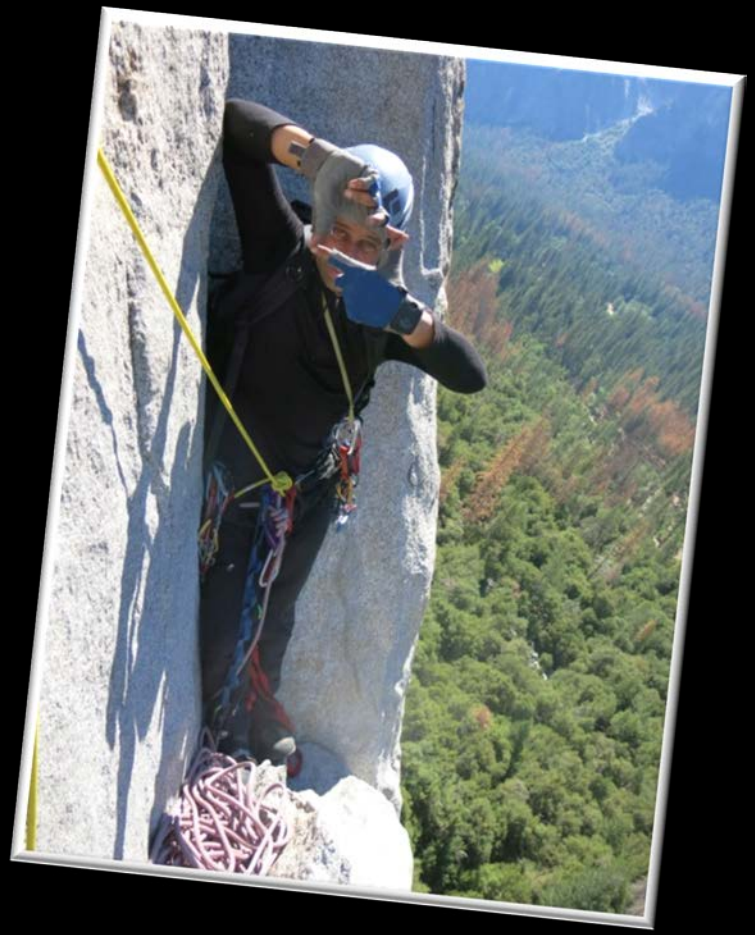
The paper is just a palimpsest

Bamberg, September 2016

Vernon Gayle
University of Edinburgh
@profbigvern



A BLACK BOX

Lack of access to the research code that produced the research output

Command files
Syntax files
.do Stata files
R scripts
.sps SPSS files
Jupyter notebooks

A parchment or other writing surface on which the original text has been effaced or partially erased, and then overwritten by another; OED

palimpsest, *n.* and *adj.*

Pronunciation: Brit. /ˈpalɨm(p)sɛst/ , U.S. /ˈpæləm(p)ˌsɛst/

Sharing Research Code

Sufficient information to enable other researchers to

understand

evaluate

build upon the work

Sharing Research Code

Sufficient information to enable other researchers to understand, evaluate and build upon the work

Enough information for a third party to reproduce (i.e. duplicate) results without needing to get additional information from the authors

Sharing Research Code

Sufficient information to enable other researchers to understand, evaluate and build upon the work

Enough information for a third party to reproduce results without needing to get additional information from the authors

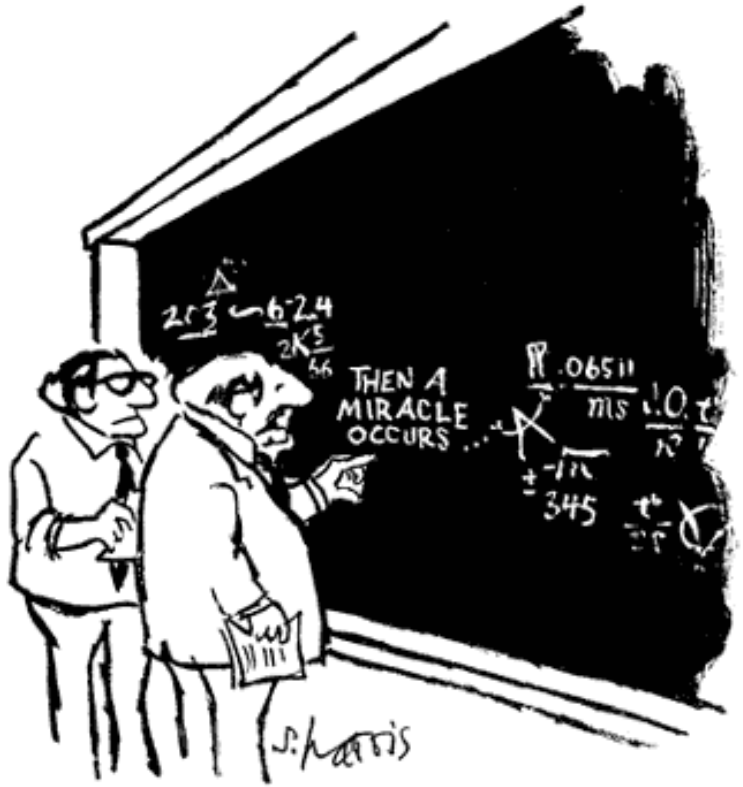
Required by *Science, Am. Econ. Rev., Econometrica, Rev. Econ. Studies*

500 journals signed up to the Transparency and Openness Promotion (TOP) Guidelines

“Show me your working out”

“Show me your working out”

The Royal Society's motto ‘*Nullius in verba*’ (take nobody’s word for it)



“I think you should be more explicit here in step two.”

Duplication

Step 1

Work can be duplicated if sufficient information is made available which ensures consistent results can be produced using the same data and the same analytical techniques

Replication

Step 2

A replication study can ‘duplicate’ the original work….

But can also further test the robustness of the original work by employing new or additional data (or measures) and alternative data analysis techniques

Duplication

Logistic Regression 5+ GCSEs (A*-C) YCS Cohort 9

	b	se	t	p
5+ GCSEs (A*- C)				
Girls	0.405	0.039	10.305	0.000
Boys	0.000	.	.	.
Chinese	2.002	0.377	5.306	0.000
Indian	1.066	0.208	5.117	0.000
White	0.643	0.171	3.757	0.000
Bangladeshi	0.766	0.345	2.222	0.026
Pakistani	0.531	0.245	2.169	0.030
Black	0.000	.	.	.
Professional/Non-Manual	2.192	0.109	20.179	0.000
Other Non-Manual	1.773	0.108	16.423	0.000
Skilled Manual	0.932	0.104	8.954	0.000
Semi-Skilled Manual	0.576	0.113	5.112	0.000
Unskilled	0.000	.	.	.
Constant	-2.208	0.198	-11.152	0.000
n	12789			

Produced in Stata using svy; Connolly (2006) used SPSS with data weighted incorrectly!

Appendix: details of binary logistic regression models derived from the three cohorts (see Table 3)

Table 5. Binary logistic regression on whether school leavers in England and Wales in 1997 gained five or more GCSE grades A*–C or not[1]

	B	S.E.	Wald	df	Sig.	Exp(B)
Gender[2]						
Girls	0.405	0.038	114.340	1	<.001	1.499
Ethnicity[3]						
Chinese	2.002	0.341	34.436	1	<.001	7.406
Indian	1.066	0.193	30.389	1	<.001	2.903
White	0.643	0.159	16.372	1	<.001	1.902
Bangladeshi	0.766	0.332	5.330	1	.021	2.151
Pakistani	0.531	0.230	5.338	1	.021	1.701
Social Class[4]						
Professional/ Managerial	2.192	0.110	396.863	1	<.001	8.954
Other Non-Manual	1.773	0.110	261.000	1	<.001	5.886
Skilled Manual	0.932	0.107	76.255	1	<.001	2.540
Semi-Skilled Manual	0.576	0.115	24.965	1	<.001	1.779
Constant	−2.208	0.189	136.885	1	<.001	0.110

[1]Source of data: Secondary analysis of data derived from first sweep of Cohort 9 of the Youth Cohort Study of England and Wales.
[2]Reference category: boys.
[3]Reference category: Black pupils.
[4]Reference category: Unskilled Manual Occupations.

Connolly, Paul. "The effects of social class and ethnicity on gender differences in GCSE attainment: a secondary analysis of the Youth Cohort Study of England and Wales 1997–2001." *British Educational Research Journal* 32.1 (2006): 3-21.

Duplication

100 lines of code (Stata syntax) to enable the data

I know the YCS very well and know Stata

Barrier to duplication

Logistic Regression 5+ GCSEs (A*-C) YCS Cohort 9

	b	se	t	p
5+ GCSEs (A*- C)				
Girls	0.405	0.039	10.305	0.000
Boys	0.000	.	.	.
Chinese	2.002	0.377	5.306	0.000
Indian	1.066	0.208	5.117	0.000
White	0.643	0.171	3.757	0.000
Bangladeshi	0.766	0.345	2.222	0.026
Pakistani	0.531	0.245	2.169	0.030
Black	0.000	.	.	.
Professional/Non-Manual	2.192	0.109	20.179	0.000
Other Non-Manual	1.773	0.108	16.423	0.000
Skilled Manual	0.932	0.104	8.954	0.000
Semi-Skilled Manual	0.576	0.113	5.112	0.000
Unskilled	0.000	.	.	.
Constant	-2.208	0.198	-11.152	0.000
n	12789			

Produced in Stata using svy: Connolly (2006) used SPSS with data weighted incorrectly!

Connolly, Paul. "The effects of social class and ethnicity on gender differences in GCSE attainment: a secondary analysis of the Youth Cohort Study of England and Wales 1997–2001. " *British Educational Research Journal* 32.1 (2006): 3-21.

Replication

Logistic Regression 5+ GCSEs (A*- C) YCS Cohort 9

	Original b		NS-SEC b	
5+ GCSEs (A*- C)				
Boys	0.000		0.000	
Girls	0.405	***	0.434	***
White	0.000		0.000	
Chinese	1.359	***	1.491	***
Indian	0.423	***	0.598	***
Bangladeshi	0.123		0.320	
Pakistani	-0.112		0.208	
Black	-0.643	***	-0.715	***
Professional/Non-Manual	0.000			
Other Non-Manual	-0.420	***		
Skilled Manual	-1.260	***		
Semi-Skilled Manual	-1.616	***		
Unskilled	-2.192	***		
Large employers and higher managerial			0.000	
Higher professional occupations			0.371	***
Lower managerial & professional			-0.417	***
Intermediate occupations			-0.848	***
Small employers and own account workers			-1.437	***
Lower supervisory and technical occupations			-1.749	***
Semi-routine occupations			-1.944	***
Routine occupations			-2.361	***
Constant	0.627	***	0.716	***
n	12789		12788	

Produced in Stata using svy: Connolly (2006) used SPSS with data weighted incorrectly!

Why bother?

- Improves transparency - don’t just trust me – I will show you
- Allows others to understand, evaluate, and build upon the work
- Checks on accuracy
- Facilitates incremental development (and comparative work)

Are we serious about what we do?
(Edinburgh University’s mission is the creation, dissemination and curation of knowledge)

The Horror of a Retraction…

The authors of a March 2015 *Journal of Health and Social Behavior* (JHSB) study, "In Sickness and in Health? Physical Illness as a Risk Factor for Marital Dissolution in Later Life" (2015, 56(1):59-73), have retracted the article.

There was a major error in the coding in their dependent variable of marital status. The conclusions of that study should be considered invalid.

A corrected version of the article will appear in the September 2015 issue of JHSB.

http://www.asanet.org/press-center/press-releases/journal-health-and-social-behavior-publish-corrected-version-study

Levels of Reproducibility

1. Unreproducible

2. Reproducible with Effort

3. Completely Reproducible

Things we should do immediately

- Deposit annotated scripts that a THIRD PARTY can use to completely duplicate ALL the results included in the published work
  - Clearly state the data source (and release)
  - Clearly state software used including versions, libraries and dependencies (even seeds)
  - Include all the script needed for data enabling
  - Deposit well annotated code books (detailing variables)
  - Check that a THIRD PARTY can genuinely duplicate the work

Things we could do

- Be clear about ‘all’ the work you did rather than just the work that has been reported

- Provide a justification of the micro-actions undertaken

  minor recodes

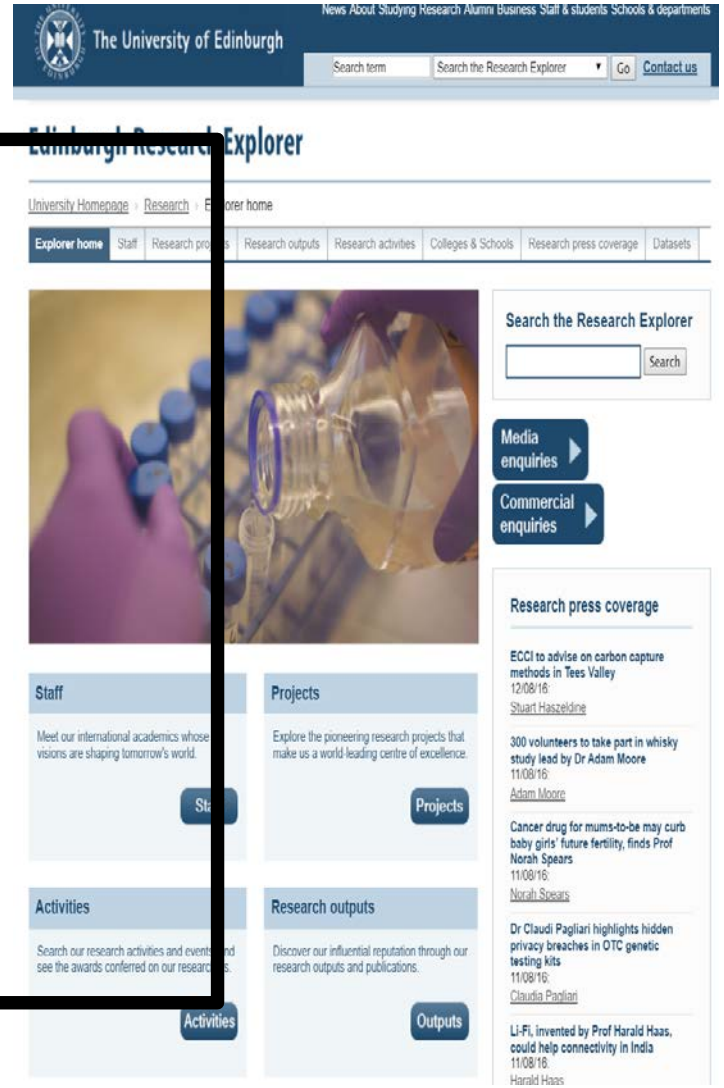
  variable choices (using wQFACHI not wQFEDHI)

- Peer programming (one player on the ball one player off the ball) ??

- More internal checking within research teams

Open the BLACK BOX

Routinely sharing research code



www.vernongayle.com/talks-recent.html

Research Talks

This page contains files related to research talks that I have delivered

---

Gayle, V. (2016) 'Is the Paper Just a Palimpsest? An appeal for reproducible social stratification research', *Social Stratification Research Seminar,* Cambridge.

Files supporting reproducibility

Stata syntax file (research code) cambridge_20160901_vg_v4.do

Stata 14 data file fake_data_20160828_vg_v1.dta

Stata 13 data file fake_data_20160828_stata13_vg_v1.dta

Excel versions of the data file fake_data_20160828_vg_v1.xlsx

Excel versions of the data file (csv) fake_data_20160828_vg_v1.csv

Vernon Gayle
Professor of Sociology & Social Statistics
University of Edinburgh
vernon.gayle@ed.ac.uk
@profbigvern