



Research Integrity

Peter Schotman

EAA PhD forum

Research integrity

- Why discuss on doctoral colloquium of EAA in Maastricht?
 - Not because we suspect PhD's are wrongdoers
 - Not because Accounting has a particularly bad reputation
 - Not because Maastricht University has had major incidents
- Good to be aware of thin line between acceptable and questionable research practices
- Sometimes PhD unwittingly victim of breaches of ethical standards by senior research supervisors
 - inflicts future career

Overview

1. Data

- Fabricated, manipulated, confidential

2. Plagiarism

- self-plagiarism, redundant publication

3. Biased research

- sponsor interests


4. Statistics

- ex-post hypotheses, p-values, exploratory data analysis

Examples from social science research (incl economics) in the Netherlands

- **Fabricated data:** Professor fills in questionnaire by himself. Phd student is handed ready-to-use data. Caught when some PhD students became suspicious about their `data`.
- **Manipulated data:** Professor changes data points to obtain significant result. Caught because too many publications with results too good to be true.
- **Unverifiable results:** Publication (student + supervisor) retracted after few years by supervisor when former student appears unwilling or unable to allow others to replicate.
- **Self-plagiarism:** Student writes chapter with supervisor using data and methods from earlier research of supervisor. Reading committee rejects dissertation. Long legal battle.

Types of data

- 
- Freely available on the web from official institutions (Central Banks, governments, ...)
 - Proprietary databases open to all researchers for a fee (Compustat, ThomsonReuters, ...)
 - Hand collected from primary sources
 - Confidential data about individuals, organisations and firms from official agencies (regulator, tax office, statistical agency, ...)
 - Experiments
 - Surveys
 - Confidential data from a private source, not available to other researchers (internal firm records, ...)

Secret data

"I have a truly marvelous regression result, but I can't show you the data and won't even show you the computer program that produced the result." - Typical paper in economics and finance.

(John Cochrane, *The Grumpy Economist* blog Dec 28, 2015)

- Who can check reliability of data?
- Who can verify the analysis?

- Big trust in integrity of researchers
 - How often will there be a problem?

Knowing about your data

Professor **F** obtained a great dataset from corporation **AX** with lots of internal details about the company. As a PhD student you work with Professor **F** on the data.

Your results provide novel insights on firm behavior. Your work will surely be publishable in a top journal.

- You only have the final data, not the way it has been collected. Professor **F** says he cannot disclose the name of corporation **AX** (afraid of bad publicity; not giving information to competitors)
- What if you know the name of the firm, but still cannot verify the data?

Data management

From the code of conduct for Dutch universities (VSNU, 2004, 2014)

- 3.1. Research must be replicable in order to verify its accuracy. The choice of research question, the research set-up, the choice of method and the references to sources used are accurately documented in a form that allows for verification of all steps in the research process.
- 3.2. The quality of data collection, data input, data storage and data processing is closely guarded. All steps taken must be properly reported and their execution must be properly monitored (lab journals, progress reports, documentation of arrangements and decisions, etc.).
- 3.3. Raw research data are stored for at least ten years. These data are made available to other academic practitioners upon request, **unless legal provisions dictate otherwise.**

[http://vsnu.nl/files/documenten/Domeinen/Onderzoek/
The_Netherlands_Code%20of_Conduct_for_Academic_Practice_2004_\(version2014\).pdf](http://vsnu.nl/files/documenten/Domeinen/Onderzoek/The_Netherlands_Code%20of_Conduct_for_Academic_Practice_2004_(version2014).pdf)

University policy

“Research data must be stored and archived in the infrastructure facilities made available by UM at the end of the research project (or earlier depending on the relevant faculty guidelines or other applicable rules). If the research data is part of an external collection managed elsewhere, the UM researcher must adequately refer to this and include when (if applicable) and by whom this data can be consulted in the external collection.”

- Data are valuable (not only for verification)
- Individual responsibility
- Will this avoid data integrity problems?
 - Manipulated data can still be stored
 - Much data will not be stored
- **Administrative burden:** who checks if stored data are valid?

The Review *of* Economics and Statistics

Journal policies

“The Review of Economics and Statistics is implementing a strict data and computer code availability policy for empirical papers. Authors of papers accepted for publication will be required to

1. post their code and programs
2. post and document their data (or document their data and include instructions for how other researchers can obtain the data when the data have been obtained under an arrangement that precludes the posting of the data)
3. post detailed readme files on-line before publication.

(...)

In general we allow the use of proprietary data as long as (1) there exists some way to apply for the data, (2) it is expected that reasonable applications will be accepted, (3) the authors will provide all the information necessary to go from the raw data to the results of the paper (including code).”

Similar policies at other economics journals, but not in Finance
(Accounting?)

Plagiarism

- Obviously unethical
 - Literal copying from others easy to detect using software
 - Less naïve forms harder to track: translation, ideas, ...
- Self-plagiarism: re-using own material
 - Redundant publications: large overlap with other papers
 - Mostly a problem for journals
 - Full disclosure to prevent problems

Dutch universities code of conduct on self-plagiarism

1.5 Academic practitioners do not republish their own previously published work or parts thereof as though it constituted a new contribution to the academic literature.

When republishing previously published findings, they indicate this with a correct reference to the source or by another means accepted within the discipline.

In many disciplines it is permissible and even customary to reprint short texts from works published with or without co-authors without a source reference when it concerns brief passages of introductory, theoretical or methodological explanation.

VSNU (2004, 2014)

Self-plagiarism as an incentive problem

- Reduce emphasis on number of publications
- Reputation based on best publications (and its citations)
- Focus on quality, not on quantity
 - Dutch standard research assessment protocol

Biased research

- Sponsored research and funding from grant institutions inevitably give direction to research questions
- Code of conduct for Dutch universities (VSNU, 2004, 2014):

- 5.4. The option to publish academic research results is assured. Arrangements with external research funders always stipulate that the academic practitioner is at liberty to publish the results within a specified, reasonable period.
- 5.5. External funders of scientific and scholarly activities are identified by name (...).
- 6.2. Academic practitioners allow themselves to be judged on the quality of their output in an honest and loyal fashion, and they cooperate in internal and external assessments of their research.

Self censorship?

You obtain confidential data from firm **AX**. In exploring the data you detect a pattern that suggests its clients are not well served. It would explain how the industry works.

For publication a journal editor and referee ask for more details on the data, including the name of the firm.

- What if disclosing its identity would be costly to firm **AX**?
- What if your results may be used as evidence about illegal practices?

Obtaining significant results

- Publication bias: accept studies that show significant effects
 - false discoveries, exacerbated by variations in test design
 - p -hacking: keep on testing until significant
 - ex-post hypotheses
- When does this become fraud?



Slim by Chocolate

- Deliberately fake study to prove a point (and more)
- Measure many attributes of people without any hypothesis
- See which correlations turn out significant
- Formulate a hypothesis
- Present result starting with hypothesis

I Fooled Millions Into Thinking Chocolate Helps Weight Loss. Here's How.



John Bohannon

5/27/15 4:23pm · Filed to: DEBUNKERY



1.1M



558



280



<http://io9.gizmodo.com/i-fooled-millions-into-thinking-chocolate-helps-weight-1707251800>

Statistical tricks

- Playing with regressions
 - search over different control variates and report best result: “sinning in the basement” (Leamer, 1978, 1983)
 - systematic specification searches: nonlinearities, interactions
- Data transformation
 - Combining different items of a survey to form a composite score
 - Winsorizing: outliers can be both problematic and helpful
- Good and bad sides of data mining
 - + learning from the analysis
 - inflating significance
- **Proper scientific reporting**
 - Extensive sensitivity analyses and robustness checks

A final observation