How many citations? A comparison of Web of Science and Google Scholar

Michael E Dewey*

July 2010†

1 Introduction

Aim To compare various methods for obtaining citation counts

Objectives To compare:

- Web of Science and Google Scholar as sources of citation information
- To examine the influence of article characteristics on any differences found

There are three main sources of citation information: ISI which is usually accessed via Web of Science, Scopus and Google Scholar. This document compares ISI with Google Scholar.

Increasing interest in using citation counts as a proxy for scientific productivity and impact has also been followed by examination of the properties of the competing sources of that information.

*Section of Epidemiology, PO 60, Institute of Psychiatry, De Crespigny Park, London, SE5 8AF, UK. mailto:m.dewey@iop.kcl.ac.uk

†Printed August 2, 2010 at 11:3

2 Methods

I will compare two of the sources: ISI which is a commercial product and Google Scholar which is freely available. Although free it is not open source and the coverage, updating frequency and algorithms used are not published. I accessed ISI via two proprietary products: Web of Science and ResearcherID. I accessed Google Scholar via the free software Publish or Perish available from `http://www.harzing.com/pop.htm` and also directly.

I accessed the sources in July 2010 and searched for articles bearing my name or articles for which I am a member of a corporate group.

I modelled the log of the ratio of citations from the two sources using study field (mental health, neurology, other health), study type (survey, cohort, systematic review, collaborative reanalysis, trial, experiment, measurement) and age (number of years since publication, zero equals 2010) as covariates. The choice of study field is in some cases quite arbitrary. I have included studies on dementia as mental health and studies on stroke as neurology. Both of these choices could be argued. Some things are difficult to classify, sleep for example, and I have classified these according to overall study aim.

For articles with small numbers of citations the ratio is not very stable and probably not very interesting either so for my analyses I restricted to those cited at least five times according to both ISI and Google Scholar. This also removes the influence of articles which for one reason or another cannot be found in one database.

3 Results

4 The data

Figure 1 shows that overall Google Scholar returns more citations and that this effect is more pronounced for more recent articles at least for mental health and neurology.

4.1 Modelling

The effects shown in Figure 1 can be confirmed by formal modelling using linear regression. I did not weight the studies although ratios based on smaller

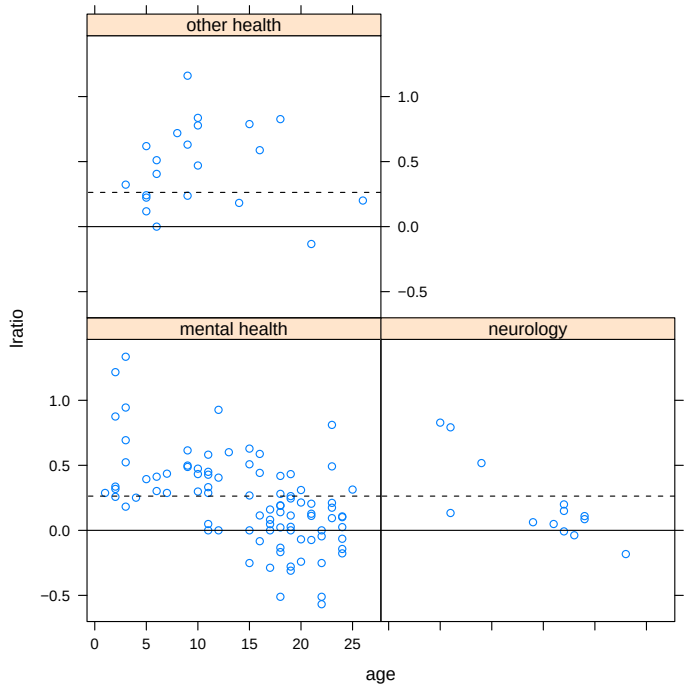

Figure 1: Ratio (on log scale) of Google Scholar citations to ISI citations by age of publication. The solid horizontal line represents equality, the dashed line the over all mean

numbers would be more variable.

	Coeff	SE	cib	ciub
mental health	0.645	0.073	0.501	0.789
mental health:age	-0.028	0.004	-0.037	-0.019
neurology	0.834	0.218	0.401	1.266
neurology:age	-0.044	0.014	-0.072	-0.016
other health	0.486	0.127	0.235	0.738
other health:age	-0.002	0.011	-0.024	0.019

Table 1: Coefficients from linear regression showing separate intercepts and slopes for the three study fields.

Study type does not seem to have any influence on the ratio and was removed from the model. There is evidence of an interaction between study field and years since publication and the final model chosen has separate intercepts and slopes for each study field. Results of the modelling are shown in Table 1. So for example the estimated log ratio for a mental health article of age

zero is 0.645 and it declines each year by 0.028.

There was no evidence of non–linearity in additional models fitted using fractional polynomials (Royston and Altman, 1994). Diagnostics reveal no unusual features.

4.2 Article retrieval

Not all papers can be retrieved.

Difficulties with ResearcherID

For instance Elkan et al. (2000) is not indexed in Web of Science under my name although it is indexed under the first author's name. I am an author, a fact which can be checked on the HTA website, but this article cannot be linked to my name in ResearcherID. There are a number of other examples which have in common that the View record field in Web of Science is not available. They amount to about 5% of the articles which I have co–authored.

Difficulties with Google Scholar

Google Scholar relies on Google's search and can only find what it can recognise. Some issues are that articles may have their author lists truncated (at either end), the end of the author list may be seen as the start of the title, the year of publication may be incorrect (for some reason an issue for me with Psychological Medicine).

Group authorship

Neither system deals with corporate authors without some manual assistance. Once found they can be added to ResearcherID. They have to be found separately in Google Scholar each time.

Variant forms

Both systems sometimes find the same article under different variants of the citation. In Web of Science this may reflect failures by original authors to cite the article correctly. Google scholar relies on automatic methods and

these give rise to other problems. One thing which is quite common is for corporate authors to be seen as part of the title perhaps because they do not look like human author names. For instance Leonardi-Bee et al. (2005) is found 71 times with a correct citation, 14 with the corporate author as the first part of the title and a further twice.

4.3 Which papers stand out?

Table 2 shows the ten articles with the largest residual from the model but showing only those from the last ten years. Using the model residual to select means that the effect of article age can be removed, these are not the articles with the largest ratio.

Article	ISI	Google	log ratio	residual
Walker et al. (2004)	24	53	0.79	0.22
Chilvers et al. (2001)	93	172	0.61	0.22
Siriwardena et al. (2002)	19	39	0.72	0.25
Dias et al. (2008)	5	12	0.88	0.29
Elkan et al. (2000)	62	135	0.78	0.31
Kendrick et al. (2000)	39	90	0.84	0.37
Castro-Costa et al. (2007)	7	18	0.94	0.38
Rodriguez et al. (2008)	8	27	1.22	0.63
Elkan et al. (2001)	73	233	1.16	0.69
McDougall et al. (2007)	5	19	1.34	0.77

Table 2: The ten articles with the largest residuals from the model from the last ten years

Coverage

Some of the differences may be due to different coverage. Google Scholar also counts citations from books and reports but my impression is that many of the extra citations are from journals which ISI does not search.

Citations from books and reports are found by Google Scholar. For instance Walker et al. (2004) is cited in various guidelines and ISI finds about half as many as Google Scholar as they are not all journal publicatons.

Google Scholar seems to find more language other than English citations. For instance it finds citations in at least seven other languages for Walker et al. (2004) although my impression is that some of these may be translations of an original.

Google Scholar clearly does limit the scope of the search. For instance I have all my publications listed on my web page but these are not found by Google Scholar.

The h–index

A scientist has index h if h of his or her N_p papers have at least h citations each, and the other $(N_p - h)$ papers have at most h citations each (Hirsch, 2005),

Calculating these from the various sources gives different values as shown in Table 3.

Method	h
ResearcherID	41
Plus manual searches using Web of Science	42
Publish or Perish	46
Plus manual search fro group authroship	48

Table 3: My h–index according to various sources of citation counts

I conclude that the idea of a single h–index is difficult to maintain.

5 Discussion

Much of what I have found concords with previous work. A summary can be found in Harzing and van der Wal (2008) and I shall not repeat it here. Finding more in Google Scholar is well known my results are different from the suggestion that the effect is greater in social science but not found in natural and health sciences (Harzing and van der Wal, 2008).

My impression looking at the ten articles with the largest residuals from the model is that these include some of the articles which I would have expected to have the greatest policy implications. There is no very good way of measuring this but Elkan et al. (2000, 2001); Kendrick et al. (2000) all came from a review of the evidence base for domicilary health visiting, a topic of considerable policy interest at the time, Walker et al. (2004) is a review relevant to community therapy services for stroke and Chilvers et al. (2001); Siriwardena et al. (2002); Dias et al. (2008) are all trials in areas where the evidence base was quite limited at the time, primary care treatment of

depression, educational interventions for prescribing, and dementia care in low and middle income countries respectively. If this is a more general finding then using ISI leads to a bias against such articles.

References

E Castro-Costa, M E Dewey, R Stewart, S Banerjee, F Huppert, C Mendonca-Lima, C Bula, F Reischies, J Wancata, K Ritchie, M Tsolaki, R Mateos, and M Prince. Prevalence of depressive symptoms and syndromes in later life in ten European countries. *British Journal of Psychiatry*, 191:393–401, 2007.

C Chilvers, M E Dewey, K Fielding, V Gretton, P Miller, B Palmer, D Weller, R Churchill, I Williams, N Bedi, C Duggan, A Lee, G Harrison, and the Counselling versus Antidepressants in Primary Care Study Group. Antidepressant drugs and generic counselling for treatment of major depression in primary care: randomised trial with patient preference arms. *British Medical Journal*, 322:772–775, 2001.

A Dias, M E Dewey, J D'Souza, R Dhume, D D Motghare, K S Shaji, R Menon, M Prince, and V Patel. The effectiveness of a home care program for supporting caregivers of persons with dementia in developing countries: a randomised controlled trial from Goa, India. *PLoS ONE*, 3:e2333, 2008.

R Elkan, D Kendrick, M Hewitt, J J A Robinson, K Tolley, M Blair, M E Dewey, D Williams, and K Brummell. The effectiveness of domiciliary health visiting: a systematic review of international studies and a selective review of the British literature. *Health Technology Assessment*, 4(13), 2000.

R Elkan, D Kendrick, M E Dewey, M Hewitt, J Robinson, M Blair, D Williams, and K Brummell. Effectiveness of home based support for older people: systematic review and meta-analysis. *British Medical Journal*, 323:719–725, 2001.

A-W Harzing and R van der Wal. Google Scholar: the democratization of citation analysis. *Ethics in Science and Environmental Politics*, 8:62–71, 2008.

J E Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102:16569–16572, 2005.

D Kendrick, R Elkan, M Hewitt, M E Dewey, M Blair, J Robinson, D Williams, and K Brummell. Does home visiting improve parenting and

the quality of the home environment? A systematic review and meta-analysis. *Archives of Disease in Childhood*, 82:443–451, 2000.

J Leonardi-Bee, P M W Bath, M-G Bousser, A Davalos, H-C Diener, B Guiraud-Chaumeil, J Sivenius, F Yatsu, M E Dewey, and the Dipyridamole in Stroke Collaboration (DISC). Dipyridamole for preventing recurrent stroke and other vascular events: a meta–analysis of individual patient data from randomised controlled trials. *Stroke*, 36:162–168, 2005.

F A McDougall, K Kvaal, F E Matthews, E Paykel, P B Jones, M E Dewey, C Brayne, and CFAS. Prevalence of depression in older people in England and Wales: the MRC CFAS study. *Psychological Medicine*, 37:1787–1795, 2007.

J L L Rodriguez, C P Ferri, D Acosta, M Guerra, Y-Q Huang, K S Jacob, E S Krishnamoorthy, A Salas, A L Sosa, I Acosta, M E Dewey, C Gaona, A T Jotheeswaran, S-R Li, D Rodriguez, G Rodriguez, P Senthil Kumar, A Valhuerdi, M Prince, and for the 10/66 Dementia Research Group. Prevalence of dementia in Latin America, India and China. A 10/66 Dementia Research Group population-based survey. *Lancet*, 372:464–474, 2008.

P Royston and D G Altman. Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. *Applied Statistics*, 43:429–467, 1994.

A N Siriwardena, A Rashid, M R D Johnson, and M E Dewey. Cluster randomised controlled trial of an educational outreach visit to improve influenza and penumococcal immunisation rates in primary care. *British Journal of General Practice*, 52:735–740, 2002.

M F Walker, J Leonardi-Bee, P Bath, P Langhorne, M E Dewey, S Corr, A Drummond, L Gilbertson, J R F Gladman, L Jongbloed, P Logan, and C Parker. An individual patient data meta-analysis of randomised controlled trials of community occupational therapy for stroke patients. *Stroke*, 35:2226–2232, 2004.