Running the REF on a rainy Sunday Afternoon:

The relative merits of peer review vs metrics in national research evaluations

Anne-Wil Harzing,
Middlesex University, London
Tilburg University, the Netherlands
www.harzing.com, @AWHarzing

Presentation outline

- General background of metrics vs. peer review
- The actual study: Running the REF on a rainy Sunday afternoon
  - Prior research, Methods
  - Results, Conclusions
- Wider reflections on metrics vs. peer review and the "national and cultural" embeddedness of research evaluation systems
- More detail and further reading in the hand-outs
  - Includes a summary of my papers on various data-sources and metrics in the hand-outs
  - Sorry for the self-promotion, but this is probably the only time ever I can afford attending a conference in this field ☺

Metrics vs. peer review: an increasing audit culture

- Increasing "audit culture" in academia, where universities, departments and individuals are constantly monitored and ranked
  - National research assessment exercises, such as the ERA (Australia) and the REF (UK), are becoming increasingly important
  - Unlike most European countries, both these national systems combine funding allocation with assessment of research quality in **one** and the same national evaluation

- Publications in these national exercises are normally assessed by peer review, esp. for SSH
  - The argument for not using citation metrics in SSH is typically that coverage for these disciplines is deemed insufficient in WoS and Scopus

What is the danger of peer review? (1)

- Peer review might lead to harsher verdicts than bibliometric evidence
  - especially for disciplines that do not have unified paradigms, such as the Social Sciences and Humanities
- In Australia (ERA 2010) the average rating for the Social Sciences was only about 60% of that of the Sciences and Life Sciences
  - despite the fact that on a cites-per-paper basis Australia's worldwide rank is similar in all disciplines
- The low ERA-ranking led to widespread popular commentary that government funding for the Social Sciences should be reduced or removed altogether
  - Similarly negative assessment of the credibility of SSH can be found in the UK (and no doubt in many other countries)

What is the danger of peer review? (2)

- More generally, peer review might lead to what I have called "promise over proof"
  - Harzing, A.W.; Mijnhardt, W. (2015) **Proof over promise: Towards a more inclusive ranking of Dutch academics in Economics & Business**, *Scientometrics,* vol. 102, no. 1, pp. 727-749

- Assessment of the quality of a publication might be (subconsciously) influenced by the "promise" of:
  - the journal in which it is published
  - the reputation of the author's affiliation, very problematic in Anglo countries that typically have **highly stratified** university systems: the "wrong" university automatically devalues your paper
  - the sub-discipline (theoretical/modeling vs. applied, hard vs. soft)
  - (or even) the gender and ethnicity of the author

What can we do?

- Remain critical about the increasing audit culture
  - But: be realistic, we are unlikely to see a reversal of this trend
- Raise awareness about
  - Alternative data sources for citation analysis that are more inclusive (e.g. including books, local & regional journals, reports, working papers)
  - Difficulty of comparing metrics across disciplines because of different publication and citation practices
- Investigate alternative data sources and metrics
  - Google Scholar, Microsoft Academic [Dimensions, Lens, Crossref]
  - hIa (Individual annualised h-index), i.e. h-index corrected for career length and number of co-authors
    - average number of single-author equivalent impactful publications published in a year (usually well below 1.0)

Running the REF on a rainy Sunday Afternoon

- Born out of sheer frustration about:
  - The amount of time wasted on REF related work and decision-making, which is crowding out mentoring and other more productive activities
    - Papers **already** peer-reviewed by expert journal reviewers are peer-reviewed **again** by non-expert colleagues and **again** by semi-expert external academics trying to all second guess **another** round of semi-expert peer-review by the REF panels
  - These REF panels are small and typically not very representative of the wider university sector and have to "burn their papers" after the event, leading to a lack of transparency
  - The misguided hero-worshipping of peer review, which in my view is confusing an idealised form of peer review with the hurried semi-expert peer review done by the REF panel

- Facilitated by the fact that
  - The new Microsoft Academic data source provides good coverage across disciplines (Harzing, 2016, Harzing & Alakangas, 2017a/b)
  - Publish or Perish has easy affiliation-level search for MA

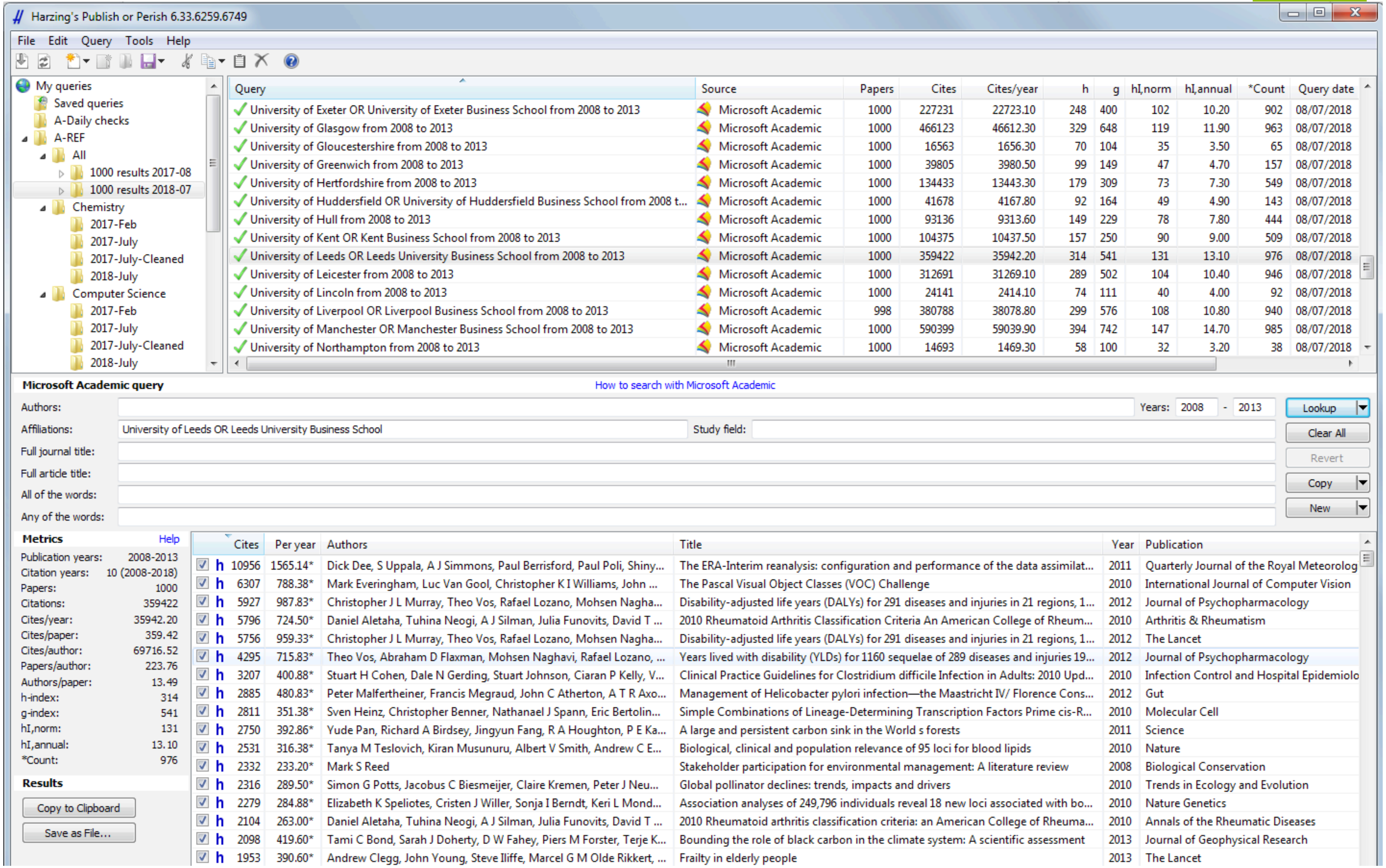
Prior research into peer review vs metrics

- Many earlier studies find strong correlations between peer review and citation rankings at an institutional level, but they:
  - Usually employed time-consuming data collection
  - Used WoS and Scopus, which do not offer sufficient coverage for the Social Sciences and Humanities
    - Recent study in Scientometrics based on Google Scholar (Mingers et al. 2017) used GS Profiles, but uptake of these is varied across institutions/disciplines
- I propose an analysis that **literally** can be done on a Sunday afternoon
  - Correlating MA total citations/hI-annual with REF Power rating
  - Proof-of-concept study that shows excellent potential
    - Fine-tuning can be done, this is really about flagging the possibility

Methods (1): Data collection

- Data collected with Publish or Perish using MA affiliation search
  - "All publications" search and "top-1000 publications only" (this literally took **only ½ hour** for the total sample after I had defined the queries!)
  - Used university variant names where needed
  - Gathered citations for publications between 2008-2013
  - Very minimal data cleaning needed
- Repeated the analysis after a year (on a very sunny Sunday afternoon)
  - Results substantively similar
  - Unlike peer review, bibliometrics analysis is not influenced by irrelevant variance e.g. the weather, lack of sleep, decision before/after lunch, bad temper, or anchoring effects
- For REF data I used the Power rather than Quality ranking
  - REF Power rating/ranking (size dependent) rather than Quality rating (size independent and heavily gamed)

A quick look at the data collection



Query	Source	Papers	Cites	Cites/year	h	g	hI,norm	hI,annual	*Count	Query date
✓ University of Exeter OR University of Exeter Business School from 2008 to 2013	Microsoft Academic	1000	227231	22723.10	248	400	102	10.20	902	08/07/2018
✓ University of Glasgow from 2008 to 2013	Microsoft Academic	1000	466123	46612.30	329	648	119	11.90	963	08/07/2018
✓ University of Gloucestershire from 2008 to 2013	Microsoft Academic	1000	16563	1656.30	70	104	35	3.50	65	08/07/2018
✓ University of Greenwich from 2008 to 2013	Microsoft Academic	1000	39805	3980.50	99	149	47	4.70	157	08/07/2018
✓ University of Hertfordshire from 2008 to 2013	Microsoft Academic	1000	134433	13443.30	179	309	73	7.30	549	08/07/2018
✓ University of Huddersfield OR University of Huddersfield Business School from 2008 t...	Microsoft Academic	1000	41678	4167.80	92	164	49	4.90	143	08/07/2018
✓ University of Hull from 2008 to 2013	Microsoft Academic	1000	93136	9313.60	149	229	78	7.80	444	08/07/2018
✓ University of Kent OR Kent Business School from 2008 to 2013	Microsoft Academic	1000	104375	10437.50	157	250	90	9.00	509	08/07/2018
✓ University of Leeds OR Leeds University Business School from 2008 to 2013	Microsoft Academic	1000	359422	35942.20	314	541	131	13.10	976	08/07/2018
✓ University of Leicester from 2008 to 2013	Microsoft Academic	1000	312691	31269.10	289	502	104	10.40	946	08/07/2018
✓ University of Lincoln from 2008 to 2013	Microsoft Academic	1000	24141	2414.10	74	111	40	4.00	92	08/07/2018
✓ University of Liverpool OR Liverpool Business School from 2008 to 2013	Microsoft Academic	998	380788	38078.80	299	576	108	10.80	940	08/07/2018
✓ University of Manchester OR Manchester Business School from 2008 to 2013	Microsoft Academic	1000	590399	59039.90	394	742	147	14.70	985	08/07/2018
✓ University of Northampton from 2008 to 2013	Microsoft Academic	1000	14693	1469.30	58	100	32	3.20	38	08/07/2018

Microsoft Academic query

How to search with Microsoft Academic

Authors:

Years: 2008 - 2013

Affiliations: University of Leeds OR Leeds University Business School

Study field:

Full journal title:

Full article title:

All of the words:

Any of the words:

Metrics Help

Publication years:	2008-2013
Citation years:	10 (2008-2018)
Papers:	1000
Citations:	359422
Cites/year:	35942.20
Cites/paper:	359.42
Cites/author:	69716.52
Papers/author:	223.76
Authors/paper:	13.49
h-index:	314
g-index:	541
hI,norm:	131
hI,annual:	13.10
*Count:	976

Results

		Cites	Per year	Authors	Title	Year	Publication
☑	h	10956	1565.14*	Dick Dee, S Uppala, A J Simmons, Paul Berrisford, Paul Poli, Shiny...	The ERA-Interim reanalysis: configuration and performance of the data assimilat...	2011	Quarterly Journal of the Royal Meteorolog
☑	h	6307	788.38*	Mark Everingham, Luc Van Gool, Christopher K I Williams, John ...	The Pascal Visual Object Classes (VOC) Challenge	2010	International Journal of Computer Vision
☑	h	5927	987.83*	Christopher J L Murray, Theo Vos, Rafael Lozano, Mohsen Nagha...	Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1...	2012	Journal of Psychopharmacology
☑	h	5796	724.50*	Daniel Aletaha, Tuhina Neogi, A J Silman, Julia Funovits, David T ...	2010 Rheumatoid Arthritis Classification Criteria An American College of Rheum...	2010	Arthritis & Rheumatism
☑	h	5756	959.33*	Christopher J L Murray, Theo Vos, Rafael Lozano, Mohsen Nagha...	Disability-adjusted life years (DALYs) for 291 diseases and injuries in 21 regions, 1...	2012	The Lancet
☑	h	4295	715.83*	Theo Vos, Abraham D Flaxman, Mohsen Naghavi, Rafael Lozano, ...	Years lived with disability (YLDs) for 1160 sequelae of 289 diseases and injuries 19...	2012	Journal of Psychopharmacology
☑	h	3207	400.88*	Stuart H Cohen, Dale N Gerding, Stuart Johnson, Ciaran P Kelly, V...	Clinical Practice Guidelines for Clostridium difficile Infection in Adults: 2010 Upd...	2010	Infection Control and Hospital Epidemiolo
☑	h	2885	480.83*	Peter Malfertheiner, Francis Megraud, John C Atherton, A T R Axo...	Management of Helicobacter pylori infection—the Maastricht IV/ Florence Cons...	2012	Gut
☑	h	2811	351.38*	Sven Heinz, Christopher Benner, Nathanael J Spann, Eric Bertolin...	Simple Combinations of Lineage-Determining Transcription Factors Prime cis-R...	2010	Molecular Cell
☑	h	2750	392.86*	Yude Pan, Richard A Birdsey, Jingyun Fang, R A Houghton, P E Ka...	A large and persistent carbon sink in the World s forests	2011	Science
☑	h	2531	316.38*	Tanya M Teslovich, Kiran Musunuru, Albert V Smith, Andrew C E...	Biological, clinical and population relevance of 95 loci for blood lipids	2010	Nature
☑	h	2332	233.20*	Mark S Reed	Stakeholder participation for environmental management: A literature review	2008	Biological Conservation
☑	h	2316	289.50*	Simon G Potts, Jacobus C Biesmeijer, Claire Kremen, Peter J Neu...	Global pollinator declines: trends, impacts and drivers	2010	Trends in Ecology and Evolution
☑	h	2279	284.88*	Elizabeth K Speliotes, Cristen J Willer, Sonja I Berndt, Keri L Mond...	Association analyses of 249,796 individuals reveal 18 new loci associated with bo...	2010	Nature Genetics
☑	h	2104	263.00*	Daniel Aletaha, Tuhina Neogi, A J Silman, Julia Funovits, David T ...	2010 Rheumatoid arthritis classification criteria: an American College of Rheuma...	2010	Annals of the Rheumatic Diseases
☑	h	2098	419.60*	Tami C Bond, Sarah J Doherty, D W Fahey, Piers M Forster, Terje K...	Bounding the role of black carbon in the climate system: A scientific assessment	2013	Journal of Geophysical Research
☑	h	1953	390.60*	Andrew Clegg, John Young, Steve Iliffe, Marcel G M Olde Rikkert, ...	Frailty in elderly people	2013	The Lancet

Methods (2): Differences in methods REF vs. MA citations

1. REF includes *non-academic impact* and *research environment*, my approach doesn’t (this could/should be evaluated separately!)
2. REF requires *disciplinary* choice (submit to specific UoA), my approach doesn’t, no problem with *multidisciplinary* research
3. REF includes a *selection* of academics, my approach includes *all* academics in the institution
4. REF includes only academics *employed* at the *census date*, my approach includes *all academics’* papers with university’s affiliation
5. REF includes max. *four publications per academic*, my approach includes *all publications*
6. REF output included mostly *journal publications*. My approach included *all publications*, incl. books, conference papers, software
7. REF allows publications *accepted*, my approach only includes *published* papers
8. REF was conducted in 2014, I counted citations in 2017/2018

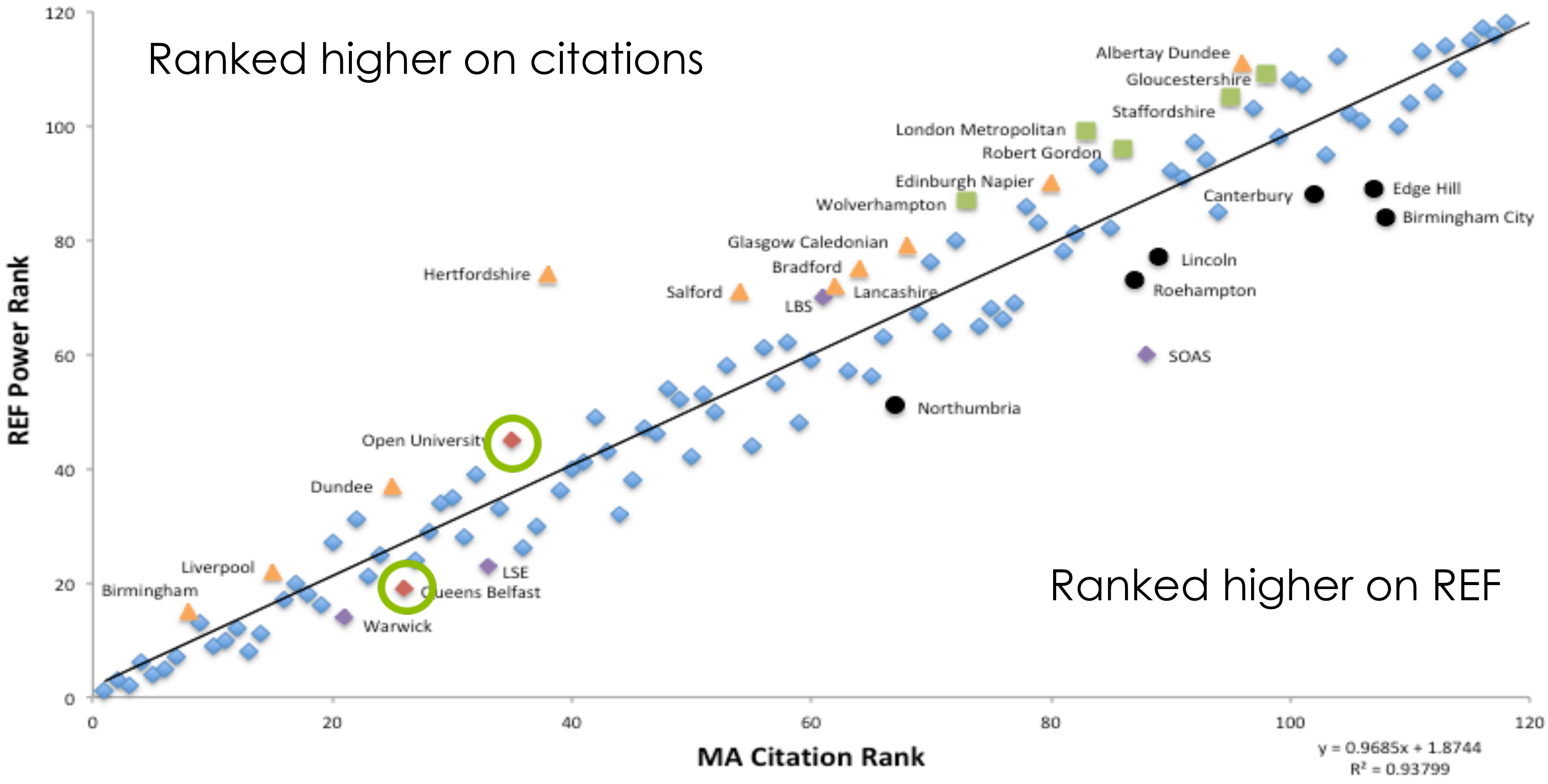
Results (1): High correlation between REF and citations

- Correlation of 0.97 between REF power rating (ranking) and MA citations (ranking)

- Most universities cluster around the regression line
  - Average difference 6.8 places out of 118 universities
  - However, there were some notable deviations

- Major deviations fall in three main categories
  1. MA errors [can probably be fixed], red diamonds
     - Problems in searching for some institutions: **Open University** (incl. Dutch and Israeli OU), **Queens University Belfast** (many pubs ascribed to Queens University); too many for OU, too few for QUB
     - Lack of affiliation data for a proportion of publications [fine as long as omission is not systematic]

REF power rank by MA citation rank

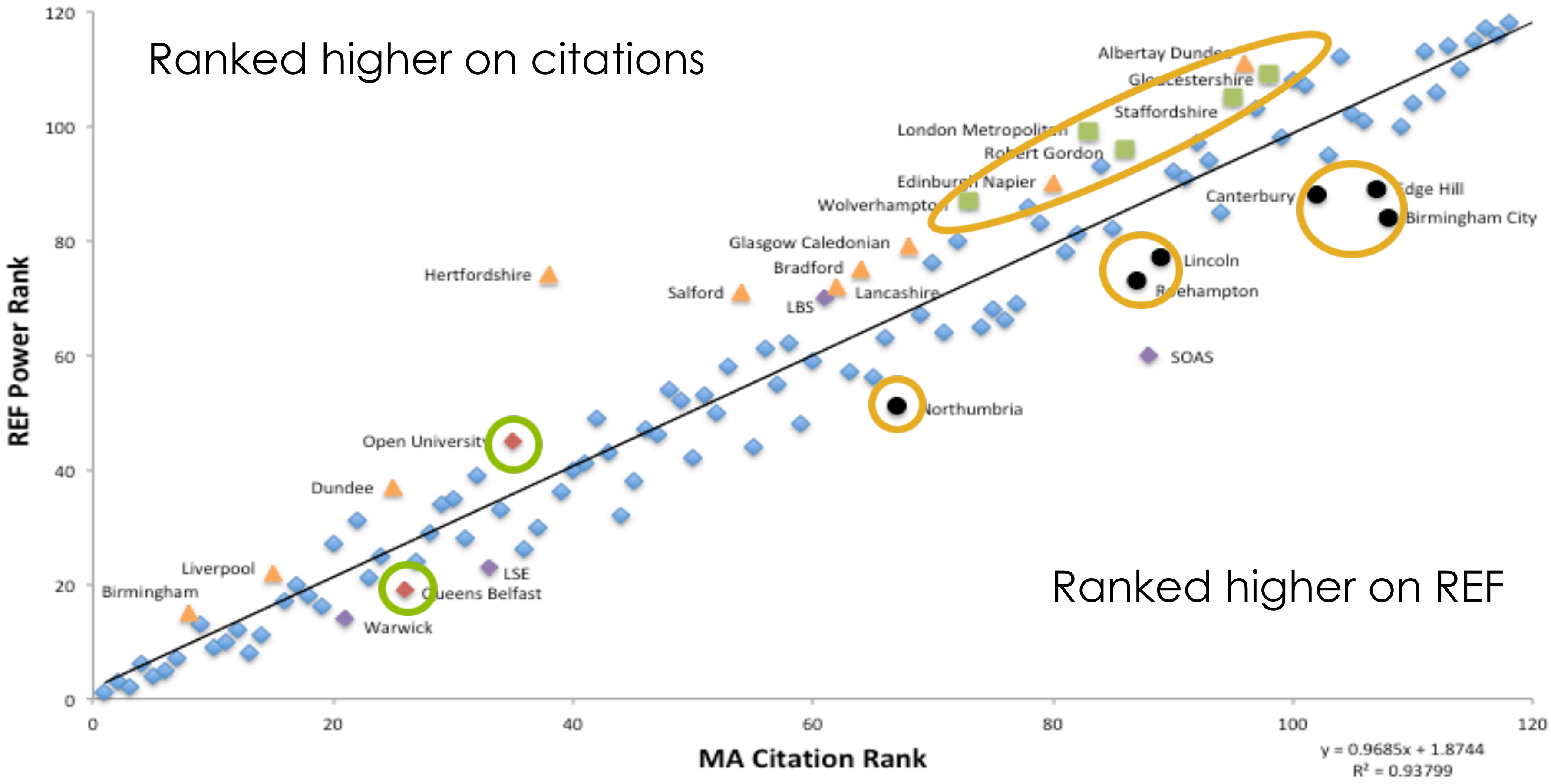


Results (2): deviation #2: post 92 universities

- One group [black circle] scores higher on REF ranking than on citation ranking
  - most likely caused by their scores on REF (societal) impact case studies
  - supported by the fact that most improved substantially since 2008 [when impact case studies were not included]

- Another group [green square] scores higher on citation ranking than on REF ranking
  - Citations might have been inflated because of "small numbers game"
    - individual highly-cited staff [e.g. Mike Thelwall]
    - highly cited textbooks

REF power rank by MA citation rank



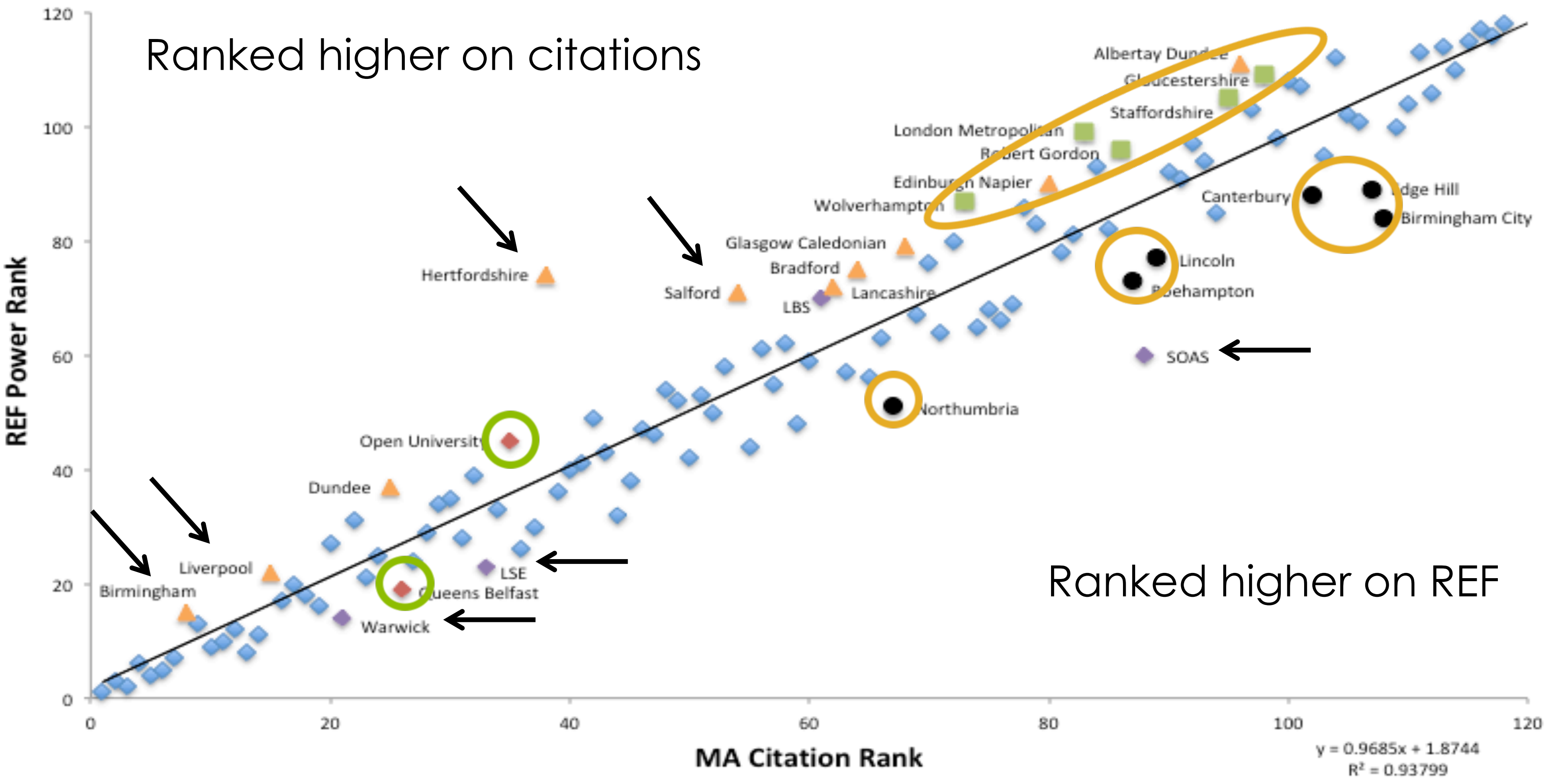
Ranked higher on citations

Ranked higher on REF

Results (3): deviation #3: Disciplinary differences

- Citation practices differ by discipline and cites are much higher in the (Life) Sciences than in the Social Sciences & Humanities
  - Universities with higher REF rank than citation rank [purple diamond]
    - tend to have more staff working in Social Sciences and Humanities
    - e.g. SOAS, LSE have a relatively low citation rank
  - Universities with higher citation rank than REF rank [orange triangle]
    - participation in huge consortia in e.g. particle physics or gene technology with highly cited papers
- Solution: use hIa or other discipline-corrected metric instead of raw citations
  - SOAS moves closer to regression line and LSE now ranks higher on metrics than on peer review

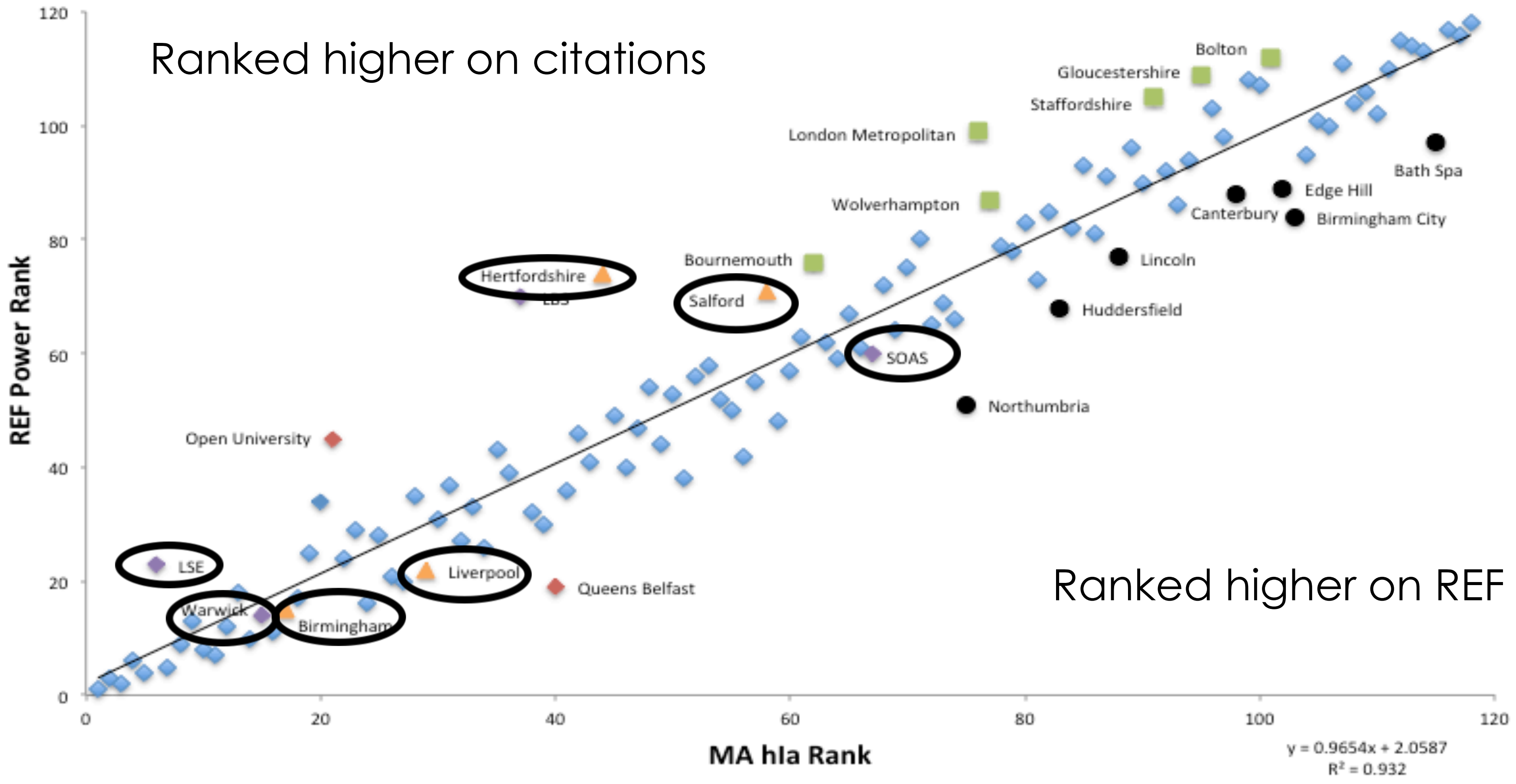
REF power rank by MA citation rank



Ranked higher on citations

Ranked higher on REF

REF power rank & MA hIa rank: Smaller disciplinary differences



Ranked higher on citations

Ranked higher on REF

Conclusion

- Peer review and metrics are highly correlated at the institutional level
  - Where differences occur these might be due to flaws in peer review just as much as flaws in metrics
- Consider separating research evaluation and funding allocation
  - The UK is one of the few countries that combines both in the same exercise
  - The two purposes are better served by different methods
    - Funding allocation can be done efficiently by metrics
    - Research evaluation is more suited to peer review, supported by metrics
  - Letting metrics do the "heavy lifting" saves time and money for a more meaningful evaluation of research quality than the current REF is able to offer

Recent evidence (1): The REF from an intl perspective

- Stern Review does **not** question the use of peer review for allocation of research funding
- Highlights five additional goals of the REF:
  - Informs strategic decision making
  - Informs local resource allocation
  - Provides accountability and transparency
  - Provides performance incentives
  - Contributes to the formation of the institution's reputation.
- "[…] **all of these goals could be reached without evaluating the performance of individual researchers**" as is currently done
- **"organizational-level evaluation with peer review as one of several tools could perhaps meet these goals even more efficiently and accurately"**

Sivertsen G (2017) Unique, but still best practice? The Research Excellence Framework (REF) from an international perspective. Palgrave Communications. 3:17078 doi: 10.1057/palcomms.2017.78.

Recent evidence (2): Knowledge Media Institute @ OU: 2nd study with MA data

- Pride & Knoth (2018) compared institutional GPA (app. Quality rating) with citations at the UoA level and concluded that:
  - "citation-based indicators are **sufficiently aligned with peer review** results at the **institutional level** to be used **to lessen the overall burden** of peer review on national evaluation exercises leading to **considerable cost savings**".

- Study is very critical of the hero-worshipping of peer review
  - *Several studies including The Metric Tide [4], The Stern Report [14] and the HEFCE pilot study [15] all state that metrics should be used as an additional component in research evaluation, with peer review remaining as the central pillar.*
  - *Yet, **peer review has been shown** by [16], [17] and [18] amongst others to **exhibit many forms of bias** including institutional bias, gender/age related bias and bias against interdisciplinary research.*
  - ***All of the above biases exist even when peer review is carried out to the highest international standards**. There were close to 1,000 peer review experts recruited by the REF, however the sheer volume of outputs requiring review calls into question the [... exactitude of the whole process.*

Pride, D., & Knoth, P. (2018). Peer review and citation data in predicting university rankings, a large-scale analysis. *arXiv preprint arXiv:1805.08529*.

Will anything change? Probably not: Individual push-back

- The research community as a whole doesn’t seem to support metrics; metrics tap into basic human fears and suffer from flaws of reasoning

- Fear of the unknown, many academics:
  - are not quantitatively minded and do not understand metrics [esp in SSH]
  - are convinced metrics don’t work in their fields (largely because they only know WoS and JIFs)

- Fear of “machines”, many academics:
  - have an (irrational) “fear of machines” and automation
  - prefer (flawed) human evaluation to (less flawed) automatic evaluation

- Flaws of reasoning
  - Level of analysis: peer review gold standard at individual level, aggregating this must surely be best for institutional/national-level evaluation?
  - Anecdata: reasoning from just **one** idiosyncratic example: my “best” paper isn’t highly cited, so..., I suspect he just cites his friends, so..., one of my citations is missing in GS/MA, so... we can’t use citations

Higher-level push-back (1)
The metric tide report

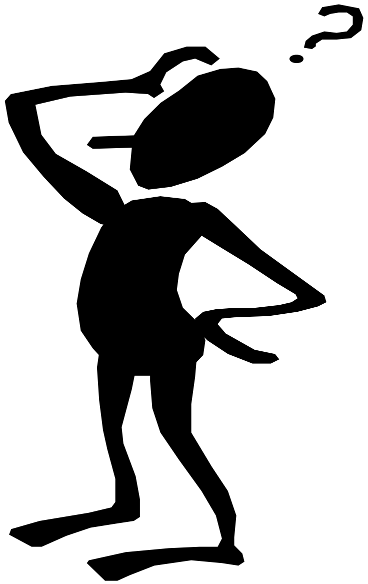
- The main push-back is a collectivized form of the individual concerns, based on the finding that **at a paper** level metrics correlate poorly with quality judgements
  - This is obviously well-known among bibliometricians
  - Metrics are meant for evaluation at higher levels of aggregation
- One of the Metric Tide's report main recommendations is:
  - Peer review is not perfect, but it is the least worst form of academic governance we have [note the implied comparison to democracy, this further legitimizes the choice], and should remain the primary basis for
  - assessing research papers [yes, absolutely]
  - research proposals [yes, sure thing]
  - and individuals [yes, obviously]
  - and for national assessment exercises like the REF **[no, not necessarily]**

Will anything change? Higher-level push-back (2)

- There are probably too many vested interests
  - Complete cotton-industry of consultancies supporting the REF submissions
  - Many (Research) Deans wouldn't know how to manage people without it ☺
  - Groups of academics who do well in the current prestige based system (4*/JoD publications) might not do as well in citations

- The current REF seems to fit the British [research] culture to a tee and might even [subconsciously!] tap into deeply held national cultural values ☺ ☺ ☺ [tongue-in-cheek, from someone who actually loves the British culture]
  - Path dependency + reluctance to change, which seems to suit the British sense of traditionalism and conserving the past [just re-watch Humphrey Appleby in Yes (Prime) Minister]
  - Reproduces the current "class system" [one of the most defining features of the British society] of universities nicely; who knows what metrics might bring?
  - Provides plenty of opportunity for ritualistic & heroic suffering and "muddling through", which the Brits seem to like so much
  - Supports the preferred reliance on gut feeling/negotiation/individual idiosyncracies over the more "Germanic" approach of hard data, systems, and structures

Thank you!



Any questions or comments?

My work on Google Scholar as a source for citation data

- Harzing, A.W.; Wal, R. van der (2008) **Google Scholar as a new source for citation analysis?**, *Ethics in Science and Environmental Politics*, 8(1): 62-71
- Harzing, A.W.; Wal, R. van der (2009) **A Google Scholar h-index for Journals: An alternative metric to measure journal impact in Economics & Business?**, *Journal of the American Society for Information Science and Technology*, 60(1): 41-46
- Harzing, A.W. (2013) **A preliminary test of Google Scholar as a source for citation data: A longitudinal study of Nobel Prize winners**, *Scientometrics*, 93(3): 1057-1075
- Harzing, A.W. (2014) **A longitudinal study of Google Scholar coverage between 2012 and 2013**, *Scientometrics*, 98(1): 565-575
- Harzing, A.W.; Alakangas, S. (2016) **Google Scholar, Scopus and the Web of Science: A longitudinal and cross-disciplinary comparison**, *Scientometrics*,106(2): 787-804

My work on Microsoft Academic

- Harzing, A.W. (2016) **Microsoft Academic (Search): a Phoenix arisen from the ashes?**, *Scientometrics,* 108(3):1637-1647
- Harzing, A.W.; Alakangas, S. (2017) **Microsoft Academic: Is the Phoenix getting wings?**, *Scientometrics,* vol. 110, no. 1, pp. 371-383
- Harzing, A.W.; Alakangas, S. (2017) **Microsoft Academic is one year old: the Phoenix is ready to leave the nest**, *Scientometrics*, vol. 112, no. 3, pp. 1887-1894.
- https://harzing.com/blog/2017/04/how-to-conduct-searches-with-microsoft-academic

My work on problems with the Web of Science

- Harzing, A.W. (2013) **Document categories in the ISI Web of Knowledge: Misunderstanding the Social Sciences?**, *Scientometrics,* 93(1): 23-34
- Harzing, A.W. (2015) **Health warning: Might contain multiple personalities. The problem of homonyms in Thomson Reuters Essential Science Indicators**, *Scientometrics,*105(3): 2259-2270
- https://harzing.com/blog/2016/09/how-to-get-listed-on-the-esi-ranking-of-highly-cited-authors
- https://harzing.com/blog/2017/02/web-of-science-to-be-robbed-of-10-years-of-citations-in-one-week
- https://harzing.com/blog/2017/09/bank-error-in-your-favour-how-to-gain-3000-citations-in-a-week

My work on new metrics

- Harzing, A.W.; Alakangas, S.; Adams, D. (2014) **hIa: An individual annual h-index to accommodate disciplinary and career length differences**, *Scientometrics*, 99(3): 811-821
- Harzing, A.W.; Mijnhardt, W. (2015) **Proof over promise: Towards a more inclusive ranking of Dutch academics in Economics & Business**, *Scientometrics*, 102(1): 727-749
- https://harzing.com/blog/2016/07/from-hindex-to-hia-the-ins-and-outs-of-research-metrics
- https://harzing.com/blog/2016/09/replication-study-gives-thumbs-up-for-the-individual-annual-hindex