Journal of Informetrics 3 (2009) 27–35

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Convergent validity of bibliometric Google Scholar data in the field of chemistry—Citation counts for papers that were accepted by *Angewandte Chemie International Edition* or rejected but published elsewhere, using Google Scholar, Science Citation Index, Scopus, and Chemical Abstracts

Lutz Bornmann[a,*], Werner Marx[b], Hermann Schier[b], Erhard Rahm[c], Andreas Thor[c], Hans-Dieter Daniel[a,d]

[a] *ETH Zurich, Professorship for Social Psychology and Research on Higher Education, Zähringerstr. 24, CH-8092 Zurich, Switzerland*
[b] *Max Planck Institute for Solid State Research, Heisenbergstraße 1, D-70569 Stuttgart, Germany*
[c] *University of Leipzig, Department of Computer Science, PF 100920, D-04009 Leipzig, Germany*
[d] *University of Zurich, Evaluation Office, Mühlegasse 21, CH-8001 Zurich, Switzerland*

ARTICLE INFO

Article history:
Received 18 August 2008
Received in revised form 3 November 2008
Accepted 6 November 2008

Keywords:
Web of Science
Science Citation Index
Scopus
Chemical Abstracts
Google Scholar
Citation analysis

ABSTRACT

Examining a comprehensive set of papers ($n = 1837$) that were accepted for publication by the journal *Angewandte Chemie International Edition* (one of the prime chemistry journals in the world) or rejected by the journal but then published elsewhere, this study tested the extent to which the use of the freely available database Google Scholar (GS) can be expected to yield valid citation counts in the field of chemistry. Analyses of citations for the set of papers returned by three fee-based databases – Science Citation Index, Scopus, and Chemical Abstracts – were compared to the analysis of citations found using GS data. Whereas the analyses using citations returned by the three fee-based databases show very similar results, the results of the analysis using GS citation data differed greatly from the findings using citations from the fee-based databases. Our study therefore supports, on the one hand, the convergent validity of citation analyses based on data from the fee-based databases and, on the other hand, the *lack of* convergent validity of the citation analysis based on the GS data.

1. Introduction

For many years the main resources for citation analysis were citation databases, in particular the Science Citation Index (SCI), that are accessible through Web of Science (WoS) available through the research platform ISI Web of Knowledge from the scientific division of Thomson Reuters (Harzing & van der Wal, 2008; Thomson Reuters, 2008). In addition to its multi-disciplinary nature, citation indexing was the major reason why WoS had an unique position among bibliographic databases (Neuhaus & Daniel, 2008). In recent years, however, citation-enhanced databases from several database producers have entered the market. In 2004, two primary competitors to the Thomson Reuters citation indexes became available: Scopus (www.scopus.com) from Elsevier (headquartered in Amsterdam) and Google Scholar (GS) (http://scholar.google.com/) from

* Corresponding author.
E-mail address: bornmann@gess.ethz.ch (L. Bornmann).

1751-1577/$ – see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.joi.2008.11.001

Google Inc. (headquartered in Mountain View, California). Alongside the multidisciplinary-oriented citation indexes from Thomson Reuters, Elsevier, and Google, some discipline-oriented bibliographic databases have also introduced citation indexing (Neuhaus & Daniel, 2008). With the database Chemical Abstracts (CA), for example, Chemical Abstracts Service (CAS), a division of the American Chemical Society, provides one of the most largest databases of published research in the field of chemistry (Marx & Schier, 2005).

GS is particularly interesting for conducting citation analyses, because in contrast to the other databases it can be accessed for free (Neuhaus & Daniel, 2008). According to Harzing and van der Wal (2008) "an important practical advantage of GS is that it is freely available to anyone with an Internet connection and is generally praised for its speed . . . The WoS is only available to those academics whose institutions are able and willing to bear the (quite substantial) subscription costs of the WoS and other databases in Thomson ISI's Web of Knowledge" (p. 62). In addition, GS does not search only peer-reviewed research journals (as WoS does, for example): "It searches lots of non-traditional sources, including preprint archives, conference proceedings and institutional repositories, often locating free versions of articles on author websites" (Giles, 2005, p. 554).

Some recent studies have compared the results of citation analyses using various databases, including especially GS, for sets of publications (for an overview, see Bar-Ilan, 2008; Harzing & van der Wal, 2008; Kousha & Thelwall, 2007; Meho & Rogers, 2008; Vucovich, Baker, & Smith, 2008). According to Kousha and Thelwall (2008) "the few articles that have compared Google Scholar with [Thomson] ISI citation data are exploratory and small-scale in nature" (p. 278). These studies have yielded very different results: while Pauly and Stergiou (2005), for instance, find nearly equal patterns of citations (using the WoS citation databases and GS) for a set of papers across a wide range of disciplines, other studies concluded that different databases returned unique material (see Vucovich et al., 2008). Giles (2005) found almost 14,000 citations in WoS for a paper in *Science* by Saiki et al. (1988) on the polymerase chain reaction, identifying it as the most highly cited paper ever to appear in that journal, whereas GS returned just under 3000 citations.

Considering the few studies that compared the databases with regard to results of citation analyses, Kousha and Thelwall (2007) conclude that more comparisons are needed. Harzing and van der Wal (2008) find that more detailed comparisons are necessary, especially for the field of chemistry. In the context of a comprehensive research project on the peer review process in science, Bornmann and Daniel (2008a, 2008b) investigated the quality of selection decisions on papers submitted for publication to the journal *Angewandte Chemie International Edition* (AC-IE). AC-IE is one of the prime chemistry journals in the world, and it has a higher annual Journal Impact Factor (JIF, provided by Thomson Reuters) than the JIFs of comparable journals (at 10.031 in the 2007 Journal Citation Reports, Science Edition). AC-IE is a journal of the German Chemical Society (Gesellschaft Deutscher Chemiker (GDCh)) and is published by Wiley–VCH (based in Weinheim, Germany). It introduced peer review in 1982, primarily in conjunction with one of the document types published in the journal, "Communications," which are short reports on work in progress or recently concluded experimental or theoretical investigations. What the editors of AC-IE look for most of all is excellence in chemical research. Manuscripts that referees deem to be of high quality are selected for publication. Manuscripts that do not meet the high standards are rejected.

As there is broad support for citation counts of scientific publications as a measure of the impact of scientific research (Cole, 2000; Daniel, 2005; van Raan, 2004), Bornmann and Daniel (2008a, 2008b) tested the assumption that rejected manuscripts earn lower citation counts than accepted papers (see for the validity of funding decisions in grant peer review Marsh, Jayasinghe, & Bond, 2008). As AC-IE also archives manuscripts that have been rejected for publication, the study explored this hypothesis via the citation counts of the accepted papers and also of papers that were rejected by AC-IE but published elsewhere. The citation analyses based on the total number of citations from time of publication up to the end of 2006 for all 1837 manuscripts that were reviewed in the year 2000 and published in a journal (in AC-IE or another chemistry journal) between 2000 and 2006. The citation counts for each publication were extracted from SCI, Scopus, and CA. As the results of Bornmann and Daniel (2008a, 2008b) show, papers accepted by AC-IE were on average statistically significantly more frequently cited than manuscripts rejected by the journal and published elsewhere—independently of from which of the three databases the citation counts were gathered.

The present study looks into the extent to which GS citation data can be used for evaluative purposes in the field of chemistry in a similar way to citation data from fee-based databases (Narin, 1976). The study compares the results of citation analyses using data provided by SCI, Scopus, and CA. If the citation analysis using GS citation data for the total of 1837 papers in the very large publication set were to yield results similar to the results of the citation analyses using the three other databases, that would be an indication of the convergent validity of the GS data for the field of chemistry. In that case, GS, to which access is free, could be used instead of the fee-based databases.

2. Methods: citation searching in the four different databases

2.1. Citation searching in SCI (from Thomson Reuters' WoS) and in CA (from CAS)

With the introduction of the SCI by the Institute for Scientific Information (now Thomson Reuters) in the early 1960s, systematic analyses of the impact and influence of scholarly work became available (Neuhaus & Daniel, 2008). "*Web of Science* now includes not only the *Science Citation Index*, but also the *Social Sciences Citation Index*, the *Arts and Humanities Citation Index*, *Index Chemicus*, and *Current Chemical Reactions*, resulting in a truly multidisciplinary citation resource. *Web of Science*, which now covers nearly 9300 high-quality, core journals from every field, is used by over 3400 organizations and universities in more than 90 countries around the world" (Thomson Reuters, 2008). Thomson Reuters extracts the references

from all of the indexed journals, and the cited reference interface lists all citations to publications (Bar-Ilan, 2008). Even though for the citation indexes in WoS Thomson Reuters continuously covers more than 9000 journals, this is still only a selected set of the journals currently available (see here Neuhaus & Daniel, 2008). Today, Ulrich's Periodicals Directory lists more than 22,000 peer-reviewed academic and scholarly journals and 2300 electronic journals. According to Marx and Schier (2005) the journals indexed in the WoS databases cover mainly the core disciplines in the natural sciences, and they do not sufficiently cover fields such as computer science or engineering science.

The CA scientific literature and patents database from CAS is a prime example of a database offering also intellectually analyzed content obtained from journals and patents, and because of that, it is relatively expensive. The CA is one of the largest literature databases, containing chemistry-related primary literature from more than 10,000 major scientific journals worldwide as well as all chemistry-related patents, books, and research reports (Marx & Schier, 2005). CA represents the world's most important compendia of chemistry and related sciences such as biology and life sciences, engineering sciences, materials sciences, medical sciences, and the substance-related fields in physics (Neuhaus & Daniel, 2008). The citation searches in SCI and in CA for the publication set of this study was conducted via the online database service STN International (http://www.stn-international.de/) operated by FIZ Karlsruhe in Germany. Using the STN retrieval system and the STN statistics functions, SCI (SCIsearch) and CA (HCAplus) could be utilized for comprehensive citation analyses. For our study, two STN-specific upload files were generated, two separate files for SCI and CA. This was necessary, because the databases transcribe some authors' names differently (especially hyphenated surnames and names containing an *Umlaut*). For the automated search process for publications in the databases, first author name, volume number, first page number, and year of publication were used. When a publication was found, the citations for the period from date of publication to the end of 2006 were determined and tabulated for the further analyses.

2.2. Citation searching in Scopus

"In 2004, Elsevier released its ambitious Scopus abstract and indexing database covering over 15,000 peer-reviewed journal titles, including coverage of approximately 500 Open Access journals, 700 conference proceedings, and 125 book series. Altogether, Scopus indexes more journals than Thomson Scientific's citation indexes, and offers greater coverage of Open Access journals, but lacks the depth of coverage in years of journals" (Neuhaus & Daniel, 2008, p. 203). According to de Moya-Anegón et al. (2007), Scopus is the largest of the currently available databases for scientific searches. It provides full citation coverage from 1996 onwards (Bar-Ilan, 2008). According to Visser and Moed (2008) "Scopus is a genuine alternative to the Web of Science as a data source for bibliometric studies of research performance in science fields during the time period as from 1996" (p. 25).

In contrast to the citation searching in SCI and in CA, the citations for the individual publications in the publication set of this study were searched manually in Scopus. To our knowledge, there is no host like STN International offering automatic searching in Scopus for citations for a publication set. Therefore, two persons of our research team used bibliographic information to search for each individual publication in Scopus and tabulated the citations for statistical analysis.

2.3. Citation searching in Google Scholar

According to Kousha and Thelwall (2007) "the citation facility of Google Scholar is a potential new tool for bibliometrics" (p. 1057). The great advantage of GS over the other databases is the many documents covered that are not usually to be found in other, journal-dominated databases (Bar-Ilan, 2008). It is reasonable to assume that a considerable number of citations from Open Access and non-Open Access documents in different subject areas – especially those from non-journal sources – that are not found in WoS and the other fee-based databases can be retrieved through GS. While a large number of publishers have released their information (mainly full texts of publications, mostly PDFs) to GS's software robots (such as, for example, Nature Publishing Group or Springer) (Giles, 2005; Kousha & Thelwall, 2007), some large publishers have denied GS access to their archives (such as the American Chemical Society). Still, Jacso (2005) concludes, "the coverage of Google is impressively broad and includes the most important scholarly publishers' archives" (p. 209).

But a number of publications name also some weaknesses of GS that are said to make conducting citation analyses difficult. As Neuhaus and Daniel (2008) point out, "Google does not disclose any information about the sources processed, nor the document types included, nor the time span covered" (p. 200). GS also does not provide information on how frequently it is updated (Harzing & van der Wal, 2008). According to Sanderson (2008), "GS was relatively unstructured and often contained a number of duplicate entries to the same publication and the facilities to filter results were limited" (p. 1185). Beyond that, the citation counts that can be extracted from GS are inflated, containing citations from poor quality publications. In a citation analysis of a publication set Meho and Yang (2007b) found that approximately one-third of the citations were from sources that most would agree should not be included in citation studies, including "Bachelor's theses, presentations, grant and research proposals, doctoral qualifying examinations" (p. 2112) and "Master's theses, technical reports, research reports" (p. 2115).

Citation searching in GS for the publication set of this study was conducted at the start of 2008 using Online Citation Service, a Web application developed at the University of Leipzig (Thor, Aumueller, & Rahm, 2007). Using relevant keywords from the titles of the papers accepted for publication by AC-IE or rejected (and published elsewhere) as well as the author and journal names, GS was searched using automatic queries. For each paper, several search strategies were used (for example, searches used the complete title of a paper or only certain words in the title in combination with other bibliographic

information), in order to guarantee that all relevant publications were found. The search results were then compared with the list of papers in the publication set of the study so that, among other things, false hits could be eliminated.

An important aspect of post-processing the data is identifying duplicate publication entries in GS. (For example, the interested reader can search GS using the search term *intitle:survey approaches author:rahm*. Note that the search results all refer to the same real-world publication.) The reason for the many duplicates lies in the automatic generation of the GS data set (among other things, automatic extraction of references lists in PDF files), which leads to heterogeneous bibliographic information for the same publications (due to typographical errors in titles, missing authors, authors listed in incorrect order, differences in the names used for the journals or conferences, and so on). The identification and summary of all duplicates is obviously crucial with regard to the quality of a citation analysis, for otherwise, relevant citations are not considered. Automatic identification of duplicate records is a very challenging task that is receiving a lot of attention in computer science research (Thor & Rahm, 2007). At the present time, no 100% correct automatic identification of duplicates is possible, which is why after automatic searching the results have to be post-processed manually and modified if necessary. For identified duplicates, the numbers of citations provided by GS are summed.

3. Results

For the papers accepted for publication by AC-IE or rejected but published elsewhere, the following sections present the results of citation analyses performed using data from SCI, Scopus, and CA and compared to the results of the citation analysis using data from GS. The results of the individual comparisons between the databases are presented in Section 3.1. Section 3.2 looks at the reasons for the differences between the analyses using citations drawn from the different databases.

3.1. Results of the analyses of citations that were tracked in the four databases

The results of the analyses of citations that were tracked in the GS, SCI, Scopus, and CA databases are shown in Table 1. As the table shows, citations were tracked in each of the databases for a total of 1837 papers that were accepted for publication by AC-IE or rejected (and published elsewhere). All 1837 papers were found with 1 entry in SCI and CA, but 10 papers could not be found in Scopus. In GS, a total of 90 papers could not be found (4.9%); of the 1747 papers that were found in GS, 73.9% returned 1 hit and 21.2% returned from 2 to 4 hits. In many cases where several hits were returned, not all of them were about one and the same publication; some of them were references to the publication or a pre-publication working paper (see here also Aaltojarvi, Arminen, Auranen, & Pasanen, 2008). If the search for one paper returned several publication hits, the citations for the individual hits were summed.

Table 1
Differences between citation counts from each of the four databases, for papers ($n = 1837$) that were accepted or rejected (but published elsewhere) by AC-IE.

	Google Scholar (GS)	Science Citation Index (SCI)	Scopus	Chemical Abstracts (CA)
Of the total of 1837 papers, absolute and relative number of papers that were found in a database				
Absolute	1747	1837	1827	1837
Relative	95.1%	100.0%	99.5%	100.0%
Absolute and relative number of papers that were found in a database and had 0 citations				
Absolute	843	31	38	21
Relative	48.3%	1.7%	2.1%	1.1%
Total citation counts	9320	44,502	44,601	48,160
Mean citation counts and (median)				
Papers accepted for publication in AC-IE	6.48 (1)	30.55 (23)	30.81 (23)	33.10 (25)
Papers rejected and published elsewhere	4.35 (1)	18.43 (12)	18.50 (13)	19.92 (14)
Total	5.34 (1)	24.23 (17)	24.41 (17)	26.22 (19)
Difference between number of citations (absolute value) returned by a database and number of citations returned by *Google Scholar*				
Mean		19.02	19.13	21.00
Median		13	14	15
Number of papers with a difference of fewer than two citations (in percent)		116 (6.6%)	115 (6.6%)	91 (5.2%)
Difference between number of citations (absolute value) returned by a database and number of citations returned by *Science Citation Index*				
Mean			1.71	2.69
Median			1	2
Number of papers with a difference of fewer than two citations (in percent)			1168 (63.9%)	784 (42.7%)
Difference between number of citations (absolute value) returned by a database and number of citations returned by *Scopus*				
Mean				2.90
Median				2
Number of papers with a difference of fewer than two citations (in percent)				803 (44.0%)

Notes: Publication window: between 2000 and 2006. Citation window: from time of publication up to the end of 2006.

A comparison of the number of citations that were found in the four databases for the publication set reveals very clear differences between GS and the three other databases: whereas in SCI and Scopus about 2% and in CA about 1% of papers for which citation counts were extracted showed 0 citations, in GS no citations were found for nearly one-half of the papers. Accordingly, the total citation counts in Table 1 for SCI, Scopus, and CA range from 44,502 to 48,160 citations, but the total citation count for GS is only 9320.

In Table 1 the results of the analysis of GS citations are shown for the citation window from publication year to 2006 (as for the analyses of citations that were based on the other three databases). As the searches in GS were conducted in the year 2008, only those citing papers entered into the citation counts for the accepted and rejected (but published elsewhere) papers for which GS showed a publication year between 2000 and 2006. As in contrast to the other three databases GS does not show the publication year for all citing papers (Kousha & Thelwall, 2008, for example, find mainly for the field of chemistry a high percentage of Chinese language citations in GS that have no publication year in Arabic numerals), the number of *all* citing papers and the number of *all* citing papers *with usable publication years* differ. In order to be able to determine the difference between the two citation rates for the publication set of this study, in GS we identified for the 1747 papers in addition all citations (with and without a publication year) since publication up to the time point of the search at the start of 2008. As the search results show, while the number of papers with 0 citations decreases from 843 (see Table 1) to 727 and the total citation counts increase from 9320 (see Table 1) to 13,345 citations, the numbers still differ greatly, however, from the numbers for the other three databases.

In addition to the total citation counts and the number of papers with 0 citations, Table 1 also shows the mean citation counts (arithmetic mean and median) for papers accepted for publication by AC-IE and papers rejected but published elsewhere. Even though based on the data in GS papers accepted for publication by AC-IE were on average more frequently cited than manuscripts rejected by the journal and published elsewhere, the difference between the two groups is lower due to the lower GS citation counts altogether: while SCI, Scopus, and CA returned an average number of citations for accepted papers of about 30 citations and for rejected papers but published elsewhere about 20 (see here also the results in Bornmann & Daniel, 2008a, 2008b), the averages returned by GS are 6.48 (accepted papers) and 4.35 (rejected papers) citations (see Table 1) (seen as percentages, however, the difference between the average GS citation counts for accepted and rejected papers is about the same as the difference between the average citation counts for accepted and rejected papers when tracked in the fee-based databases).

The bottom three rows in Table 1 show the average differences (arithmetic mean and median of the absolute difference values) in citation rates returned by the four different databases for the papers in the data set of this study. This analysis, too, reveals clear differences between GS and the other three databases: whereas the difference in the individual citation rates between the citations returned by SCI, Scopus, and CA was on average about two to three citations, the citation rates using the GS database differed on average by about 20 citations from the citation counts returned by the other databases. Accordingly, for SCI, Scopus, and CA the percentages of papers with a difference of fewer than two citations between the citation rates from two databases are 42.7% (SCI and CA) and 63.9% (SCI and Scopus). But if the citation counts from these three databases are compared with citation counts returned by GS, it is found that only between 5.2% and 6.6% of the papers show a small or no difference in the citations.

Seen overall, the results in Table 1 indicate that there is little difference between the results of the citation analyses using SCI, Scopus, and CA data but clear differences between the results of the analysis using GS data and the analysis results using the three other databases. Bauer and Bakkalbasi (2005) found similar results (but, however, with altogether higher citation counts using GS than using fee-based databases) for articles published in the *Journal of the American Society for Information Science and Technology* (JASIST): "For JASIST articles published in 2000, Google Scholar provided statistically significant higher citation counts than either Web of Science or Scopus, while there was no significant difference between Web of Science and Scopus." Clear differences in citation counts returned by GS as compared to citation counts returned by WoS and/or Scopus were also reported by Bakkalbasi, Bauer, Glover, and Wang (2006), Bar-Ilan (2008), Meho (2007), Meho and Yang (2007b), Noruzi (2005), and Vaughan and Shaw (2006). Moreover, a comparison of *h* index values of a list of highly cited Israeli researchers (on the *h* index, see Bornmann & Daniel, 2007; Hirsch, 2005) based on citation counts retrieved from the WoS, Scopus, and GS revealed that "the results obtained through Google Scholar are considerably different from the results based on the Web of Science and Scopus" (Bar-Ilan, 2008, p. 257).

3.2. Reasons for the differences in the citation data from the four databases

Section 3.1 above presented the differences in the results of the analysis of citations returned by GS and the fee-based databases. The present section looks at the reasons for these differences. Basically we have to differentiate between reasons for failing to identify source items and the reasons of poor covering of the citing articles. In general Google has access to the bibliographic information of papers published by all publishers, but not for the citations embedded. Citations are part of the full text which is not available for free but protected by intellectual properties rights. For indexing of citations Google has to rely on cooperation with the publishers. This may be an active cooperation where the publisher is pushing the citations continuously to Google or the classic passive strategy by allowing access to full text for the Google web crawlers. However, looking for the concrete reasons of differences in the results of the analysis of citations returned by GS and the fee-based databases is a very difficult undertaking overall, as Google—in contrast to the fee-based database providers, provides only minimal information about the content of GS, informa-

Table 2
Publisher of the journals in which papers accepted or rejected (but published elsewhere) by AC-IE were published, by papers that were not found in Google Scholar (GS) and that showed 0 or at least one citation in GS and in Science Citation Index (SCI).

Publisher of the journal in which the paper was published	Paper was not found in GS	Paper with 0 citations in SCI and in GS	Paper with 0 citations in GS (and at least 1 citation in SCI)	Paper with at least 1 citation in GS (and in SCI)	Total
Wiley–VCH	73	5	451	531	1060
American Chemical Society	10	3	160	200	373
Royal Society of Chemistry	3	5	83	64	155
Elsevier	1	7	75	72	155
Chemical Society of Japan	0	0	15	11	26
Thieme	1	0	16	3	20
Other publisher	2	7	15	23	47
Total	90	27	815	904	1836

Notes: In the data set there is only one paper for which 0 citations were found in SCI but more than 0 citations (namely, 4) were found in GS.

tion such as publisher and journal lists, time span, or the disciplinary distribution of records (Kousha & Thelwall, 2007).

3.2.1. Reasons that have to do with the cited papers

As was shown in Section 3.1, for papers accepted and rejected (but published elsewhere) by AC-IE it is conspicuous that GS returns 0 citations for almost 50% of them (the fee-based databases returned 0 citations for only 1–2% of the papers). This result could be an indication that for certain papers in the field of chemistry (papers, for example, that were published within a certain publication window) GS can return no citation counts. The extent to which this might be true was tested based on various characteristics of the papers in the data set of this study: publisher of the journal, the journal in which a paper was published, and the publication year of a paper. Table 2 shows the results for publisher of the journal in which papers were published that were accepted or rejected (but published elsewhere) by AC-IE. The citation counts returned by SCI and GS are compared. Papers that were published by one and the same publisher (some publishers are grouped together under the category "other publisher") are shown according to whether they (1) were not found in GS, or (2) had 0 citations in SCI and in GS, (3) had 0 citations only in GS (and had at least one citation in SCI), or (4) had at least one citation in GS (and in SCI).

As the results of Table 2 show, 73 (81.1%) of a total of 90 papers that were not found in GS were published in a journal published by Wiley–VCH (mainly in AC-IE). However, as approximately 60% of the papers in the data set of this study were published in a journal published by Wiley–VCH (mainly in AC-IE), this result is not surprising. For 27 papers in the data set, both GS and SCI returned 0 citations. With regard to these papers, the table shows no conspicuous frequencies for particular publishers. The same is true of papers having 0 citations only in GS (see column 4 in Table 2) and having at least one citation in GS (and in SCI) (see column 5 in Table 2). For example, 451 papers that were published in a journal published by Wiley–VCH returned 0 citations in GS, while 531 papers (published in a journal by the same publisher) returned at least one citation in GS. The analysis of the cited papers according to the publisher of the journal in which the papers were published thus provides little information on why the citations counts returned by GS differ greatly from the citation counts returned by SCI.

The analyses for which the results are shown in Table 2 were not conducted only for publisher but also for journal and for publication year of the papers in the data set of this study. In addition to the SCI citation counts, also the Scopus and CA citation counts were analyzed in combination with the GS citation counts. However, as the analyses with these citation counts – similar to those in Table 2 – did not reveal any conspicuous rates for particular journals and publication years, the results are not shown in a table.

3.2.2. Reasons that have to do with the citing papers

The reasons for clearly lower citation counts returned by GS as compared to the fee-based databases apparently relate to the papers that cited the chemistry papers. In a comprehensive study, Kousha and Thelwall (2008) examined a sample of 882 papers from 39 Open Access journals (indexed by Thomson Reuters) published in 2001 in the fields of biology, chemistry, physics, and computing and classified the type, language, publication year, and accessibility of the GS and WoS unique citing sources. For the articles in the field of chemistry ($n = 276$) they found a relatively low number of GS citations ($n = 279$) and a much higher number of WoS citations ($n = 668$) (see here also the results in Evidence Ltd., 2007, p. 21). This result agrees with the results found for the analysis of the WoS and GS citations for the publication set of this study.

As the classification of the WoS *unique* citations (that is, those citations that are returned by WoS *only* and not by GS) in the study by Kousha and Thelwall (2008) showed, 41% of these citations are found in a journal published by Elsevier. "One explanation for such missing citations in Google Scholar is that it couldn't directly access the Elsevier publication database . . . in order to index the 'citing references' to Open Access chemical journal articles" (Kousha & Thelwall, 2008, p. 288). For the articles in the field of chemistry, 28% of *unique* citations returned by WoS came from articles in journals that are published by the American Chemical Society (ACS): "Although Google Scholar directly indexes bibliographic information and abstracts of journal articles from ACS publications, it couldn't access the references of those articles which targeted OA articles in this study" (Kousha & Thelwall, 2008, p. 288).

Table 3
Publishers of the citing papers of 373 manuscripts rejected by AC-IE but published in ACS journals using Science Citation Index and Google Scholar.

Database	American Chemical Society	Elsevier	Wiley–VCH	Other publisher	Total
Science Citation Index	2961 (34%)	1923 (22%)	1327 (15%)	2512 (29%)	8723
Google Scholar	29 (1%)	589 (21%)	461 (16%)	1725 (62%)	2804

Table 4
IP-domains of all source papers/citing papers found in Google Scholar (GS) according to three major Publishers.

Source paper	*n*	Citing paper				
		wiley.com	elsevier.com	acs.org	Other IP-domain	Total
wiley.com	544	755	815	34	2660	4264
elsevier.com	167	55	119	2	369	545
acs.org	267	288	375	21	1131	1815
Other IP-domain	739	954	1096	41	848	2941
Total	1717	2052	2405	98	5010	9565

Notes: The IP-domains were extracted from the URL provided by GS for every single paper. An URL was not available for all source papers found in GS: for 30 of the total of 1747 papers that were accepted or rejected (but published elsewhere) by AC-IE (see Table 1) an IP-domain was missing. With $n = 9565$ the number of total citation counts in the table differs from the number of total citation counts in Table 1 ($n = 9320$), because we considered in Table 4 only the cases where both the cited and the citing paper have an IP-domain. Additionally, we did not restrict the analysis here on citing paper with usable publication years (as we did it in the analyses for the findings in Table 1).

With our data set we can show that this is true not only for OA articles, but also holds for papers that are published by traditional print journals. We verified this by analyzing the journal-to-journal citations, e.g. from *Organometallics* (published by ACS) to *Organometallics*, or in general by analyzing the publisher-to-publisher citations as given in Table 3: the citing papers for the manuscripts rejected by AC-IE but published by ACS journals ($n = 373$) are primarily published by journals of the three major publishers ACS, Elsevier and Wiley–VCH, where we found 6211 (71%) citing papers in SCI of the total citation counts (SCI, $n = 8723$), but only 1079 (38%) citing papers in GS of the total citation counts (GS, $n = 2804$). The results of the IP-domain analysis of all combinations of source papers/citing papers of the three major publishers found in GS are given in Table 4. It is clearly visible that combinations including acs.org are significantly fewer existent than combinations including wiley.com or elsevier.com. Further analyzing the few GS citations coming from the ACS domain (acs.org, $n = 98$, see Table 4) reveals that almost all links are pointing to free accessible teaser papers.

4. Discussion

Google Scholar has been available on the Internet for free since 2004 as beta release, and it "provides every academic with access to citation data regardless of their institution's financial means" (Harzing & van der Wal, 2008, p. 72). Because of the fee-based databases' poor coverage of certain fields and the difficulty of citation analysis for publications that are not published in journals indexed by Thomson Reuters (or in Scopus), the analysis of GS data can be a great advantage. A study by Rahm and Thor (2005), for instance, which analyzed the citation counts for two main database conferences (SIGMOD and VLDB) and three database journals (TODS, VLDB Journal and Sigmod Record) over 10 years, demonstrated the high usefulness of GS for this specific subject area. However, "data preparation, data cleaning and integrating data from several sources are important to achieve useful and correct results" (Rahm & Thor, 2005, p. 53).

For a comprehensive set of papers that were accepted for publication by AC-IE or rejected (but published elsewhere), the study reported here investigated the extent to which GS can be used for evaluative purposes in the field of chemistry and yield valid results. The results of citation analyses using three fee-based databases were compared with the results of a citation analysis based on GS data. Whereas the analyses using citations returned by the three fee-based databases show very similar results, the results of the analysis using GS citation data differed greatly from the results using citations from the fee-based databases. The results therefore support, on the one hand, the convergent validity of citation analyses based on data from the fee-based databases and, on the other hand, the *lack of* convergent validity of the citation analysis based on the GS data.

All in all, the results support Moed's (2005) claim that the SCI, in contrast to GS, is suitable for research evaluation in chemistry (see also Kousha & Thelwall, 2008). This also holds – as the results of the present study show – for Scopus and CA. The peer-reviewed journal literature, which plays a very important role in the exchange of research findings in the field of chemistry, is covered comprehensively by the fee-based databases. The results of the citation analysis using GS data are obviously decisively limited due to major publishers' lack of cooperation with Google. ACS, as one of the dominating publishers in chemistry, does not cooperate with Google which causes a significant loss of citations and prohibits the use of GS in citation analysis. Unfortunately information about cooperations belongs to the confidential company policies and most publishers are not willing to give clear statements. Using GS for citation analysis might be beneficial for the fields of "(1) business, administration, finance & economics; (2) engineering, computer science & mathematics; (3) social sciences, arts & humanities" (Harzing & van der Wal, 2008, p. 65), where peer-reviewed journal literature does *not* dominate the

formal scientific communication (and where instead, books, contributions to anthologies, conference papers and so on are published frequently), but it is not beneficial for the field of chemistry.

However, independently of the field or discipline, anyone using GS must be aware that the database is still in beta testing (Bar-Ilan, 2008). According to an overview by Bar-Ilan (2008), neither the Boolean operators nor the range operator (for limiting the date of publication) work properly. Furthermore, as also this study showed, it is not always possible "to correctly identify the publication year of the item, and citations are not always attributed to the correct publication" (Bar-Ilan, 2008, p. 260). For Jacso (2008a) GS "does a really horrible job matching cited and citing references" and "often can't tell apart a page number from a publication year, part of the title of a book from a journal name, and dumps at you absurd data" (see also Perkel, 2005). In addition, "the hit counts and the citation counts of Google Scholar keep changing dramatically. If they were increasing, it could be chalked up to adding new records, but often these counts decrease because of deleting records from the database" (Jacso, 2008b, p. 270). Meho and Yang (2007a) conclude that overall, GS is "not conducive for large-scale comparative citation analyses" (p. 579). As it is not always clear what is being counted in GS (Falagas, Pitsouni, Malietzis, & Pappas, 2008), it is not always possible to be sure, in a group comparison for evaluative purposes based on citation counts, what is actually being compared to what (Pringle, 2008). Upon the background of these weaknesses it seems justified that Gardner and Eng (2005) conclude that Google should improve GS significantly in the beta testing phase before it becomes fully operational.

References

Aaltojarvi, I., Arminen, I., Auranen, O., & Pasanen, H. M. (2008). Scientific productivity, web visibility and citation patterns in sixteen Nordic sociology departments. *Acta Sociologica*, *51*(1), 5–22.

Bakkalbasi, N., Bauer, K., Glover, J., & Wang, L. (2006). Three options for citation tracking: Google Scholar, Scopus and Web of Science. *Biomedical Digital Libraries*, *3*(7).

Bar-Ilan, J. (2008). Which *h*-index?—A comparison of WoS, Scopus and Google Scholar. *Scientometrics*, *74*(2), 257–271.

Bauer, K., & Bakkalbasi, N. (2005). An examination of citation counts in a new scholarly communication environment. *D-Lib Magazine*, *11*(9).

Bornmann, L., & Daniel, H.-D. (2007). What do we know about the *h* index? *Journal of the American Society for Information Science and Technology*, *58*(9), 1381–1385.

Bornmann, L., & Daniel, H.-D. (2008a). The effectiveness of the peer review process: Inter-referee agreement and predictive validity of manuscript refereeing at Angewandte Chemie. *Angewandte Chemie International Edition*, *47*(38), 7173–7178.

Bornmann, L., & Daniel, H.-D. (2008b). Selecting manuscripts for a high impact journal through peer review: A citation analysis of Communications that were accepted by Angewandte Chemie International Edition, or rejected but published elsewhere. *Journal of the American Society for Information Science and Technology*, *59*(11), 1841–1852.

Cole, J. R. (2000). A short history of the use of citations as a measure of the impact of scientific and scholarly work. In B. Cronin & H. B. Atkins (Eds.), *The web of knowledge. A Festschrift in honor of Eugene Garfield* (pp. 281–300). Medford, NJ, USA: Information Today.

Daniel, H.-D. (2005). Publications as a measure of scientific advancement and of scientists' productivity. *Learned Publishing*, *18*, 143–148.

de Moya-Anegón, F., Chinchilla-Rodríguez, Z., Vargas-Quesada, B., Corera-Álvarez, E., Muñoz-Fernández, F., González-Molina, A., et al. (2007). Coverage analysis of Scopus: A journal metric approach. *Scientometrics*, *73*(1), 53–78.

Evidence Ltd. (2007). *The use of bibliometrics to measure research quality in UK higher education institutions*. London, UK: Universities UK.

Falagas, M. E., Pitsouni, E. I., Malietzis, G. A., & Pappas, G. (2008). Comparison of PubMed, Scopus, Web of Science, and Google Scholar: Strengths and weaknesses. *FASEB Journal*, *22*(2), 338–342.

Gardner, S., & Eng, S. (2005). Gaga over Google? Scholar in the social sciences. *Library Hi Tech News*, 22(8).

Giles, J. (2005). Science in the web age: Start your engines. *Nature*, *438*(7068), 554–555.

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

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(46), 16569–16572.

Jacso, P. (2005). Google Scholar: The pros and the cons. *Online Information Review*, *29*(2), 208–214.

Jacso, P. (2008). The plausibility of computing the *h*-index of scholarly productivity and impact using reference-enhanced databases. *Online Information Review*, *32*(2), 266–283.

Jacso, P. (2008b). Google Scholar and The Scientist. Retrieved June 5, 2008, from http://www2.hawaii.edu/~jacso/extra/gs/.

Kousha, K., & Thelwall, M. (2007). Google Scholar citations and Google Web/URL citations: A multi-discipline exploratory analysis. *Journal of the American Society for Information Science and Technology*, *58*(7), 1055–1065.

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

Marsh, H. W., Jayasinghe, U. W., & Bond, N. W. (2008). Improving the peer-review process for grant applications—reliability, validity, bias, and generalizability. *American Psychologist*, *63*(3), 160–168.

Marx, W., & Schier, H. (2005). CAS kontra Google. *Nachrichten aus der Chemie*, *53*, 1228–1232.

Meho, L. I. (2007). The rise and rise of citation analysis. *Physics World*, *20*(1), 32–36.

Meho, L. I., & Rogers, Y. (2008). Citation counting, citation ranking, and *h*-index of human-computer interaction researchers: A comparison of Scopus and Web of Science. *Journal of the American Society for Information Science and Technology*, *59*(11), 1711–1726.

Meho, L. I., & Yang, K. (2007a). Fusion approach to citation-based quality assessment. In D. Torres-Salinas & H.F. Moed (Eds.) *Proceedings of the 11th Conference of the International Society for Scientometrics and Informetrics* (Vol. 2, pp. 568–581). Madrid, Spain: Spanish Research Council (CSIC).

Meho, L. I., & Yang, K. (2007b). Impact of data sources on citation counts and rankings of LIS faculty: Web of Science versus Scopus and Google Scholar. *Journal of the American Society for Information Science and Technology*, *58*(13), 2105–2125.

Moed, H. F. (2005). *Citation analysis in research evaluation*. Dordrecht, The Netherlands: Springer.

Narin, F. (1976). *Evaluative bibliometrics: the use of publication and citation analysis in the evaluation of scientific activity*. Cherry Hill, NJ, USA: Computer Horizons.

Neuhaus, C., & Daniel, H.-D. (2008). Data sources for performing citation analysis—an overview. *Journal of Documentation*, *64*(2), 193–210.

Noruzi, A. (2005). Google Scholar: The new generation of citation indexes. *Libri*, *55*(4), 170–180.

Pauly, D., & Stergiou, K. I. (2005). Equivalence of results from two citation analyses: Thomson ISI's Citation Index and Google's Scholar service. *Ethics in Science and Environmental Politics*, *5*, 33–35.

Perkel, J. M. (2005). The future of citation analysis. *The Scientist*, *19*(20), 24.

Pringle, J. (2008). Trends in the use of ISI citation databases for evaluation. *Learned Publishing*, *21*(2), 85–91.

Rahm, E., & Thor, A. (2005). Citation analysis of database publications. *SIGMOD Record*, *34*(4), 48–53.

Saiki, R. K., Gelfand, D. H., Stoffel, S., Scharf, S. J., Higuchi, R., Horn, G. T., et al. (1988). Primer-directed enzymatic amplification of DNA with a thermostable DNA-polymerase. *Science*, *239*(4839), 487–491.

Sanderson, M. (2008). Revisiting *h* measured on UK LIS and IR academics. *Journal of the American Society for Information Science and Technology*, *59*(7), 1184–1190.

Thomson Reuters. (2008). *Using bibliometrics: a guide to evaluating research performance with citation data*. Philadelphia, PA, USA: Thomson Reuters.

Thor, A., Aumueller, D., & Rahm, E. (2007). Data integration support for mashups. In U. Nambiar & Z. Nie (Eds.), *Proceedings of the Sixth International AAAI Workshop on Information Integration on the Web* (pp. 104–109). Vancouver, Canada: AAAI Press.

Thor, A., & Rahm, E. (2007). MOMA—a Mapping-based Object Matching System. In *Proceedings of the Third Biennial Conference on Innovative Data Systems Research* (pp. 247–258). Asilomar, CA, USA: www.cidrdb.org.

van Raan, A. F. J. (2004). Measuring science. Capita selecta of current main issues. In H. F. Moed, W. Glänzel, & U. Schmoch (Eds.), *Handbook of quantitative science and technology research. The use of publication and patent statistics in studies of S&T systems* (pp. 19–50). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Vaughan, L., & Shaw, D. (2006, 7–9 September). *Comparison of citations from ISI, Google, and Google Scholar: seeking Web indicators of impact*. Paper presented at the Ninth International Conference on Science and Technology Indicators, Leuven, Belgium.

Visser, M. S., & Moed, H. F. (2008). Comparing Web of Science and Scopus on a paper-by-paper basis. In J. Gorraiz & E. Schiebel (Eds.), *Excellence and emergence. A new challenge for the combination of quantitative and qualitative approaches. Proceedings of the 10th International Conference on Science and Technology Indicators* (pp. 23–25). Vienna, Austria: Austrian Research Centers (ARC).

Vucovich, L. A., Baker, J. B., & Smith, J. T. (2008). Analyzing the impact of an author's publications. *Journal of the Medical Library Association*, *96*(1), 63–66.