A pilot study exploring spreadsheet risks in scientific research

Simon Thorne
Ghada AlTarawneh

Neuroscience

- *“the study of the nervous system, how it affects behaviour, and how it is affected by disease. The goal of neuroscience is to define and understand the continuum from molecular to cell to behaviour”* (US Congress, 1984).
- A relatively new science using state of the art equipment for quantitative data capture and analysis
  - MRI, cellular imaging, computational modelling, molecular genetics, animal and human behaviour, psychophysics

Institute of Neuroscience
Newcastle University

- The department comprises 17 researchers at undergraduate, postgraduate and post-doctorial levels
- Largely quantitative experiments using various different measurement options
- Statistical analysis of data
- Publishing in leading journals: *Journal of Neuroscience, Brain and Language* and *Physics life reviews*.

Research Design

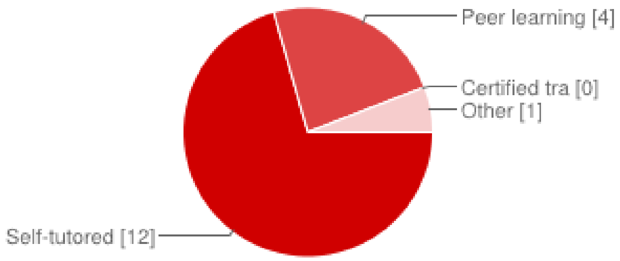
- Questionnaire distributed amongst all 17 members of the group
- Two in-depth interviews with the post-doctorial students
  - Post doctorial only since these members have the most experience and knowledge of how the lab runs

Questionnaire themes

Question area	Questions posed
Demographics	Age, Sex, Education and Role in the organisation
The Importance of research data	Frequency of spreadsheet use; size of spreadsheets in the department; number of users per spreadsheet and motivations for spreadsheet use.
Spreadsheet knowledge and experience	Methods of learning spreadsheets; self-assessed proficiency and willingness to train
Spreadsheets and other statistics software	How useful are spreadsheets for data analysis; other potential statistics software and personal advantages of spreadsheet software
The spreadsheet lifecycle	Approaches to design, separation of input, calculation and output; use of guidelines in development; approaches to testing; documentation
Spreadsheet backup and security	Organisational backup strategies; cell protection; password protection;

Results: Training and proficiency

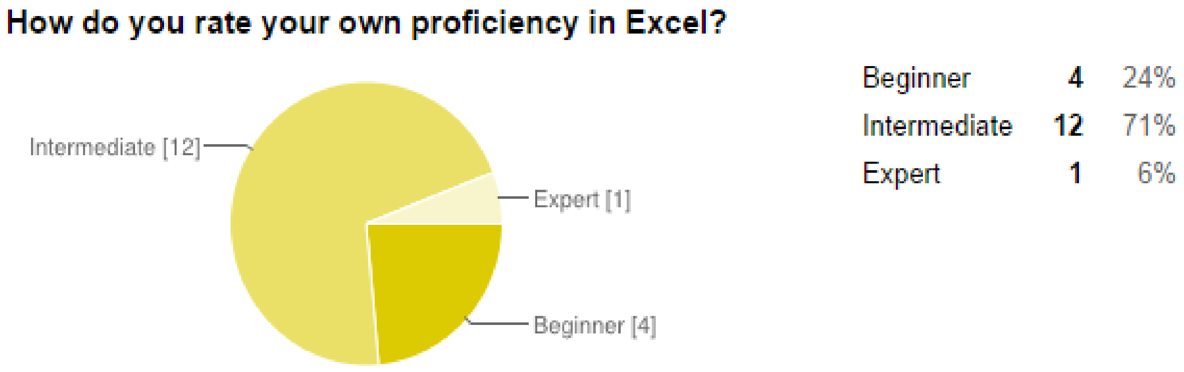
How did you learn how to use spreadsheets?



Self-tutored	12	71%
Peer learning	4	24%
Certified training courses	0	0%
Other	1	6%

- Most (71%) are ‘self-tutored’ which is typical of other spreadsheet surveys - Taylor *et al* 1998, SERP 2006

Proficiency



How do you rate your own proficiency in Excel?

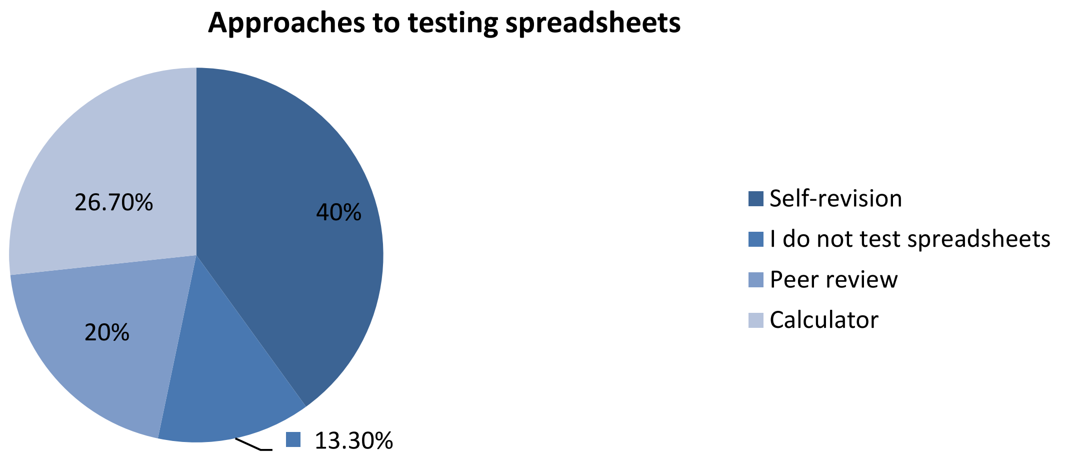
Beginner	4	24%
Intermediate	12	71%
Expert	1	6%

- 71% Intermediate, 24% Beginner, 6% Expert
- Most rate themselves as intermediate users
- Typical of other studies, SERP 2006
- Overconfidence could be an issue

Design

- Almost all of the researchers (94%) start designing their own spreadsheets by directly inputting data into the computer
- Typical of most spreadsheet user surveys (SERP 2006, Taylor *et al.* 1998)
- A particularly risky behaviour common amongst spreadsheet modellers

Testing



- Self revision is the most popular way of checking and correcting mistakes
- Research shows that self-auditing finds 60% of mistakes are (Panko, 2008)
- Team auditing being the best at correcting mistakes (80%) (Panko, 2008)

Documentation

- A third of the respondents didn’t document at all
  - A risky practice
- Half used cell comments to provide some remarks on the spreadsheet
  - The equivalent to annotated coding in software engineering.
  - Good practice but alone it is probably not enough
  - Also depends on the quality of those comments
  - Combined with a conceptual model, cell comments could be an highly effective approach to documenting

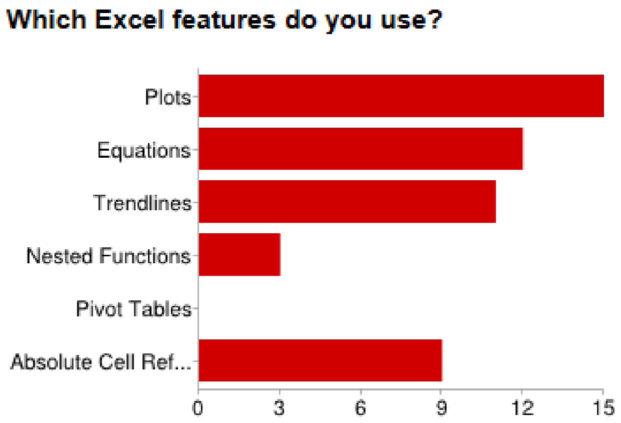
Security

- 82% didn’t use any form of security on their spreadsheets
- Those that did use passwords had them written down in the office
- Mitigated perhaps by the fact that the Neuroscience lab is a secure facility
- However, if spreadsheets are being transmitted electronically, they are vulnerable

Backup

- 82% of respondents took no measures to back up their work from their own machines
- A few indicated they used external storage and the universities shared infrastructure
- Mitigated by the universities backup plan, however it does not cover the hard disks of individuals, only shared drive space

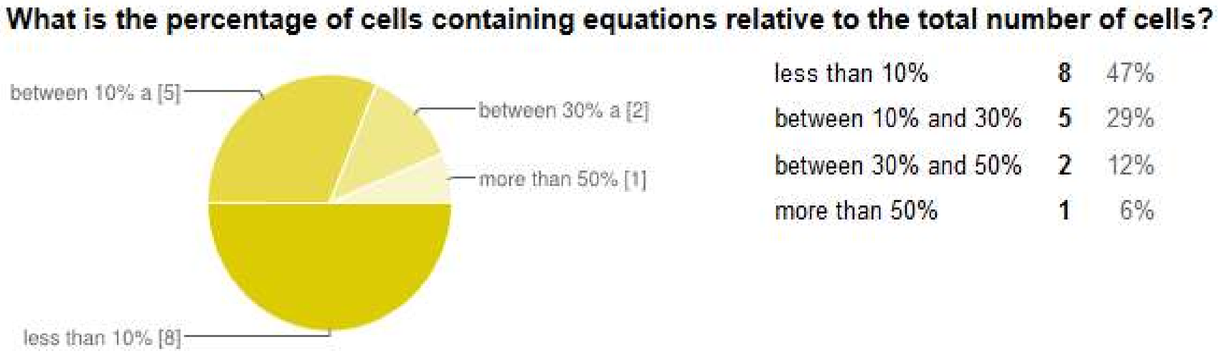
Functions used



Plots	**15**	88%
Equations	**12**	71%
Trendlines	**11**	65%
Nested Functions	**3**	18%
Pivot Tables	**0**	0%
Absolute Cell References (e.g. D3)	**9**	53%

- Fits broadly in line with other studies (Chan and Storey 1996, Ballinger *et al.* 2003, Thorne and Ball 2008)

Calculative cell to data ratio



What is the percentage of cells containing equations relative to the total number of cells?

less than 10%	**8**	47%
between 10% and 30%	**5**	29%
between 30% and 50%	**2**	12%
more than 50%	**1**	6%

- 47% indicate that the number of calculative cells is less than 10%
- Suggests that the models being produced are typical of field audits in composition (Panko, 2008)

Conclusions on data

- Although the sample is small, it would appear the activities at the neuroscience lab are broadly similar to other areas of spreadsheet activity
- Spreadsheets are the key tool in the capture and analysis of data
- The composition of spreadsheets seems broadly similar too (Chan and Storey 1996, Ballinger *et al.* 2003, Thorne and Ball 2008)

Board risks to the neuroscience lab

- Risks to the lab come in several guises
  - Lost data
    - Re-run of experiments, financial
    - Re-run of experiments, reduction in ‘measurable outputs’ (Publications), knock on effect with funding decision making
  - Erroneous analysis/data
    - Misleading journal articles, damaging to reputation of the institute, the individual and to the wider field
    - Retraction of journal/conference articles, damage to reputation of the institute, the individual and the wider field
    - Both could also have wider implications in the bidding for research monies
  - Fraud
    - Although rare, fraud is a concern for academic institutions. Very damaging to the scientific field and to the institutions involved

Conclusions

- The risks to the Neuroscience institute are not that far removed from business
  - Lost data, erroneous analysis, unsupported conclusions and fraud
- Incidence of committing errors must be broadly similar too
  - Lots of research shows that spreadsheet modellers tend to make the same mistakes regardless of ability or experience (Panko, 2008)
- Why then are retractions in academic journals very rare?
  - Where are all the academic spreadsheet horror stories?
  - Reinhardt and Rogoff

The academic safety net

- At the core of academic process is the idea of peer review.
  - We peer review all of the submissions to the EuSpRIG conference!
- Peer review means that research is assessed by a competent contemporary in a variety of settings with a variety of different possible outcomes
  - Formal and informal settings, outcomes vary depending on the context of the review

Peer review in the Neuroscience lab

- Peer review comes in a variety of forms at the Neuroscience institute
  - ‘Informal’ peer review for postgraduate students (PhD students)
    - Student – supervisor relationship, PhD student and a senior academic are paired up
    - The supervisor will extensively check the students data, analysis, conclusions and hypotheses
    - Mistakes in each area will be identified early on and corrective action can be advised

Peer review in the Neuroscience lab

- Peer review comes in a variety of forms at the Neuroscience institute
  - Collegiate review
    - Post doc researchers and senior academics will both check each others work and challenge each other over inconsistencies and possible erroneous data/analysis/conclusions/hypothesis
  - Co-authoring
    - Senior academics often work with several others to write and frame research
    - Co-authoring is like an on-going peer review through colleagues deconstructing and challenging each others analysis, assumptions and conclusions

External peer review

- In addition to internal process, work submitted to journals goes under another formal detailed critique
  - This process will be more detail than internal review processes
  - This may well include scrutiny of data, methods and the validity of the conclusions
  - Blind reviewed, the author and institute is unknown which allows for less chance for bias

How can business learn from this example?

- Peer review does exist as part of some spreadsheet modelling processes
  - Code-inspection has been shown to be effective at catching mistakes (although only 60% of them)
  - However, it is not enough to just consider the spreadsheet itself
  - One needs to understand the context and assumptions of the model
  - Of course, time is a significant limiting factor

How can business learn from this example?

- Business could have a central quality control aspect to their spreadsheet modelling activities
  - To reduce the time burden, this should also be linked to a risk assessment procedure for each spreadsheet written
  - If the spreadsheet is sufficiently risky, a peer review process should be instigated that examines the model and all of the assumptions that feed into this model

How can business learn from this example?

- Of course, the downside to such a system is time and cost.
- However, judging from the relatively few number of retractions in academic journals, this process does seem to work
- Reserving this process for the most risky is perhaps the best option for business
- Some of this may happen informally at present but a clear organisational framework could be a market leader

Conclusions

- Neuroscience broadly shares the same types of risk seen in other spreadsheet research
  - Errors, poor quality analysis, poor conclusions
- Broadly speaking, the spreadsheet models and modellers are similar to business too
  - Lack of training, lack of documentation, lack of testing
- However, the peer review process significantly reduces the likelihood of these mistakes being transmitted externally
  - A good model for reducing mistakes

Thank you

- Any questions?

- sthorne@cardiffmet.ac.uk

References

Taylor. M, Moynihan. P, Wood-Harper. T, (1998), *'End User Computing and information systems methodologies'*, Information systems Journal, **8**, pp 85-96

SERP. (2006). *'Spreadsheet Engineering Research Project'*. [Online] http://mba.tuck.dartmouth.edu/spreadsheet/, accessed 24/6/2015

Panko, R., (2008), 'What We Know About Spreadsheet errors'. *Journal of End User Computing,* 10(2), pp. 15-21.

Herndon, Thomas; Ash, Michael; Pollin, Robert. (2014). "Does High Public Debt Consistently Stifle Economic Growth? A Critique of Reinhart and Rogoff". *Cambridge Journal of Economics*.

Lacroix, Z. & Critchlow, T., (2003), '*Bioinformatics: Managing Scientific Data',* 1st ed, Morgan Kaufmann.