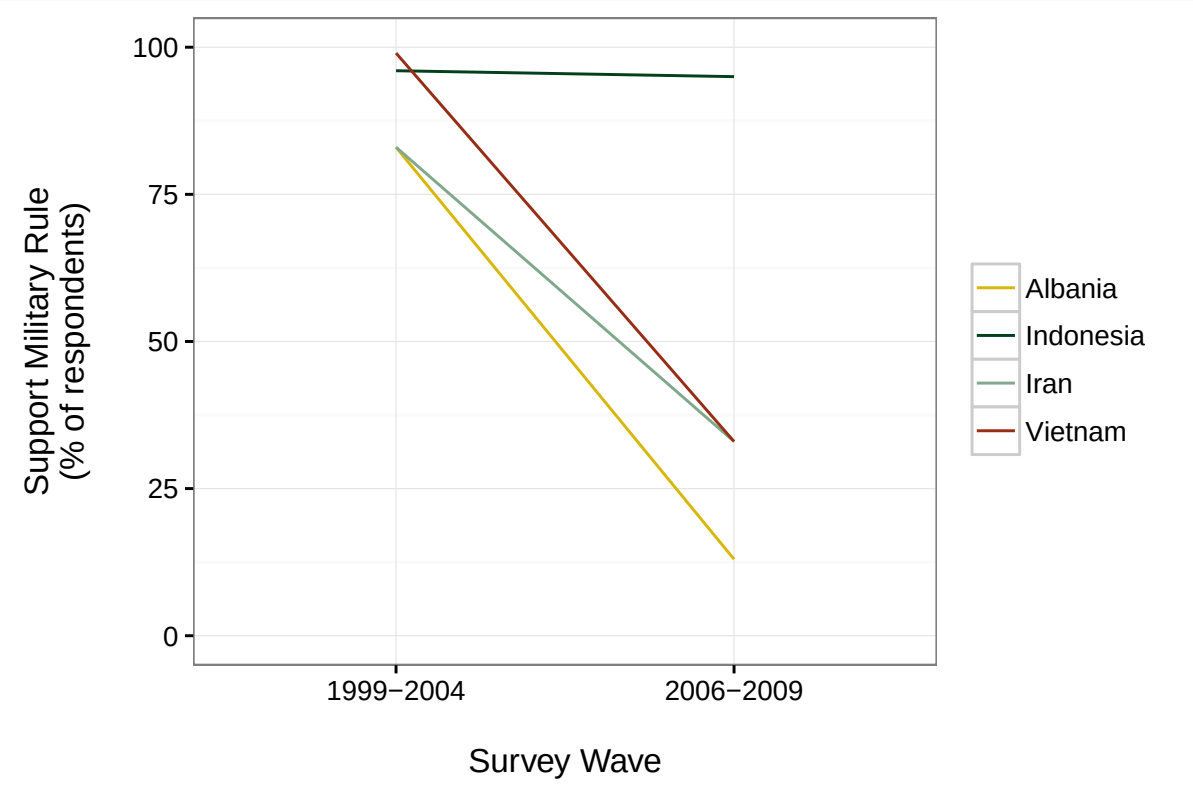
Scientific Scrutiny and the Data Pipeline

Making your data pipeline really reproducible helps others to more **closely evaluate your research claims**.

Example: World Values Survey

Background: the World Values Survey is a large, repeated cross-national survey of political and social values.

Original research finding:



Why did support for military rule decline so much in Albania, Iran and Vietnam over only a few years?

Possible explanation

Inglehart and Welzel (2005) argue that: “If younger generations are socialized under significantly different conditions from those that shaped earlier generations, the values of the entire society will gradually change through intergenerational replacement.”

More likely...

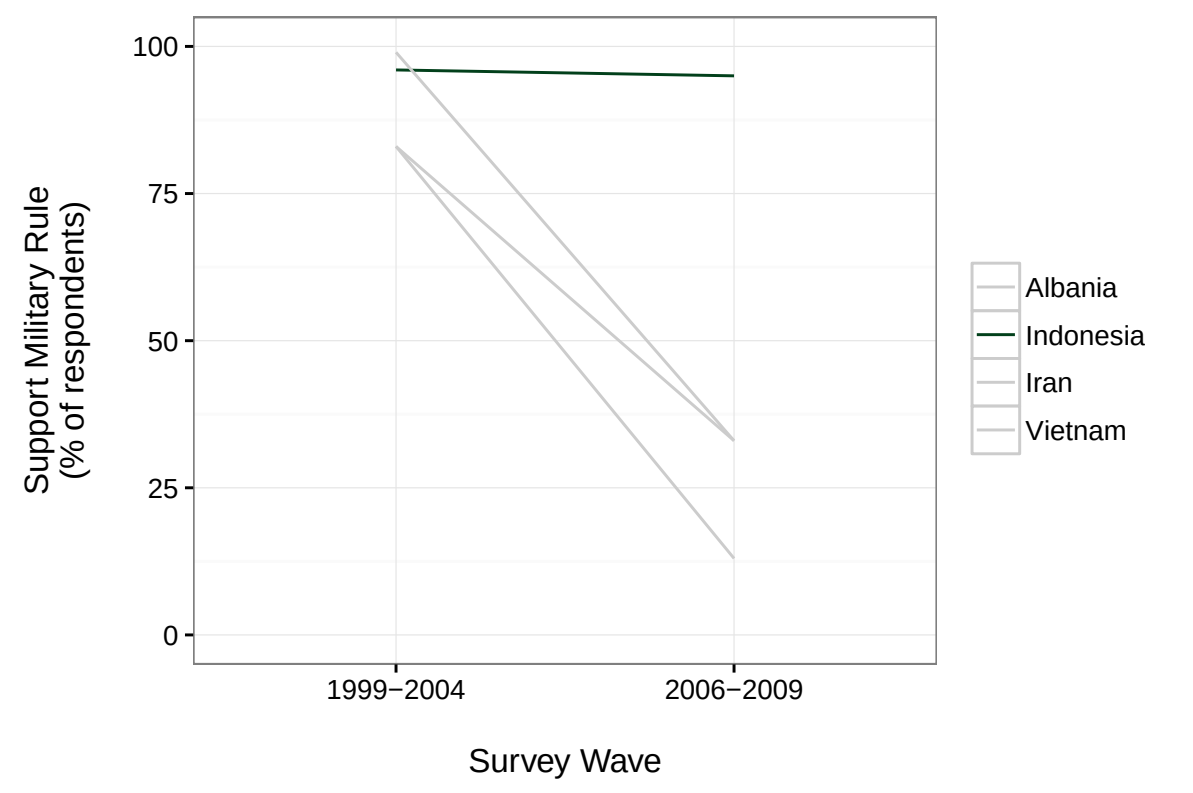
Kurzman (2014) argues that it is more likely that the **question translation changed**.

	1999-2005	2006-2009
English	having the military rule	having the military rule

More likely...

Kurzman (2014) argues that it is more likely that the **question translation changed**.

	1999-2005	2006-2009
English	having the military rule	having the military rule
Indonesia	Memiliki peraturan yang jelas tentang angkatan bersenjata (**having clear military rules**)	Memiliki peraturan tentang angkatan bersenjata (**having military rules**)


Support Military Rule
(% of respondents)
100
75
50
25
0
1999−2004
2006−2009
Survey Wave
Albania
Indonesia
Iran
Vietnam

More likely...

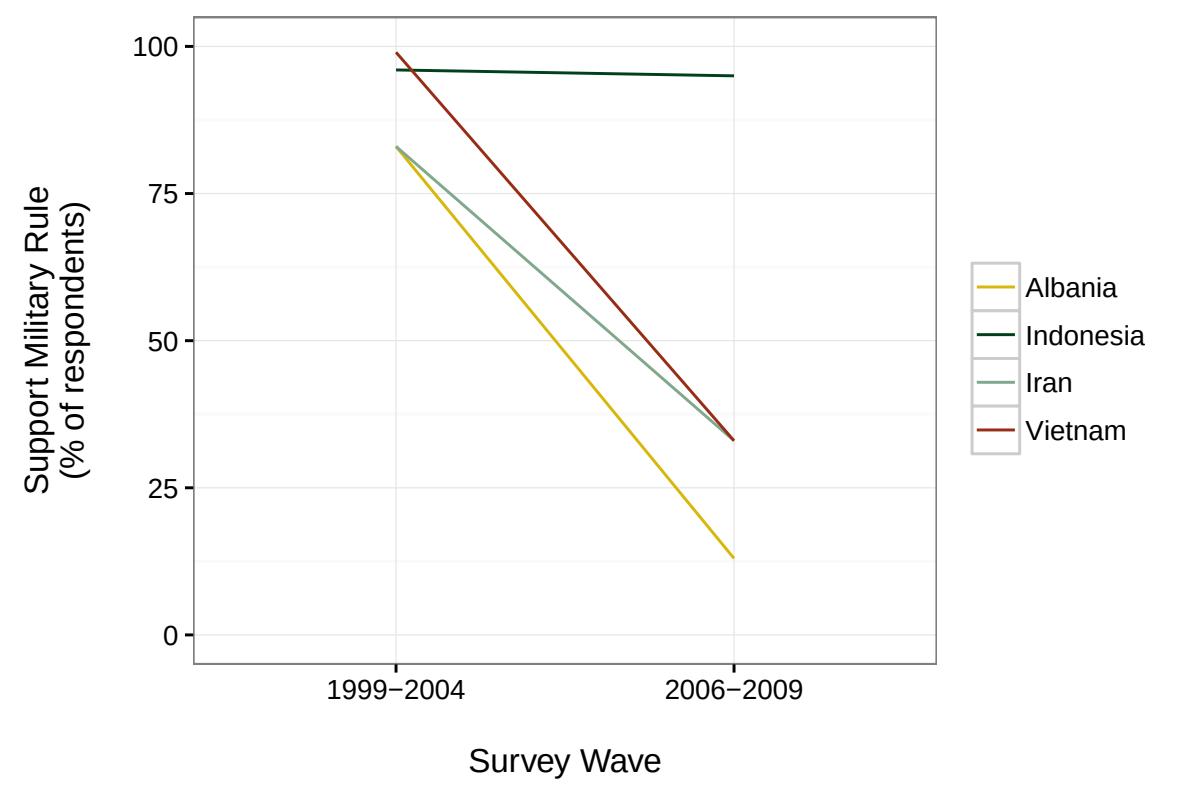
Kurzman (2014) argues that it is more likely that the **question translation changed**.

	1999-2005	2006-2009
English	having the military rule	having the military rule
Indonesia	Memiliki peraturan yang jelas tentang angkatan bersenjata (**having clear military rules**)	Memiliki peraturan tentang angkatan bersenjata (**having military rules**)
Albanian	Të kesh rregulla të ushtrisë (**having military rules**)	Të kesh regjim ushtarak (**having a military regime**)

More likely...

Kurzman (2014) argues that it is more likely that the **question translation changed**.

	1999-2005	2006-2009
English	having the military rule	having the military rule
Indonesia	Memiliki peraturan yang jelas tentang angkatan bersenjata (**having clear military rules**)	Memiliki peraturan tentang angkatan bersenjata (**having military rules**)
Albanian	Të kesh rregulla të ushtrisë (**having military rules**)	Të kesh regjim ushtarak (**having a military regime**)
Vietnamese	Vai trò cùa quân đội (**role of the military**)	[Unavailable]


Support Military Rule
(% of respondents)
100
75
50
25
0
1999−2004
2006−2009
Survey Wave
Albania
Indonesia
Iran
Vietnam

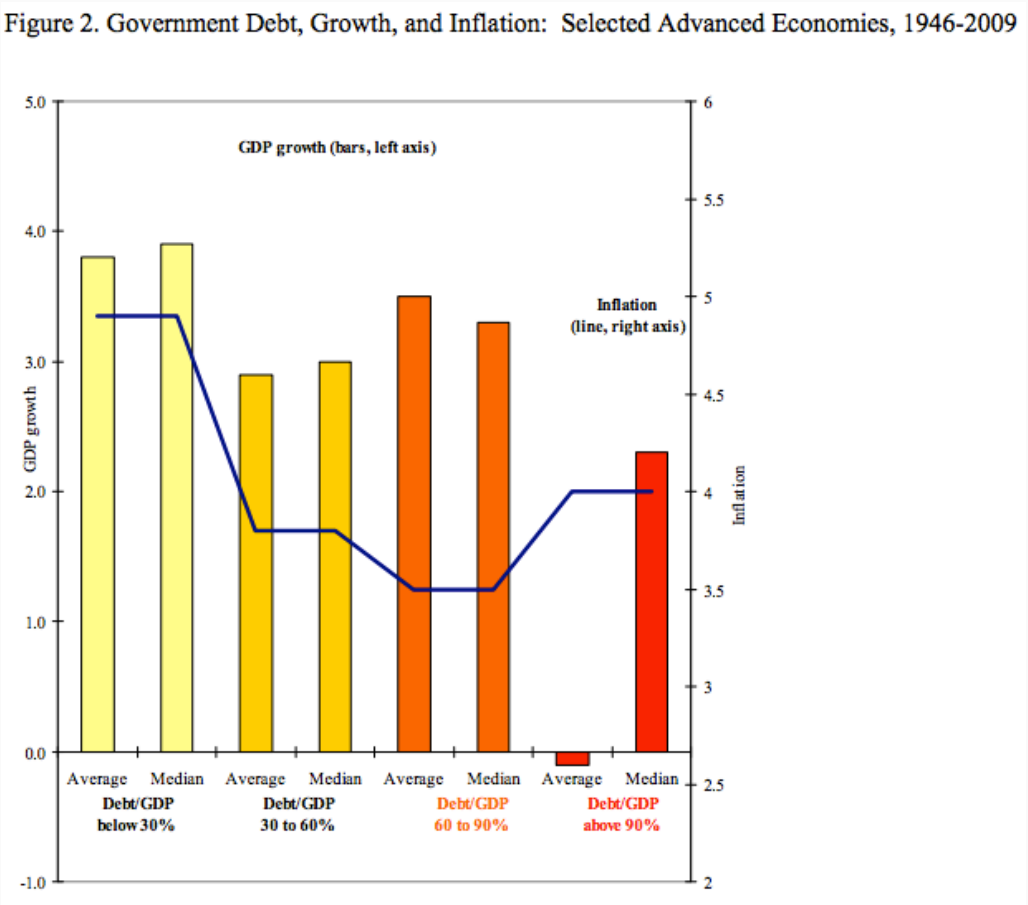
Reproducible research practices–the questionnaires and translations were available–**made it possible to find** these errors.

Example: Reinhart and Rogoff (2010)

Background: Reinhart and Rogoff (2010) in a highly influential study (in academics and government) found a threshold effect at 90% Public debt/GDP on economic growth. Above 90% growth slowed dramatically.

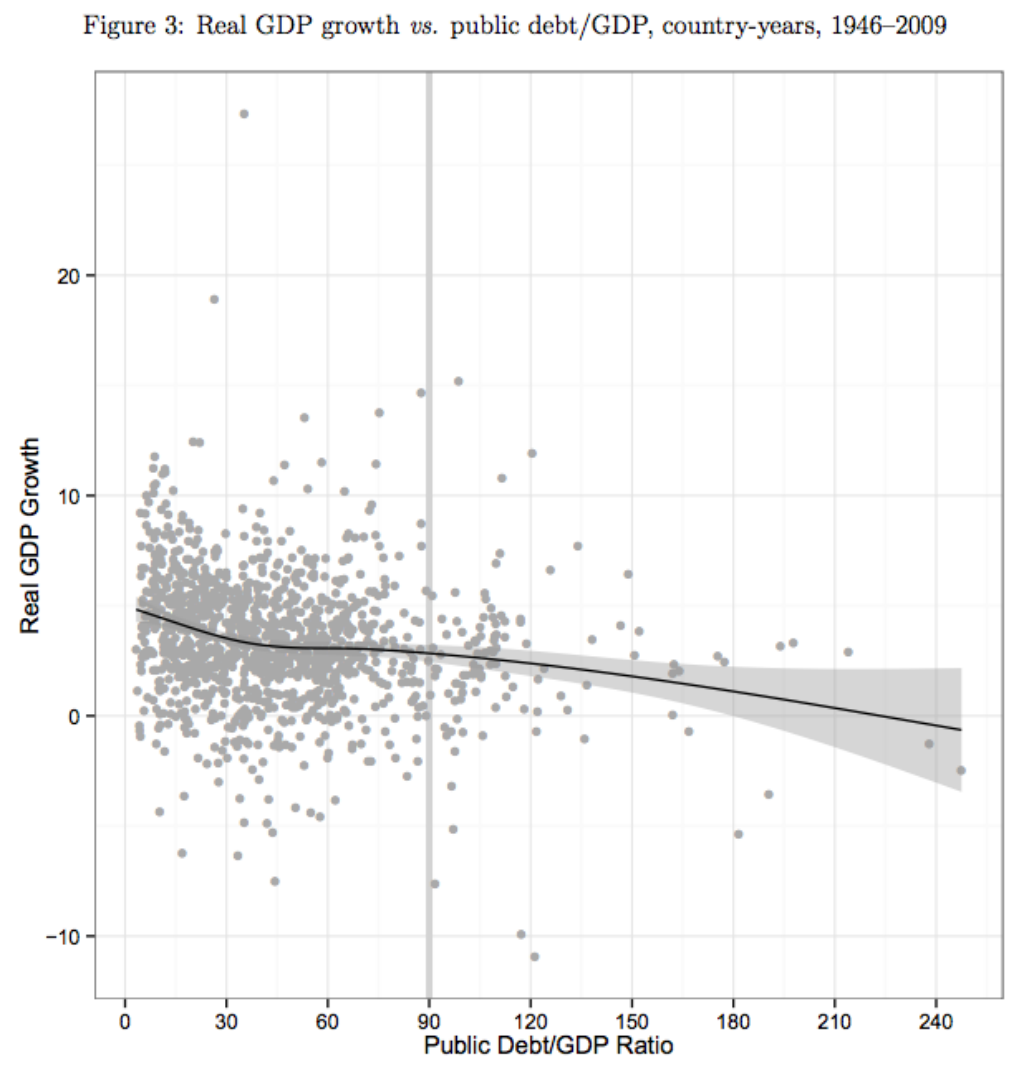
Original finding:

Figure 2. Government Debt, Growth, and Inflation: Selected Advanced Economies, 1946-2009



Replication: But, Herndon et al. (2014) found that an Excel coding error had dropped Australia, Austria, Belgium, Canada, and Denmark.

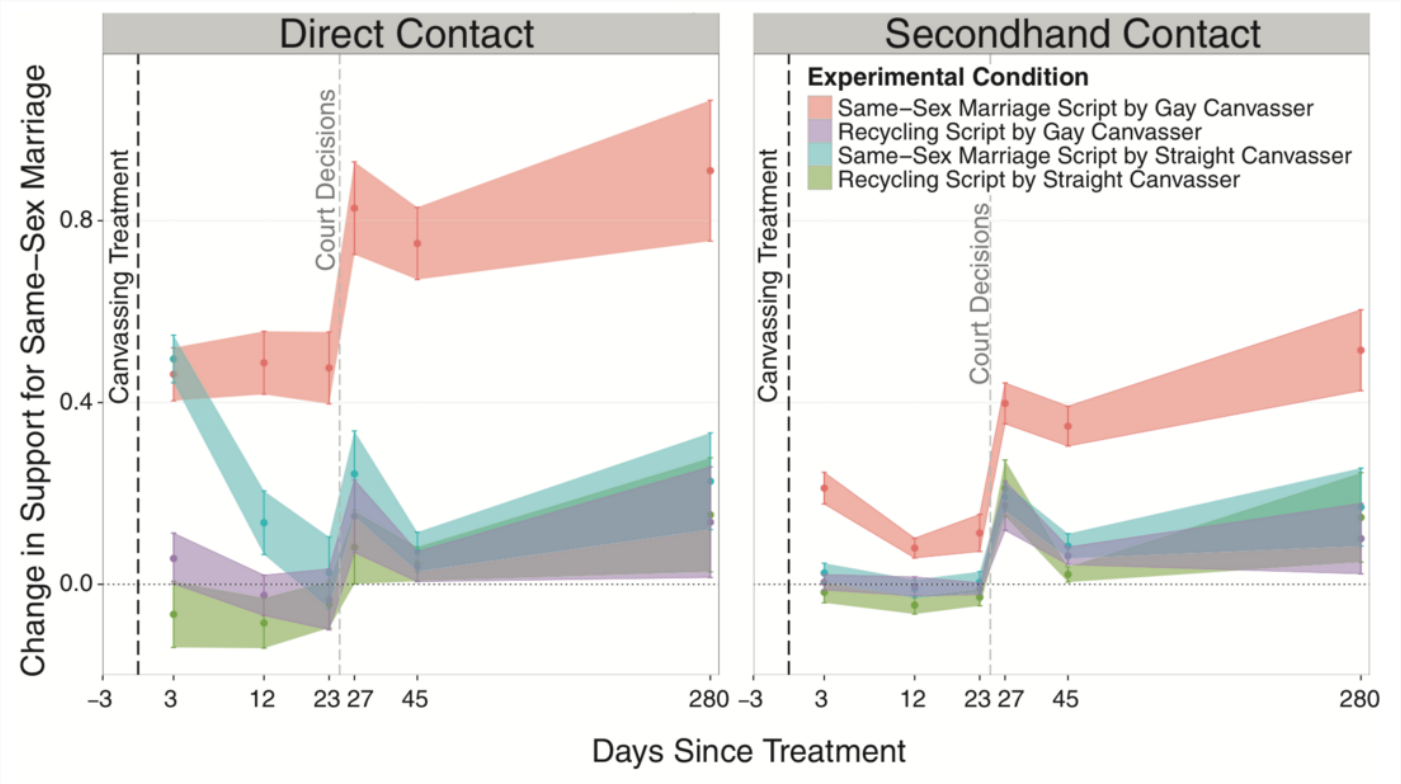
Corrected finding:



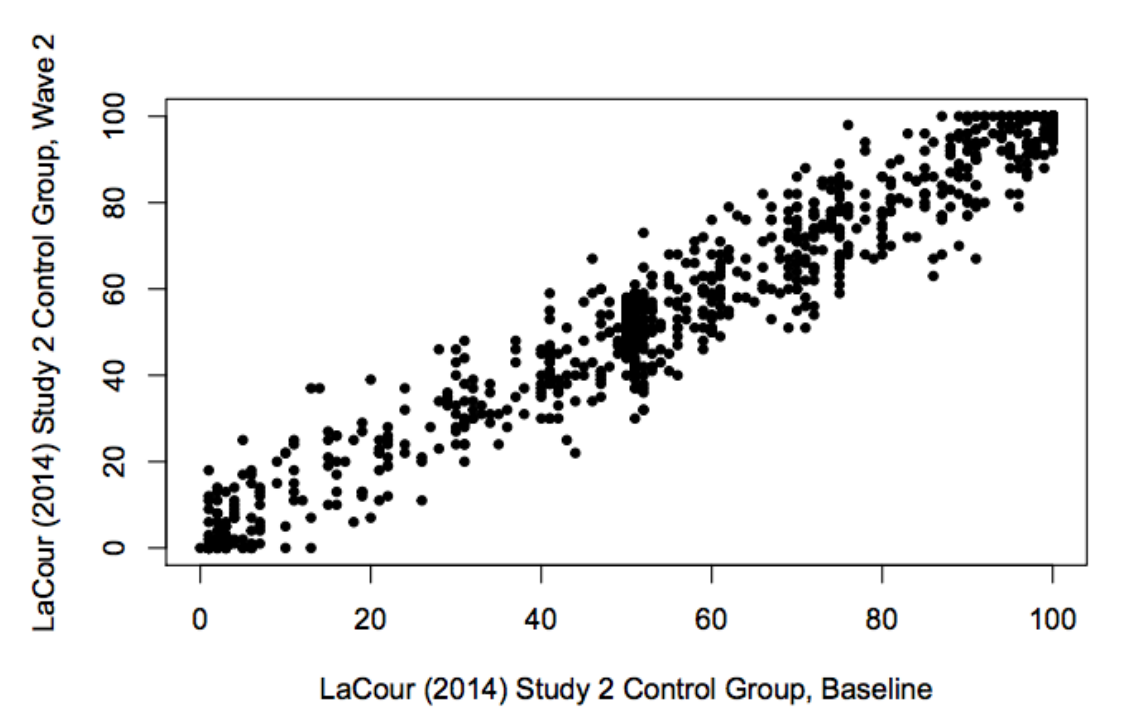
Example: LaCour Affair

Background: Lacour and Green (2014) found that having a conversation with a gay advocate had a very strong positive long-term effect on support for same-sex for marriage. This was **widely reported** in the popular press and advocacy groups began to use similar techniques.

Original finding:



Broockman et al. (2015) found that LaCour had **fabricated** the survey data. (Tech details: he took someone else's survey data and added random noise to the subsequent survey "waves".)



For science

Reproducible research also benefits science by:

- **Avoiding scientists wasting time** trying to understand things that do not exist.

- **Avoiding effort duplication** by cutting down the time the scientific community spends gathering data/developing analytic procedures.

How to do really reproducible research

Reproducible Research

When there is sufficient information available for independent researchers to make the same findings, using the same procedures with the *same data*.

Reproducible Research

When there is sufficient information available for independent researchers to make the same findings, using the same procedures with the *same data*.

How do we do this?

Doing reproducible research

Some key tips:

- Document everything!
- All of your files should be human-readable.
- From the start of your research project, have a plan to organise, store, and make your files accessible.
- Explicitly tie your files together.

Document Everything!

Need to fully document the steps we took and the rationale for these steps.

- Documentation *both* in the presentation document (usually discussion of general steps) and appendix files (e.g. data cleaning and analysis code, survey questionnaires, raw data).

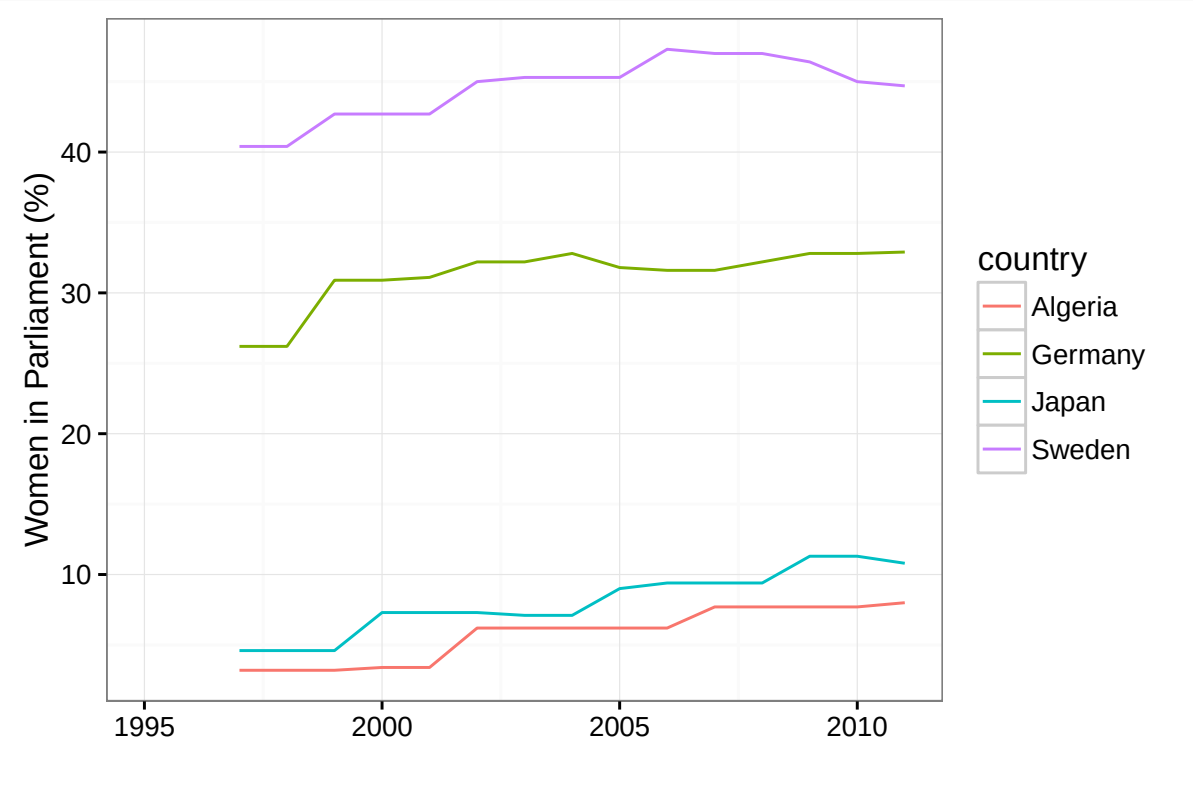
Document everything specifics: source code

It is important to record in detail the steps you used to clean and analyse your data, ideally with the original source code (e.g. SPSS syntax or R source).

Minimal (messy) R Example

```
library(WDI)
library(dplyr)
library(ggplot2)
setwd('U:\\research/group_project')
women<-WDI(indicator='SG.GEN.PARL.ZS',start=1995)
women<-rename(women_in_parl=SG.GEN.PARL.ZS)
women<-filter(women,country%in%c('Algeria','Germany','Japan
    ','Sweden'))
ggplot(women,aes(x=year,y=women_in_parl,colour=country))+
geom_line()+ylab('WomeninParliament(%)')+xlab('')+theme_bw()
```

Result


Women in Parliament (%)
40
30
20
10
1995
2000
2005
2010
country
Algeria
Germany
Japan
Sweden

Human-readable

Even files that you intend a computer to run should be **human-readable**.

So that another person (and yourself in the future) can **understand** what you did, even if the computer program no longer exists (helps **future-proof** your work).

So ...

Formatting and Comment characters

Write your source code with the **intention** that it will be **read by a person**.

- Use a **consistent style** (just as you would in a presentation document).
- Use comment characters that allow you to write information that humans can read and the computer will ignore.
  - R comment character: `#`
  - SPSS comments: Begin with `*` and end with `.`

Minimal (human-readable) R Example (1)

```
1  ##############################
2  # Gather women in parliament data from WDI and plot subset
3  # Christopher Gandrud
4  ##############################
5
6  # Load required packages
7  library(WDI)
8  library(dplyr)
9  library(ggplot2)
10
11 # Set working directory. Changed as needed
12 setwd('U:\\research/group_project')
13
14 # Download women in parliament data from
15 # World Bank Development indicators
16 # from 1995. Indicator ID = SG.GEN.PARL.ZS
17 women <- WDI(indicator = 'SG.GEN.PARL.ZS', start = 1995)
```

Minimal (human readable) R Example (2)

```
# Rename women in parliament indicator.
women <- rename(women, women_in_parl = SG.GEN.PARL.ZS)

# Select data from Algeria, Germany, Japan, and Sweden.
women <- filter(women, country %in%
                  c('Algeria', 'Germany', 'Japan', 'Sweden'))

# Create a comparative line plot of the data.
ggplot(women, aes(x = year, y = women_in_parl,
                      colour = country)) +
geom_line() +
ylab('Women in Parliament (%)') + xlab('') +
theme_bw()
```

Minimal SPSS Example

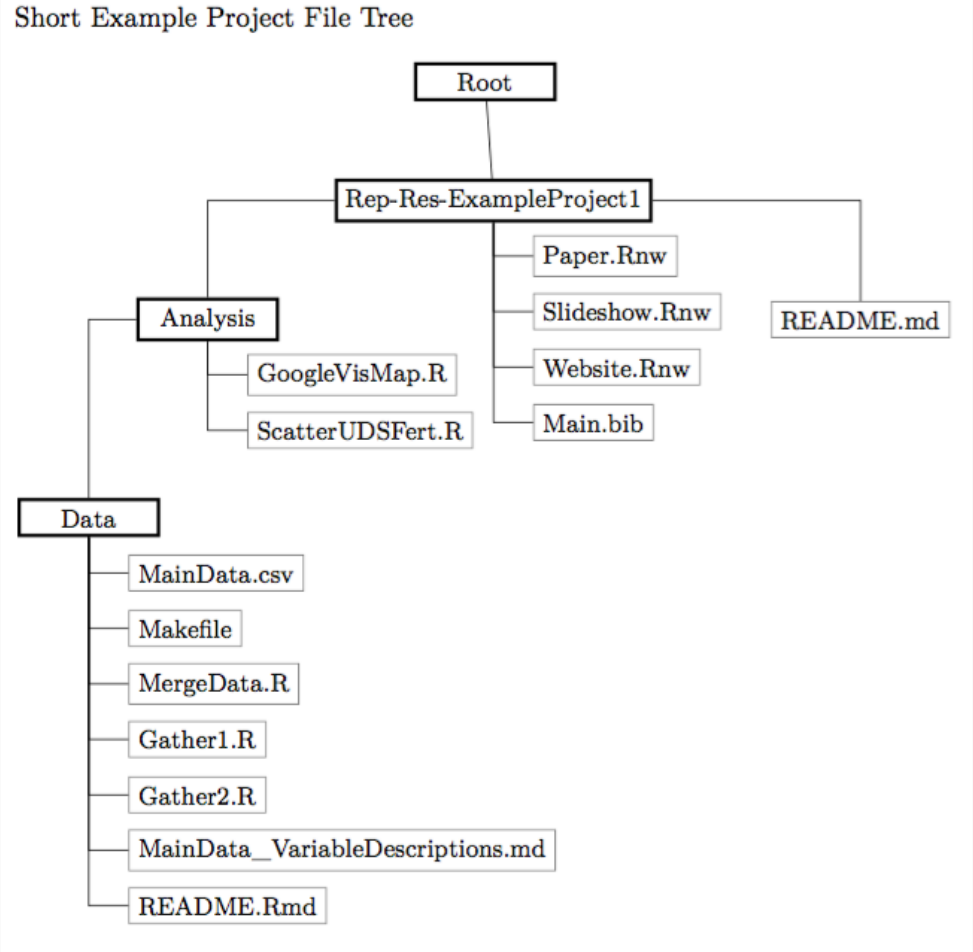
```
* Load employee survey data & recode age.
* Christopher Gandrud.

* Load data from employee_data.sav file in the C drive.
GET FILE='C:\group_project\data\employee_data.sav'.

*Recode respondent age in years into a 3-category variable.
RECODE agea (Lowest thru 25=1) (26 thru 64=2) (65 thru
    Highest=3) INTO age3.
VARIABLE LABELS  age3 'age - 3 categories'.
VALUE LABELS age3
1 '25 or under'
2 '26-64'
3 '65 or over'.
EXECUTE.
```

Organise your files into an understandable hierarchy

Start your research with a plan to organise, store, and make your files available. Your files should be arranged in a **hierarchy** that makes the **purpose** of the files clear to independent researchers.



Organise your files

Give your files **unique and understandable names** that make sense within the file hierarchy.

Bad	Better
`File1.sps`	`recode_survey.sps`

Organise your files

Give your files **unique and understandable names** that make sense within the file hierarchy.

Bad	Better
`File1.sps`	`recode_survey.sps`
`source.R`	`download_wdi_data.R`

Organise your files

Give your files **unique and understandable names** that make sense within the file hierarchy.

Bad	Better
`File1.sps`	`recode_survey.sps`
`source.R`	`download_wdi_data.R`
`data.sav`	`brexit_survey.sav`

Organise your files

Give your files **unique and understandable names** that make sense within the file hierarchy.

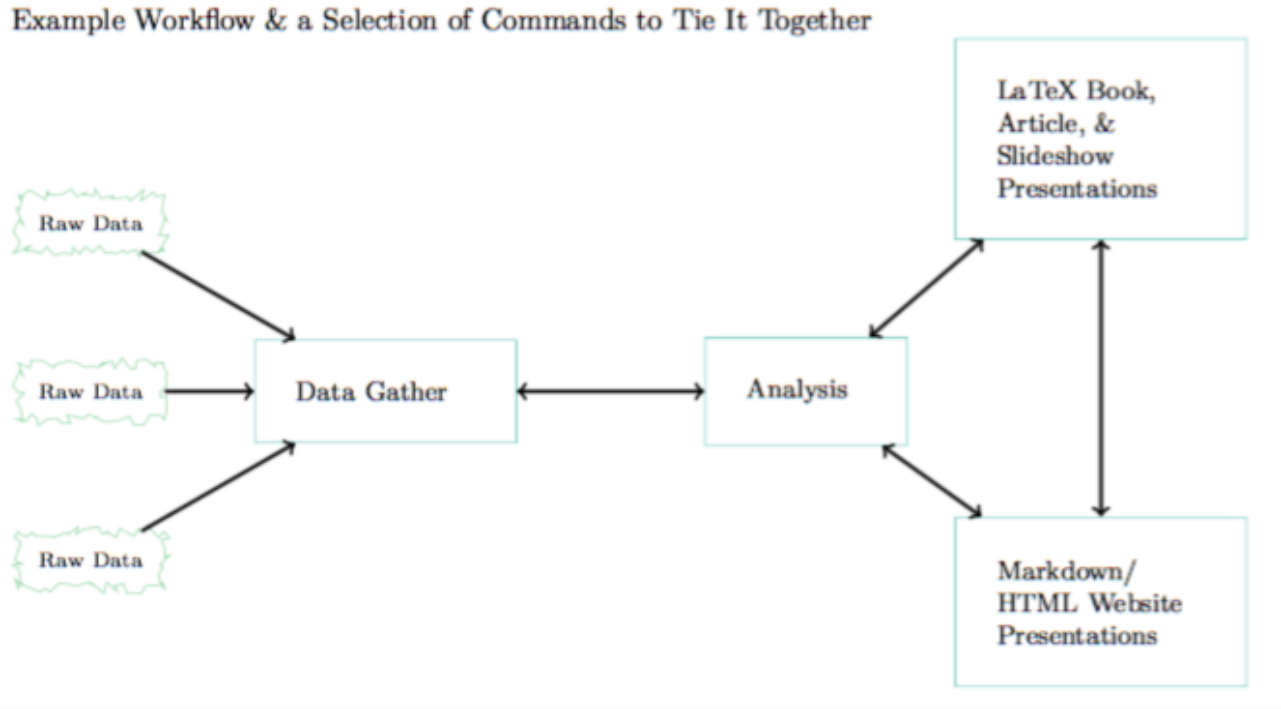
Bad	Better
`File1.sps`	`recode_survey.sps`
`source.R`	`download_wdi_data.R`
`data.sav`	`brexit_survey.sav`
`data final 6.csv`	`women_in_parliament_v6.csv`

Organise your files

Include a **README** file that **explains your file hierarchy**, so that others can figure out what the files do.

File ties

Your files should be **tied** together in a **documented** and **understandable** way, so that others (and your future self) will understand your research process.



Gandrud (2015, 21)

Note, we don't cover all of the tools to do all of this in this course.

Minimal R Example

```r
##############################
# Gather women in parliament data from WDI and save locally
# Christopher Gandrud
##############################

# Load required packages
library(WDI)
library(dplyr)
library(rio)

# Set working directory. Changed as needed
setwd('U:\\research/group_project')

# Download women in parliament data from
# World Bank Development indicators
# from 1995. Indicator ID = SG.GEN.PARL.ZS
women <- WDI(indicator = 'SG.GEN.PARL.ZS', start = 1995)

# Save data in CSV format
export(women, file = 'data/women_in_parl_wdi.csv')
```

Minimal SPSS Example

```
* Load employee survey data, recode age, and save.
* Christopher Gandrud.

* Load data from employee_data.sav file in the C drive.
GET FILE='C:\group_project\data\employee_data.sav'.

*Recode respondent age in years into a 3-category variable.
RECODE agea (Lowest thru 25=1) (26 thru 64=2) (65 thru
    Highest=3) INTO age3.
VARIABLE LABELS  age3 'age - 3 categories'.
VALUE LABELS age3
1 '25 or under'
2 '26-64'
3 '65 or over'.
EXECUTE.

* Save subsetted data set.
SAVE OUTFILE='C:\group_project\data\employee_data_cleaner.
    sav'.
```

The reproducible research ideal

Ideally, data gathering, analysis, and presentation can be **dynamically** linked.

We don't cover these tools in this class, but (FYI) they go by names such as RMarkdown and Sweave.

This . . .

```
---
title: "Example Dynamically Linked Research Project"
author: "Christopher Gandrud"
date: "10 March 2016"
output: pdf_document
---

```{r setup, include=FALSE}
knitr::opts_chunk$set(echo = FALSE)
knitr::opts_chunk$set(message = FALSE)
knitr::opts_chunk$set(warning = FALSE)

```

R Markdown

Lorem ipsum dolor sit amet, reprehendunt interpretaris his cu. Fierent blandit ei vel, mel aliquip habemus no, eos novum eligendi ea. Eu unum facete salutatus sit. Has eu assum equidem deseruisse.

Sumo dicant nonumy his ea, usu veri utinam at, duo ex ludus voluptua. Eos et modus prompta, quo ut noluisse sensibus. Quem possit splendide ea duo, usu lorem audiam ne, ei sea delenit lobortis disputando. Te scripta epicuri convenire ius, eum fabulas admodum quaestio in.

Ut alia discere conclusionemque eos, in mei mollis assentior. Mel ea dolor urbanitas. Te commodo suscipit nec, ei quo detraxit gloriatur. Vel ancillae propriae temporibus ei. Vel ne solet doming definitionem, vim iudico possit pericula ex.


```{r}
# Load required packages
library(WDI)
library(dplyr)
library(ggplot2)

# Download women in parliament data from World Bank Development indicators
# from 1995. Indicator ID = SG.GEN.PARL.ZS
women <- WDI(indicator = 'SG.GEN.PARL.ZS', start = 1995)

# Clean up: (1) rename women in parliament indicator,
# (2) select data from Algeria, Germany, Japan, and Sweden
women <- rename(women, women_in_parl = SG.GEN.PARL.ZS)
women <- filter(women, country %in% c('Algeria', 'Germany', 'Japan', 'Sweden'))
```

Including Plots

Dicat fuisset aliquando vix cu, malis velit persequeris per et.

```{r}
# Create a comparative line plot of the data
ggplot(women, aes(x = year, y = women_in_parl, colour = country)) +
geom_line() +
ylab('Women in Parliament (%)') + xlab('') +
theme_bw()
```

Nemore senserit assueverit eum no, qui vocent option accusam ut. Integre detraxit ad vel, mei libris noluisse suavitate ad. Stet novum dicant ius cu, quot doming iudicabit ea mel. Rebum iracundia vix ne, ex probo labores feugait cum, nihil choro et nec. Novum vocibus deterruisset ei mea.
¶
```

Is all you need to create this . . .

Example Dynamically Linked Research Project

Christopher Gandrud

10 March 2016

R Markdown

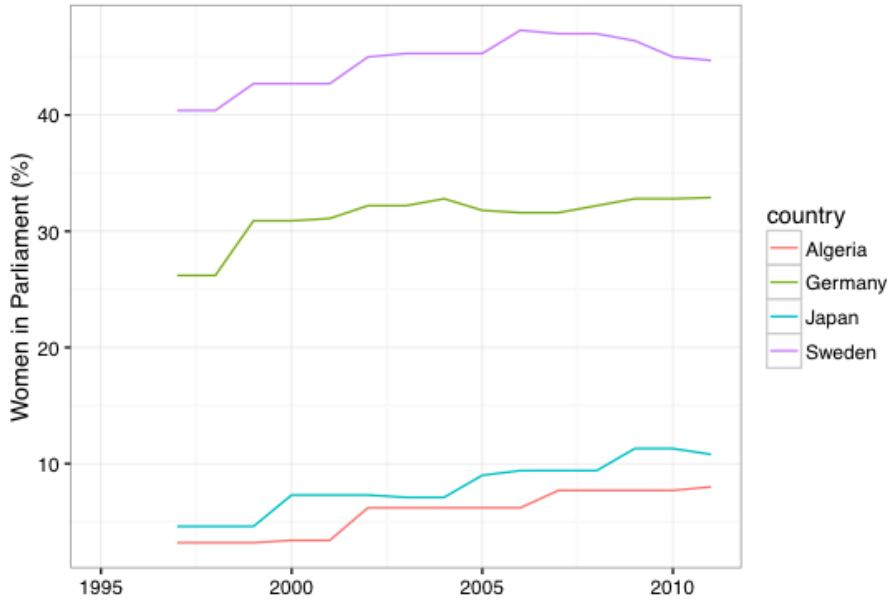
Lorem ipsum dolor sit amet, reprehendunt interpretaris his cu. Fierent blandit ei vel, mel aliquip habemus no, eos novum eligendi ea. Eu unum facete salutatus sit. Has eu assum equidem deseruisse.

Sumo dicant nonumy his ea, usu veri utinam at, duo ex ludus voluptua. Eos et modus prompta, quo ut noluisse sensibus. Quem possit splendide ea duo, usu lorem audiam ne, ei sea delenit lobortis disputando. Te scripta epicuri convenire ius, eum fabulas admodum quaestio in.

Ut alia discere conclusionemque eos, in mei mollis assentior. Mel ea dolor urbanitas. Te commodo suscipit nec, ei quo detraxit gloriatur. Vel ancillae propriae temporibus ei. Vel ne solet doming definitionem, vim iudico possit pericula ex.

Including Plots

Dicat fuisset aliquando vix cu, malis velit persequeris per et.



Nemore senserit assueverit eum no, qui vocent option accusam ut. Integre detraxit ad vel, mei libris noluisse suavitate ad. Stet novum dicant ius cu, quot doming iudicabit ea mel. Rebum iracundia vix ne, ex probo labores feugait cum, nihil choro et nec. Novum vocibus deterruisset ei mea.

Your group projects

Your project should be **reproducible**.

- Your paper should have clear descriptions of your measurement instrument
  - Survey: questionnaire
  - Content analysis: coding scheme
  - Composite indicators: original variables and how they were aggregated
- A complete account of your data
  - Including as much raw data (e.g. texts for content analysis, data sets for surveys) as possible with instructions (detailed descriptions and syntax where possible) for how you accessed/set up, cleaned and analysed it
  - You should submit non-presentation files as appendices to your report

Note: In the Labs

Note: in the labs this week we will be introducing you to the **Q-Step Pathways** you can choose from.