Using Google Scholar in research evaluation of social science programs, with a comparison with Web of Science data

Ad Prins[*], Rodrigo Costas[**], Thed van Leeuwen[**] and Paul Wouters[**]

[*] info@adprins.nl
Ad Prins Support in ResearchManagement, Lage der A 31, Groningen, 9718 BL (The Netherlands)

[**] rcostas@cwts.leidenuniv.nl; leeuwen@cwts.nl; p.f.wouters@cwts.leidenuniv.nl
CTWS Centre for Science and Technology Studies, Leiden University, Wassenaarseweg 62A,
Leiden, 2333 AL (The Netherlands)

Introduction

In the Netherlands, research evaluation is organized under the combined responsibility of the Association of Universities (VSNU), the National Research Council (NOW), and the Royal Academy of Arts and Sciences (KNAW). Their combined responsibility results in a Standard Evaluation Protocol, (SEP) that describes in detail the organization of research assessment, the various aspects taken care off during research assessments, and the indicators that should be part of the reporting by the committee. In the assessment cycles, that cover a six year time span, including an international assessment as well as an internal mid-term review, peer review is the guiding principle.

In the SEP, the usage of quantitative measures, such as bibliometric indicators is not compulsory, however, in many assessment cycles mainly in the natural, life and medical sciences, bibliometric indicators are introduced to support the work of the review committee. As it is well known that in the social sciences, the humanities, and law, the application of bibliometric indicators is of relatively lesser value, due to the lower degree of coverage of the systems that form the basis for bibliometric analyses (van Leeuwen, 2013), in most of the Dutch assessments in the SSH and Law domains bibliometrics was not applied. In the past, the field of psychology applied bibliometrics, just as the fields of economics and business & management. (Nederhof, 2006) These fields stand out among the SSH and Law domains, as the communication among scholars in these domains has shifted more and more towards journals publications. However, from the SSH domains a strong concern with respect to the design and organization of research assessment has led to the report "Judging research on its' merits" (KNAW, 2005), which initiated a further thinking among the scholars in these domains on how to further elaborate the preferred way of assessing research in the SSH and Law domains. Two advisory councils were installed, and this led to two reports, one for the humanities (KNAW, 2012), and one for the social sciences (KNAW, 2013). These two reports have strongly influenced the new SEP, that has to be applied from 2015 onwards. An important shift in this new SEP is a lesser focus on productivity, and a wider focus on the impact of scholarly activities, not only in the scientific realm, but also on societal impact.

As the publication cultures differ in the social sciences and humanities (Hicks, 2004, Nederhof et al, 2010, and van Leeuwen et al, forthcoming), impact cannot be established in the regular, journal based electronic databases normally used for bibliometrics (e.g. Web of Science or Scopus). An alternative for the traditional journal-based systems is Google Scholar (hereafter

referred as GS). Although this system has been studied before (Harzing 2008, Kousha 2008, 2011), to the best of our knowledge there are no examples of its use in a real life assessment procedure.

In this paper we report on the application of GS based metric in the formal assessment of research programs in the fields of Education and Pedagogical Sciences, and in Anthropology, and on a meta-analysis on the comparison of the results based on Google Scholar and WoS.[1] Finally we discuss some issues with regard to methods in relation to the context of the assessments.

Data and methods
The assignments by deans of the participating faculties to use GS in an evaluative bibliometric context was proposed due to concerns about the representation of SSH outputs in Web of Science (hereafter referred to as WoS)[1].

Education and Pedagogical Sciences (hereafter referred as Ed/Ped) comprised 13 programs of six universities over the evaluation period 2006-2011. Anthropology comprised five programs of an equal number of universities, over the evaluation period 2004-2012. The selection of publications differs slightly for both cases. In the case of Ed/Ped, each program was asked to send in 10 publications per year (60 publications per program). Program directors were asked to send in highly valued or highly cited publications, possibly including also books. In the case of Anthropology, the selection was based on $5 - 10$ most cited publications for each year to be evaluated, related to the size of the program. A reduction of numbers of selected publications was chosen for, assuming that small programs are less likely than large programs to produce equal numbers of highly cited publications. Checking for publications that were listed in more than one program as double entries, the resulting numbers of publications were 774 for Ed/Ped (with 6 double entries) and 328 for Anthropology (four double entries).

Data collection for the publications was based on keywords of title and author, allowing for various spellings. As doubts has been cast on the reliability of GS information (Jacso, 2012) information was retrieved for the full second order GS citing data (i.e. enabling a check on the citing sources). The selection criterion was that the citing source should be verifiable, meaning that the source should be traceable in terms of a proper working URL of websites of journals, publishers or other location. Other citing sources, in particular those without proper URL, might still be valid if checked individually, but were nevertheless taken out of the data set. This was also the case with sources with defective data such as improper year of reference in comparison to the publication date of the cited reference. The net certified citations were 22887 (89,8% of gross total of Ed/Ped), and 8092 (89,7% of gross total for Anthropology).

In a further analysis of data quality, performed during the later meta analysis, specific sources of the citing data have been investigated. This analysis was based on the specific URL of each of the citing publications. Information provided by GS is based on the indexes produced by crawling specific internet sources such as electronic academic journals, academic books (Google Books), websites of academic publishers, and internet repositories such as www.jstor.org, www.cairn.info, http://papers.ssrn.com or www.academia.edu. GS indexes also university

---

[1] It is important to mention here the involvement of the research directors of the faculties as stakeholders within their field(s) of expertise in choosing a selection base for the publications to be analyzed.

libraries, as well as academic societies, governments and other sources. The majority of these sources contain verifiable meta data, either the proper (post print) academic publication itself, or the meta data of pre-prints or the version of record available repositories or university libraries. However, as is also noted by Jacso, some sources such as university libraries may contain also other referring publications, such as theses by PhD's or master theses and repositories might also include conference papers and reports by research institutes. These citing sources might therefore be considered to have a somewhat wider range of reliability in terms of academic status. The meta data were therefore classified, based on the available URL revealing characteristics of the citing source, such as the publisher or university. Classification was possible for the majority of the URL's, as these frequently shared common characteristics such as websites identifiably owned by publishers like Sage, Elsevier, Oxford or Cambridge, or from university libraries. The second order data have been classified as coming from (a) verifiable academic sources (including academic journals, academic publishers of volumes etc. and academic books), (b) university libraries, (c) repositories not identifiable as university libraries, (d) other sources than the above, including academic societies, government sites, blogs and personal webpages of researchers. In a number of cases, the available URL did not share common characteristics, occurring only once in the database. For efficiency reasons these were classified as "unknown".

Results

Table 1 shows that in both fields the academic citing sources account for more than half of the total volume of cites. Also, the volume of "unknown" sources, which may contain both "exceptional" and customary citing sources, is fairly low (11% and 7%).

Table 1: Sources for citations in GS for Anthropology and Ed/Ped

Source	Anthrop #	Anthrop %	Ed/Ped #	Ed/Ped %
Academic sources	4573	56,5%	14470	63,2%
University Libraries	1677	20,7%	4573	20,0%
repository other than ULs	616	7,6%	1236	5,4%
Other Sources	327	4,0%	1085	4,7%
Unknown	899	11,1%	1523	6,7%
Total	8092	100,0%	22887	100,0%

The results also show that the fields of Ed/Ped and Anthropology differ with regards to the document types of the cited output. Whereas in Ed/Ped the share of journal articles in the total set of publications is almost 90%, in the case of Anthropology the share of journal articles is 58%, with higher percentages for books, volumes and chapters. Differences in publication cultures are even more apparent in the volume of citations per publication types. Whereas in the case of Ed/Ped journal articles on average are the most cited publication type, in Anthropology books are the more cited type (table 2). These differences are not due to a single or a few outliers, as figures 1 and 2 show the median volume of cites per document type is higher for books in Anthropology and higher for journal articles in Ed/Ped. Even though books in Ed/Ped receive considerable

attention, for the Anthropology programs books are both important forms of output as well as important means for receiving scientific impact.

Table 2: Citations per cited document type

	Anthropology			Ed/Ped		
	Cites #	Pubs #	Cites p publ	Cites #	Pubs #	Cites p publ
Books (Monographs)	1818	44	41,3	700	28	25,0
Chapters in volume	613	38	16,1	788	42	18,8
Journal articles	3885	187	20,8	21256	695	30,6
Other	357	19	18,8	109	7	15,6
Edited Volumes	1419	39	36,4	34	2	17,0
Total	8092	327	24,7	22887	774	29,6

Figure 1: Distribution of cites per document type in Anthropology
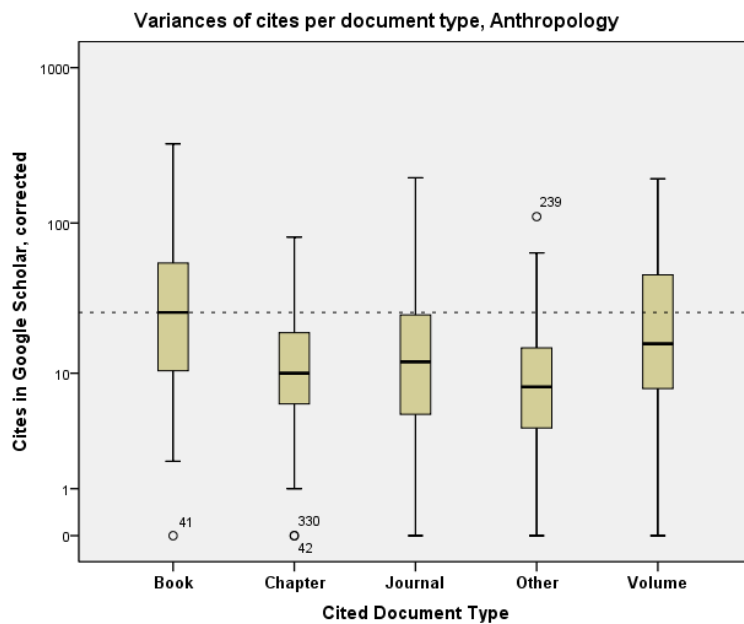
Figure 2: Distribution of cites per document type in Education and Pedagogical Sciences

Variances of cites per document type in Ed/Ped
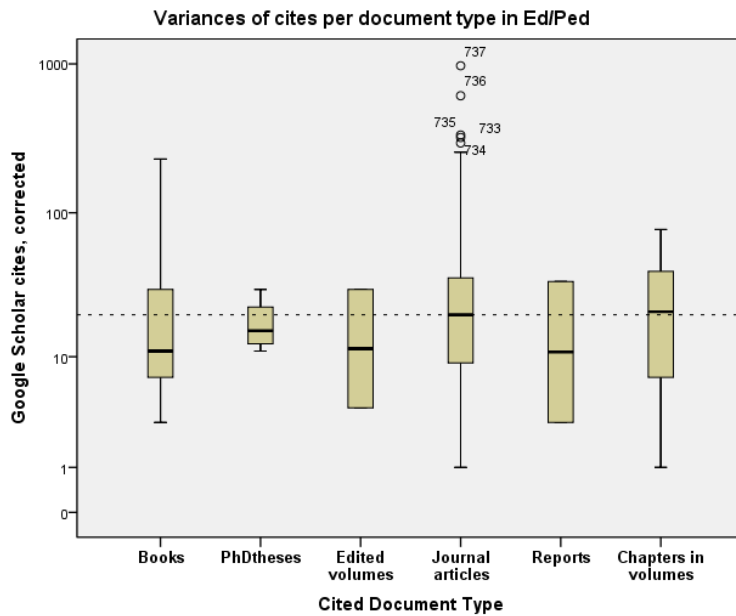
A comparison of the results for GS with data retrieved from WoS shows large differences for the studied publications in their coverage. Whereas over 80% of the publications in most programs in Ed/Ped were published in journals covered by WoS, this percentage fell to an average of 37.5% for the programs in Anthropology. Relatively lower coverage were also noted for the two programs in Ed/Ped on theory, history and philosophy in Ed/Ped, represented as extreme values with O in figure 3. (figure 3)

Figure 3: WoS Coverage of publications per program in three fields

Coverage of Program's Publications in WoS
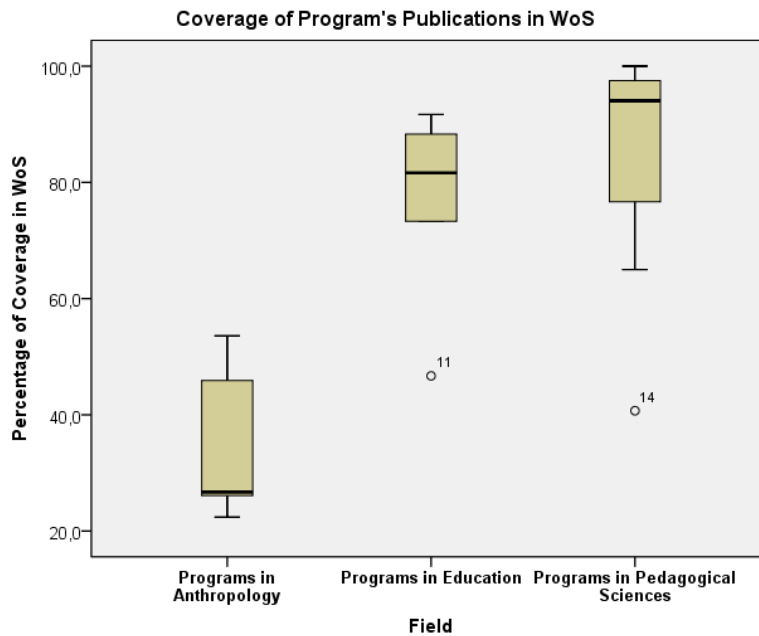
A further comparison of citations from GS with those from WoS indicate higher levels of citation information provided by GS for both fields. This is also true if only GS citations from identifiably academic sources, such as academic journals, publishers and books are considered (Table 3).

Table 3: Total cites in Google Scholar and Web of Science for two fields

Field	Cites in Google Scholar	Cites in Google Scholar, Academic sources	Cites in WoS
Anthropology	8092	4573	1097
Education & Pedagogical Sciences	22887	14470	8870

The numbers of citations per publications show a fair correlation for GS and WoS both in Ed/Ped and Anthropology. However, the correlation for Anthropology is based on a strongly reduced set, as only 37.5% of the publications were covered in WoS. In figures 4 and 5, the individual programs are correlated and the differences observed among them lead to the conclusion that programs vary with regard to how their citations are calculated based on GS or WoS.

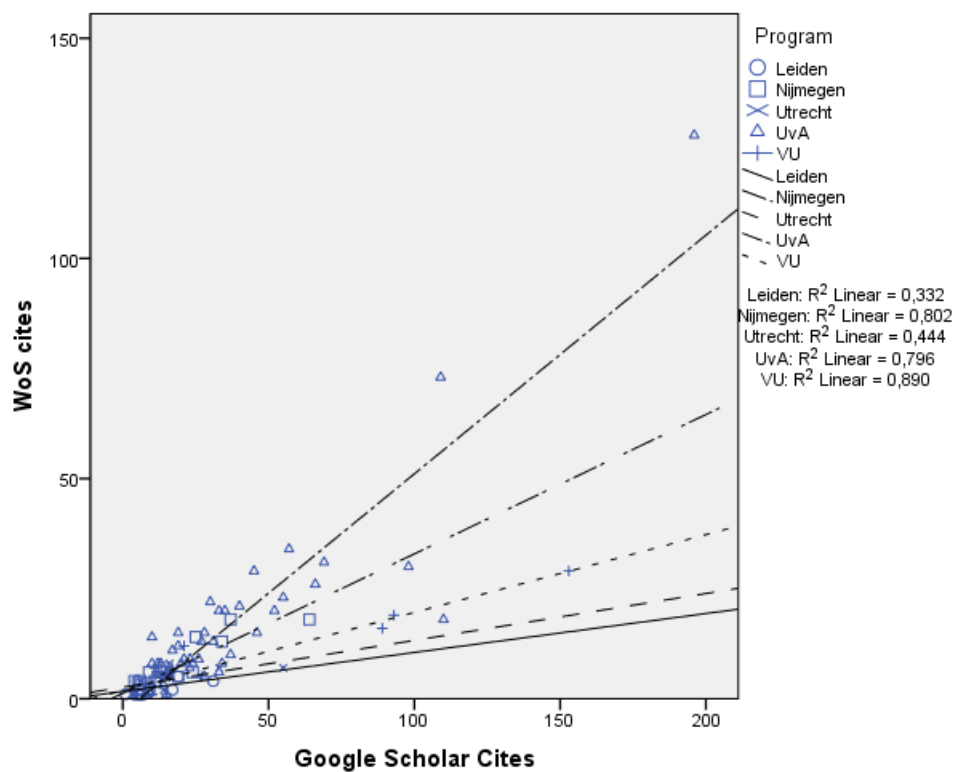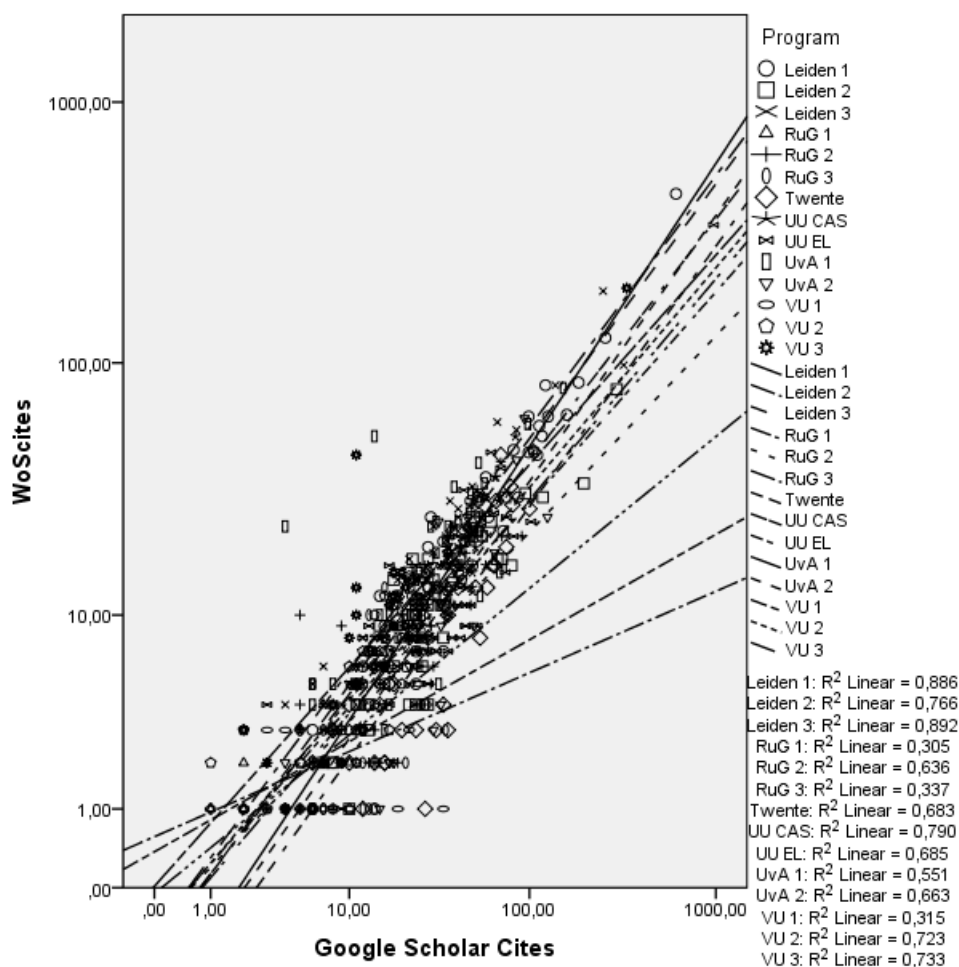Figure 4 Scatterplot WoS and GS citations for programs in Anthropology

Figure 5: Scatterplot WoS and GS citations for programs in Ed/Ped

Discussion

Our results indicate that indeed it is possible to perform bibliometric studies for evaluation purposes using GS, both with regards to data collection and data reliability, once data are based on selected publications and cleaned for erroneous data.

The comparison of GS results with WoS results indicates that it is fruitful to use GS for fields with lower degrees of coverage in WoS (Van Leeuwen, 2013), in particular fields that produce more diverse types of output than articles in journals included in WoS. As we show, in Ed/Ped and even more so in Anthropology other types of publications are important means of communication, receiving considerable impact according to GS which is missing in WoS.

In contrast to claims by critics of GS that the results are very unreliable (Jacso, 2012), the information in GS, once retrieved on the basis of existing publication data and cleaning of citation sources, indicate acceptable levels of reliability in terms of source. Also, the volume of information that can be retrieved for Anthropology increases considerably to levels comparable

with the results for programs in Education and Pedagogical Sciences. There are however several issues regarding how GS results are to be used in the context of assessments.

1. Workload and data limitations
In contrast to WoS, GS data are to be retrieved with quite considerable effort, in particular if the analysis is to be based on second order data. These data are essential in establishing the traceability of citations and the source of the citation. Recent limitations of search results to 20 per query, imposed by the GS engine contributes to this situation. The workload thus imposes limitations to how many publications can be investigated, and influences the design of robust bibliometric analysis, since labor intensive studies are costly. Also, although GS indexes are based on the available meta data of publishers and repositories, including page numbers, issues, author lists and journal title, few of this information is provided to the end user of GS, thus making the correct identification of the publications harder. WoS is in this respect a more precise source, be it that its precision is not very relevant for fields such as Anthropology.

2. Possibilities for field normalization
One important limitation is that GS provides as yet very limited opportunities for field-normalized indicators (Wouters & Costas, 2012). For this study, attempts have been made for GS based field normalization in part based on PoP (Publish or Perish) data for journals (Harzing, 2008) (data not shown). Although technically feasible, these attempts are as yet rather unsatisfactory since the data were based on averages of citations per paper per year, whereas the selection base comprised highly cited papers. Also, using tools for journal data such as PoP does not allow for a traceability check as performed in this study. Even though the percentage of non-traceable citations was small, the comparison might still be biased. In the case for Ed/Ped Sciences attempts have been made to include information of other sources such as Scimago Journal Rank (SJR) (SCImago, 2007). This is possible, but it leads to complicated procedures and methodological issues.

3. The definition of citations
Using GS implies a shift in the definition of what may count as a citation. Whereas the citations in WoS are based on references in academic journals (and increasingly also in other academic sources), the criterion is the academic nature of these forms of publications as established in this database. In GS however, citations also may include references from scientific reports, PhD theses and also student theses. Once using GS, the results inevitably include these citations too, of which one may argue that these suffice or not as tokens of academic recognition. Whether this shift is accepted in view of changing views about assessment standards such as in the new SEP remains to be seen.

4. The selection base
The selection base for publications to be analyzed is obviously relevant to the results and to the methods to be used. In the Ed/Ped case, the selection was performed by program leaders. This led to a selection of highly cited papers but included also publications that were most likely deemed very relevant to the program, but possibly not highly cited. As an indication, 4.5% of the selected publications were not found to be cited at all in GS. This situation of inefficiencies in selecting research outputs has been also observed elsewhere (Abramo et al, 2014). More importantly, although technical issues - such as workload - impose limitations in selecting higher volumes of

publications, the selection base is crucially related to the questions to be addressed in the assessment.

5. Questions to address in the assessment
The shifting ideas about assessment goals for research programs may lead to more variegated bibliometric questions, which in turn may require different research designs such as a focus on specific publications typical for the mission of institutes, or contextual bibliometrics. As the precision and transparency of data shows limitations and the workload is also high, the use of GS will impose restrictions to the possibilities to answer the desired assessment questions.

Conclusions
One of the crucial factors for not applying bibliometry in the social sciences is the coverage of output in Web of Science, which is low for fields such as Anthropology and mediate for Education or Pedagogical Sciences. There is reasonable evidence to consider GS as a valuable source for the analysis of certain fields of science, particularly in the Social Sciences and perhaps also in the Humanities, in providing more information based on a broader set of publication types. However, attention should be given to data reliability. To use GS in the context of evaluation, various ways for benchmarking or field normalization have to be worked out, for instance on the basis of available journal data, to address the issues of research assessments. These are not only technical problems, but they are also issues dependent on the questions raised in assessments. Moreover, the application of GS may find important limitations in fields that rely on even higher volumes of non-journals sources than in the case of the current programs in Anthropology.

References

Abramo, G., D'Angelo, C.A. & Di Costa, F. (2014). Inefficiency in selecting products for submission to national research exercises. *Scientometrics*. 98, 2069–2086

Harzing, A. & Van der Wal, R. (2008) Google Scholar as a new source for citation analysis, *Ethics in Science and Environmental Politics*, 8, 61-73.

Hicks, D. (2005) The Four Literatures of Social Science. In Moed, H. et al, *Handbook of Quantitative Science and Technology Research*. Springer Netherlands, 473-496.

Jacsó, P. (2005). Google Scholar: the pros and the cons. *Online information review*, *29* (2), 208-214.

KNAW (2005) Judging research on its' merits, Amsterdam

KNAW (2013) Towards a framework for the quality assessment of social science research, Amsterdam.

KNAW (2012) Quality indicators in the humanities, Amsterdam

Kousha, K. & Thelwall, M. (2008) Sources of Google Scholar citations outside the Citation Index: A comparison between four science disciplines, *Scientometrics*, 74 (2), 273-294.

Kousha, K., Thelwall, M., & Rezaie, S. (2011). Assessing the citation impact of books: The role of Google Books, Google Scholar, and Scopus. *Journal of the American Society for Information Science and Technology*, *62* (11), 2147-2164.

Nederhof, A.J. (2006) Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics* 66 (1), 81-100.

Nederhof, A. J., van Leeuwen, T. N., & van Raan, A. F. (2010). Highly cited non-journal publications in political science, economics and psychology: a first exploration. *Scientometrics*, *83*(2), 363-374.

SCImago. (2007). SJR — SCImago Journal & Country Rank. Retrieved between September 10 and October 10, 2012, from *http://www.scimagojr.com*

Van Leeuwen, T.N. (2013) Bibliometric research evaluations, Web of Science and the Social Sciences and Humanities: a problematic relationship?, Bibliometrie – Praxis & Forschung, 2013, http://www.bibliometriepf.de/article/view/173

Van Leeuwen, T.N., Van Wijk, E. & Wouters, P.F. (Forthcoming) Bibliometric analysis of output and impact based on CRIS data: A case study on the registered output of a Dutch university.

Wouters, P., & Costas, R. (2012) Users , narcissism and control – tracking the impact of scholarly publications in the 21 st century. In M. Van Berchum, & K. Russell (Eds.) Retrieved from http://sticonference.org/Proceedings/vol2/Wouters_Users_847.pdf

---

[i] The first part of the actual bibliometry has been performed by Ad Prins. The second part has been performed by all present authors.