Possible inaccuracies occurring in citation analysis

H.F. Moed and M. Vriens
LISBON Institute, Science Studies Unit, University of Leiden, Stationsplein 242, 2312 AR Leiden, The Netherlands

Received 31 August 1988

Citation analysis of scientific articles constitutes an important tool in quantitative studies of science and technology. Moreover, citation indexes are used frequently in searches for relevant scientific documents. In this article we focus on the issue of reliability of citation analysis. How accurate are citation counts to individual scientific articles? What pitfalls might occur in the process of data collection? To what extent do 'random' or 'systematic' errors affect the results of the citation analysis?

We present a detailed analysis of discrepancies between target articles and cited references with respect to author names, publication year, volume number, and starting page number. Our data consist of some 4500 target articles published in five scientific journals, and 25000 citations to these articles. Both target and citation data were obtained from the Science Citation Index, produced by the Institute for Scientific Information.

It appears that in many cases a specific error in a citation to a particular target article occurs in more than one citing publication. We present evidence that authors in compiling reference lists, may copy references from reference lists in other articles, and that this may be one of the mechanisms underlying this phenomenon of 'multiple' variations/errors.

1. Introduction

Citation analysis constitutes an important tool in quantitative studies of science and technology. Moreover, citation indexes are used frequently in searches for relevant scientific documents on a certain topic. Authors of scientific publications refer to earlier publications that bear some relevance to the claims the authors want to make. In the list of cited references the author should indicate a publication in such a way that it is uniquely defined, thus enabling a reader to look it up easily. In citation analysis, one creates a set of scientific articles, and wishes to determine the number of times each of these articles is cited during a certain period of time. To be more specific, the analyst wants to determine how frequently each article occurs in the lists of cited references of some collection of other publications from the scientific literature. The articles for which the frequency of citation is to be determined will be called target articles throughout this paper, since they constitute the 'targets' in the referencing process. The publications that cite a specific target article—that is, that contain a cited reference to the target article—will be indicated as source publications.

The above type of citation analysis is performed in most cases with the use of computerized databases on scientific literature, created by the Institute for Scientific Information (ISI), in Philadelphia. ISI compiles three large databases that contain information on the cited references in source articles, published in a large number of scientific journals: the Science Citation Index (SCI), covering the Natural Sciences and Life Sciences, the Social Science Citation Index (SSCI) for the Social Sciences, and finally the Arts and Humanities Citation Index (A&HCI). ISI publishes the contents of the databases in a printed form. In addition, these databases are made available online by several host computer organizations. These hosts store the databases on computer memory and offer telecommunication facilities and software to search the databases for relevant scientific material. The online version of the SCI is called SCISEARCH.

ISI extracts the following information from a cited reference:
- the first author;
- the title of the publication medium (journal, book, report);
- the year of publication;
- (if present) the volume number;
- (if present) the starting page number.

Suppose we have created a set of target articles, and for each article we have gathered all relevant bibliographic information, i.e. information on

0165-5515/89/$3.50 © 1989, Elsevier Science Publishers B.V.

Table 1
Example of discrepancies between target article and cited references

Target	SCHOEMAKER H, V225, P61	
Cited	SCHOEMAKER H, V225, P61	57x
Reference	SHOEMAKER H, V225, P61	6x
	SCHOEMAKER H, V255, P61	1x
	SCHOEMAKER H, V275, P61	2x

authors, journal title, publication year, volume and starting page number. Other information may be included as well, such as the title or data on the institutional affiliations of the authors (corporate source data), but these types of information are not relevant in the citation analysis itself. Generally speaking, a citation to a specific target article is identified by matching (either manually or by means of computer programs) the information on cited references extracted by ISI with the appropriate information on the target article.

In the data on target articles or cited references, errors and/or variations may occur with respect to the author names, journal title, publication year, volume and starting page number. These types of information will be called datafields. As a consequence, discrepancies arise between the bibliographic information of the target and the cited reference, although it may be clear—or at least highly probable—that the cited reference is actually a citation to a particular target.

In Table 1 we present some typical examples of discrepancies that may arise with respect to author name and volume number. The target article concerned is cited 57 times in references not showing a discrepancy. However, in six references the author name is spelled differently. In three cases, the volume number shows a variation.

In this article, we will analyse discrepancies as indicated in Table 1. Questions that will be addressed are:

- How frequently do discrepancies between targets and cited references occur?
- Which types of discrepancies do occur?
- Can we identify particular characteristics of target articles that increase the probability of being cited erroneously? For instance, are there specific author names that are often misspelled?
- Do highly cited articles receive more citations showing discrepancies than articles cited less frequently?

In the example given in Table 1 we observed that the discrepancy 'SCHOEMAKER–SHOEMAKER' occurs in more than one cited reference to the target article. We will indicate this phenomenon as 'multiple discrepancies'. In two case studies we examined whether evidence could be obtained that copying references from other articles may be a cause underlying this phenomenon.

In our analysis we match (in different ways) a dataset with data on target articles with a dataset containing citations to these targets, both extracted in computer-readable form from SCISEARCH, and we analyse 'discrepancies' between the various datafields in the two datasets. As a consequence, our analyses are primarily related to the internal consistency of the SCISEARCH database. If discrepancies between target articles and cited references in our database are detected, the following possibilities exist:

1. the data on the target articles contain a variation/error;
2. the data on the cited reference contain a variation/error.

The variations or errors may occur in several stages of the process of data-handling, and we can distinguish at least three possible sources:

a. the scientific publisher or the authors of the articles made the variation/error; if so, the variation/error is present in the printed text of the article;
b. the variation/error occurred at ISI, either when the data were key-punched into the computer, or when computer manipulations on the data were performed;
c. the error or variation occurred at the host organization that brings SCISEARCH online (in our case: DIMDI), possibly in the process of loading the database.

In order to obtain some insight into the sources of the discrepancies, we compared in a limited number of cases the information in our datasets to the 'original' printed articles in the journals.

The structure of this paper is as follows. Section 2 deals with data collection and handling. We will describe the datasets that were used. In Section 3 we give an outline of the methods applied in our analysis. The various match-keys we applied will be defined, and a typology of discrepancies will be presented. The results of our analysis and a discussion are presented in Section 4. Finally, a general discussion of the results and conclusions follows in Section 5.

2. Data collection and handling

Our data source is SCISEARCH, the online version of the Science Citation Index, implemented in the host computer of the Deutsches Institut für Medizinische Dokumentation und Information (DIMDI) at Cologne (FRG). Our set of target articles consists of articles, published in five scientific journals in one or two particular years:

- Journal of Nuclear Materials (JNM), years: 1982–1983;
- Surface Science (SS), year: 1982;
- European Journal of Pharmacology (EJP), years: 1983–1984;
- British Journal of Pharmacology (BJP), years: 1983–1984;
- Journal of Pharmacology and Experimental Therapeutics (PET), years: 1983–1984.

The first two journals are from the field of materials science, while the other three belong to the field of pharmacology.

The citation data were collected by selecting all references in which the title of the (cited) journal and the (cited) publication year correspond to those of the articles in the target dataset. Since cited journal titles are not unified in SCISEARCH, we had to determine in advance under which variations the titles of our journals appeared. Possibly, we overlooked some variations that are hardly 'predictable' and that do not occur frequently. Consequently, we may have missed some citations to our target articles. However, these omissions do not affect the analyses presented in this paper. The principal point here is that the cited references were selected *independently* of the first author name, volume number and starting page number. Therefore, our database is appropriate for an analysis of discrepancies between target articles and cited references with respect to each of these datafields.

Both target articles and selected cited references were downloaded from SCISEARCH and stored in two data libraries, containing separate variables for the first author, volume number, starting page number and publication year. The two data libraries were match-merged, using different match-keys. Details of this process of successive match-merging are given in the next section. A detailed description of our method of data collection and handling is given in [1].

3. Method

In a computerized analysis we identified first the cited references in which all datafields (full author name, publication year, starting page number and volume number) were identical to the corresponding datafields of specific articles in our target dataset. Those citations showing no discrepancy at all will be symbolized by C throughout this paper, and in some cases will be indicated as 'correct' citations. Next, we created a subset of all those cited references that showed discrepancies in at least one datafield.

Using this subset of citations not matched in the first step, we determined in the second step the cited references that contained discrepancies in one datafield only. These citations will be symbolized as C^{D}, and in some cases be indicated as one-datafield discrepancies. This analysis was performed with respect to first author name, volume number and page number. Technically, this was achieved by running several match programs in which each datafield was successively excluded from the match-key. In 5% of the cases, a cited reference was matched to several target articles rather than one. These cases were examined manually, and the most probable correct match was identified. For each individual target article we determined the number of received citations, showing the various types of discrepancies.

Finally, we created a subset of cited references not matched in the previous steps. These references contain discrepancies in at least two datafields (e.g., full author name *and* volume number). In this set we identified only citations that do not contain an indication of the volume number and the page number. These are probably citations to articles that were not yet published at the time the citing publication was printed. We will indicate these citations as 'premature'. Further analyses with these citations were not performed, since in many cases it appeared to be very problematic to establish which particular target article was intended to be cited. The results we present in the next sections relate only to discrepancies in at most one datafield.

With respect to the one-datafield discrepancies we made a typology of discrepancies occurring in the various datafields. We divided all cases into 19 sub-types, using a combined computerized and manual approach. To give a few examples, with respect to discrepancies between author names of target and of cited reference, we distinguished between 'small' differences (e.g. 'AKASU' vs. 'AKAZU'), and completely different names ('HAGGBLAD' vs. 'PEDATA'), on the basis of visual inspection. In addition, we identified cases in which the cited author is co-author but not the first author of the (most probably) corresponding target article. Similar sub-types were distinguished

Table 2
Sub-types of discrepancies between targert articles and cited references

Type of discrepancy	Typical examples target versus cited reference
A: Discrepancies between author names	
A1 The name of the target author is too long: the last name exceeds 15 characters or the number of initials exceeds 3.	BARTHESLABROUSSE MG – BARTHESLABROUSS.MG
A2 The cited author is co-author (not first author) of the target article	PEL J; SMIT H – SMIT H
A3 Target author has more initials than cited author	WEISSMAN BA – WEISSMAN B
A4 The first initials of target and cited authors are different	DAUM PR – DAUM DR
A5 Other discrepancies between initials	REISINE T – REISINE TD
A6 Target author name starts with de/van/mac etc.	VANGILST WH – GILST WHV
A7 Target author name consists of two parts	STERINBORDA L – BORDA LS
A8 Variations in two identical subsequent consonants or in three subsequent consonants	BEDDELL CR – BEDELL CR SCHOEMAKER J – SHOEMAKER J
A9 Small variations between last names of cited and target author	AKASU T – AKAZU T
A10 Target and cited author are totally different or cited author is missing	HAGGBLAD D – PEDATA F PIETERS JM –
P: Discrepancies between starting page numbers	
P1 Cited page number is equal to the last page of the target article	p. 94 to 101 – p. 101
P2 Inversion of identical numbers	p. 277 – p. 227
P3 Target article is published in letter section	p. L115 – 115; p. L115 – p 1115
P4 Small variations	p. 123 – p. 125; p. 687 – p. 697
P5 Page numbers are totally different or cited page nr. missing	p. 497 – p. 112
V: Discrepancies between volume numbers	
V1 Target article published in combined volume (e.g., vol. 108/109)	v. 108 – v. 109
V2 Inversion of identical numbers	v. 119 – v. 199
V3 Small variations	v. 80 – v. 8; v. 80 – v. 89
V4 Volume numbers totally different or cited vol.nr. missing	v. 101 – v. 65

with respect to discrepancies in other datafields. A complete overview of all sub-types is presented in Table 2. With respect to sub-type V1, we give a more detailed explanation. Scientific journals may publish proceedings of scientific meetings or conferences. In some cases such proceedings are published in two journal volumes rather than one (e.g., vol. 108 and 109). We examined whether authors, in citing articles in such combined volumes indicate the wrong volume number, in the sense that they cite vol. 108 if the target article is published in vol. 109, or vol. 109 instead of vol. 108.

For each sub-type we determined the frequency of occurrence in our datasets, in order to identify the sub-types that occur most frequently. In addition, these data were used to analyse possible causes of the substantial differences that appear to exist among the five journals considered.

4. Results and discussion

4.1. Analysis by type of discrepancy

The overall figures from our analyses are presented in Table 3. As can be seen from this table, in the first step of the matching process we matched 24433 citations to our 4514 target articles published in all five journals. It appears that, relative to this total number of citations with no discrepancies at all, 9.4% of the citations in our cited reference dataset show a discrepancy in at least one datafield. Roughly speaking, for every ten citations containing a bibliographic description identical to that of a particular article in our target dataset, one citation shows some kind of discrepancy.

Aggregating all journals, the percentages of discrepancies between either author name, page number or volume number are 2.8%, 1.4% and 1.0% respectively. The 'premature' citations contribute 2%, and all other citations account for 2.3%.

Inspecting the results for each journal separately, it appears that large variations exist among the journals. JNM shows the highest percentages in most categories. Considering the other journals, EJP and SS have a high percentage of discrepancies between author names. SS and PET show a high percentage of discrepancies between page numbers and volume numbers respectively. BJP has the highest percentage of premature citations. Possibly the period between acceptance of an article by the editor and the actual publication date is longer for this journal than for the other journals.

Table 3
Overall figures

	Journal					
	BJP	EJP	JNM	PET	SS	Total
Number of target articles	633	1288	963	960	670	4514
Number of citations showing no discrepancies	3917 (100)	7765 (100)	2017 (100)	6722 (100)	4012 (100)	24433 (100)
Discrepancies between author names	45 (1.1)	277 (3.6)	91 (4.5)	127 (1.9)	133 (3.3)	673 (2.8)
Discrepancies between page numbers	36 (0.9)	66 (0.8)	63 (3.1)	80 (1.2)	86 (2.1)	331 (1.4)
Discrepancies between volume numbers	24 (0.6)	49 (0.6)	35 (1.7)	120 (1.8)	21 (0.5)	236 (1.0)
Premature citations	154 (3.9)	159 (2.0)	23 (1.1)	132 (2.0)	24 (0.6)	492 (2.0)
All other citations	119 (3.0)	92 (1.2)	106 (5.3)	143 (2.1)	93 (2.3)	553 (2.3)
Number of citations showing at least one discrepancy	378 (9.7)	643 (8.3)	305 (15.1)	602 (9.0)	357 (8.9)	2285 (9.4)

In order to obtain more insight into the nature of the discrepancies and into possible causes underlying the differences among the journals, we divided the one-datafield discrepancies into 19 sub-types. The definition of these sub-types is given in Table 2 in Section 3. Some sub-types may be expected to occur potentially with respect to *all* target articles in our dataset, whilst other sub-types

Table 4
Discrepancies by sub-type [a]

Sub-type of discrepancy (example)	Journal														
	BJP			EJP			JNM			PET			SS		
	C_s/C (%)	C^D	C^D/C_s (%)	C_s/C (%)	C^D	C^D/C_s (%)	C_s/C (%)	C^D	C^D/C_s (%)	C_s/C (%)	C^D	C^D/C_s (%)	C_s/C (%)	C^D	C^D/C_s (%)
A1 BARTHESLA-BROUSSE MG – BARTHESLA-BROUSS.MG	0.0	0	–	0.3	83	415	0.0	1	–	0.5	5	15.0	0.0	26	–
A2 PEL J; SMIT H – SMIT H	100	0	0.0	100	0	0.0	100	2	0.1	100	0	0.0	100	20	0.5
A3 WEISSMAN BA – WEISSMAN B	61	1	0.0	49	59	1.5	49	36	3.6	73	23	0.5	45	7	0.4
A4 DAUM PR – DAUM DR	100	1	0.0	100	29	0.4	100	7	0.3	100	23	0.3	100	2	0.0
A5 REISINE T – REISINE TD	100	0	0.0	100	6	0.1	100	9	0.4	100	7	0.1	100	1	0.0
A6 VANGILST WH – GILST WHV	1.0	0	0.0	3.9	9	2.9	1.5	0	0.0	3.1	0	0.0	1.5	4	6.7
A7 STERINBORDA L – BORDA LS	?	7		?	9		?	1		?	3		?	9	
A8 BEDDELL CR – BEDELL CR	38	7	0.5	38	28	1.0	29	6	1.0	36	11	0.6	28	13	1.2
A9 AKASU T – AKAZU T	100	14	0.4	100	36	0.5	100	7	0.8	100	31	0.5	100	44	1.0
A10 HAGBLADD D – PEDATA F	100	14	0.4	100	14	0.2	100	12	0.6	100	24	0.4	100	7	0.1
P1 p.94 to 101 – P. 101	100	1	0.0	100	2	0.0	100	3	0.1	100	3	0.0	100	2	0.0
P2 p.665 – p.655	16	5	0.1	15	1	0.0	14	3	0.1	18	4	0.3	10	1	0.2
P3 p.L115 – p.115	0			0			0			0			15	12	2.0
P4 p.497 – 129	100	7	0.2	100	17	0.2	100	42	2.1	100	29	0.4	100	34	0.8
V1 Combined volumes	0			0			54	13	1.2	0			0		
V2 v.119 – v.199	0.0	0	0.0	4.9	0	0.0	38	3	0.4	80	50	0.9	79	3	0.1
V3 v.80 – v.8	100	21	0.5	100	41	0.5	100	15	0.7	100	48	0.7	100	15	0.4
V4 v.101 – v.61	100	1	0.0	100	5	0.1	100	2	0.1	100	11	0.2	100	2	0.0

[a] C: the number of 'correct' citations to all articles in the journal; C_s: the number of 'correct' citations to the subset of articles that could potentially generate the specific discrepancy; C^D: the number of citations showing the specific discrepancy and given to the articles in the particular subset.

will occur only with respect to a *specific subset* of targets. For instance, small variations in volume or starting page number (sub-type P4) can be assumed to occur in citations to each article, since all articles have a page number in their bibliographic description. However, discrepancies of type P3, in which target articles contain an 'L' in the page number field while in the cited reference the 'L' is omitted, obviously may occur only with respect to target articles published in 'letter' sections of a journal. Therefore, for each sub-type we first established the subset of target articles that could potentially generate the discrepancy of that sub-type. With respect to those sub-types for which we had no *a priori* grounds to restrict the occurrence to certain classes of target articles, we considered the set of all target articles published in a particular journal.

The results are presented in Table 4. For each journal and sub-type three parameters are given. The first (C_s/C) indicates the percentage of 'correct' citations to target articles in the subset that could potentially generate the sub-type of discrepancy (C_s), relative to the total number of 'correct' citations to all articles in the journal (C). If the subset of target articles that could generate the discrepancy is equal to the set of all articles in the journals, C equals C_s and C_s/S amounts to 100%. The second parameter (C^D) indicates the absolute number of citations—showing the specific discrepancy–to target articles in the subset. Finally, the percentage of C^D/C_s is given.

To given an example, 38% of the 'correct' citations to all articles in the journal BJP are citations to articles of which the first author name is of type 'BEDDELL' or 'SCHOEMAKER' (A8). These articles are cited seven times in cited references in which the author is of the type 'BEDELL' or 'SHOEMAKER'. These seven citations constitute 0.5% of all citations given to the specific type of targets and not showing any discrepancy. With respect to sub-type P3, one observes that 15% of all 'correct' citations to the articles in SS, are to targets that contain an 'L' in the page number field, while the other journals do not contain this type of article. The letters receive 12 citations in which the 'L' is omitted or substituted by 'l', and that constitute 2% of all 'correct' citations of these letters.

Table 3 shows that JNM has the highest percentages of discrepancies between author name, volume and page number among our five journals. Inspecting Table 4, one observes that the high percentage of discrepancies between volume numbers is partly explained by the 'combined volume problem' (sub-type V1), since only JNM contains combined volumes. However, the percentages for most sub-types related to author names and page numbers are still high, relative to the other journals. We are not able to identify any particular characteristic of the target articles that might explain the high percentages for JNM in these two datafields. Probably the explanation has to be found in characteristics of the sources citing our target articles. We suggest that in the field covered by JNM—nuclear fusion and fission research–proceedings of scientific meetings play an important role in the communication, and that JNM receives relatively more citations from camera-ready produced proceedings, published in journals processed for the SCI. In fact, JNM itself contains several of these proceedings in later years. Articles in camera-ready proceedings are probably 'checked' less accurately in the process of copy-editing than articles in typeset volumes, or if errors in the text are detected, they may not be corrected in all cases. As a consequence, cited references in such articles may contain more errors. Below, we consider only the other four journals.

The relatively high percentage of discrepancies between page numbers for SS can be explained at least partly by the fact that SS is the only journal in our sample with a letter section. PET's high percentage of discrepancies between volume numbers is explained only partly by sub-type V2. In fact, PET has a high percentage of citations to articles with volume numbers of sub-type V2, compared to BJP, EJP and JNM (80% against 0%, 5% and 38% respectively). However, for PET and SS these percentages of citations to targets in this subset are hardly different (80% against 79%), while the percentages of 'erroneous' citations to these targets differ by a factor of almost 10 (0.9 vs. 0.1). We do not have an explanation for the differences with respect to this variation among PET and SS.

A striking result is that the percentage of discrepancies between author names in PET and BJP is low compared to the other journals. We found that the articles in BJP and PET emanate mainly from the United Kingdom and the USA respectively, while the other journals publish more articles from non-English speaking countries (particularly the other European countries and Japan). We conjectured that the native language of the authors is a factor of significance, but we did not find strong evidence in favour of this hypothesis. The high overall percentage of discrepancies between author names in EJP appears to be caused to a large extent by sub-type Al. However, this sub-type is rather trivial. The discrepancies are due to the fact that the name of the cited author is in most cases abbreviated, at least in the SCISEARCH database. As a consequence, target articles with too long author names that are cited correctly, have a high probability of showing this type of discrepancy. Obviously, this should be kept in mind in citation searches using this database. The high overall percentage for SS is caused mainly by sub-type A2. We do not have an explanation why the number of citations to co-authors instead of first authors is relatively high for this journal. Possibly it is a characteristic of the field (materials science, physics), in the sense that hierarchical relationships between researchers may be stronger in this field than in biomedical fields. As a consequence, a stronger tendency exists to cite the 'senior' rather than the 'junior' researcher, even if the senior is not the first author.

We calculated the mean value of the variable $C^D/C_s(\%)$, based on the scores of the individual journals. These statistics provide a rough estimate of the percentage of citations showing discrepancies to target articles in each class. The mean percentages for the main sub-types of discrepancies are summarized in Fig. 1. The sub-types are indicated by the corresponding example from Table 2. Mean percentages of discrepancies higher than 1.0 are found for articles in letter sections (sub-type P3), for author names containing 'van', 'of', 'Mac' (A6) or with more than one initial (A4), and for targets in combined volumes (V1). It should be noted that the standard deviations are rather high and exceed in many cases the mean.

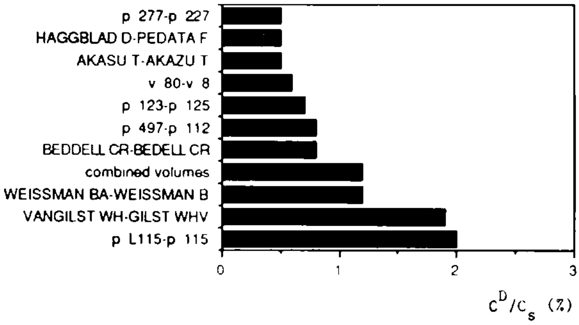

Fig. 1. Scores for main sub-types of discrepancies.
C_s: The number of 'correct' citations to the subset of articles that could potentially generate the specific discrepancy; C^D: The number of citations with the specific discrepancy. The sub-types of discrepancies are indicated by the particular examples, presented in Table 2.

4.2. *Discrepancies between publication years*

As indicated in Section 2, we selected the cited references in which the publication year corresponds to the year of publication of the target articles in our target dataset. For instance, we selected all citations to JNM containing 1982 or 1983 as cited publication year. Counting the number of citations in our cited reference dataset to JNM by volume, we detected that two particular volumes received hardly any citations. The articles in these two volumes had publication year 1982 in our target dataset. However, inspecting SCISEARCH online we found 931 citations to these volumes, but the cited publication year was 1981! For this reason these citations were not present in our dataset of cited references. In fact, these citations are not included in the analyses presented in this paper.

The percentage of citations to JNM showing this discrepancy, relative to the number of citations not showing any discrepancy is 46%. The cause of the discrepancy is the following: the two volumes are proceedings volumes and were announced to be published in 1981. Due to several reasons, they were actually published in 1982. In the heading of each article, 1981 was indicated as publication year. On the cover of the volume, however, 1982 appeared. ISI assigned 1982 as the publication year to the articles, while citing authors gave 1981 in their references.

A similar discrepancy between target and cited publication year was detected for an issue of EJP, published at the end of 1982. This issue was a part of the 1983-programme of the scientific publisher. In fact, in the heading of the printed articles, 1983 appears as the publication year, but the issue was actually published in 1982, and on the cover the year 1982 was indicated. In SCISEARCH, the articles in this issue as *source* articles have 1982 as the year of publication, while citations to these articles contain 1983. We deleted these citations from our dataset.

A further analysis of the journal JNM revealed another dimension of the 'combined volume' problem. The two volumes (vol. 103 and 104) were part of a combined volume. The volumes were clearly separated in the journal, and ISI assigned the correct volume numbers to the target articles. However, in the heading of the printed articles, 'vol. 103&104' was indicated. Citations often contained 'vol. 103', even if the target articles were formally published in vol. 104. In fact, 387 citations showed this type of discrepancy. In addition, we observed that a citation to an article in a combined volume—say, 'Vol. 103&104, pg. 117'—was processed by ISI in 37 cases in the following way: two references were key-punched: one containing 'vol. 103, pg. 117', and another one containing 'vol. 104' and either again 'pg. 117' or no indication of the page number at all.

The number of discrepancies for these two combined volumes is much higher than that for the other combined volumes in this journal for which the results were presented in Table 4. This is probably due to the fact that in the heading of the printed articles in these volumes *two* volume numbers appear (vol. 103&104), while the articles in the other combined volumes contain only one volume number in their headings.

4.3. Discrepancies related to the number of times cited

We examined whether highly cited articles receive more citations showing discrepancies than articles that are cited less frequently. Therefore, we divided all target articles into classes according to the number of received citations showing no discrepancy (C), and calculated for each class the percentage of citations showing a discrepancy in only one datafield (author name, volume and page number) relative to the total number of citations that are given to all targets in the class without any discrepancy (C^D/C (%)). Discrepancies of type A1 (too long author names) are not included.

The results are presented in Fig. 2. In our datasets, the articles receiving 0–3 citations have the highest percentage of citations showing discrepancies in one datafield (8.3%). This percentage decreases to 2% as the citation frequency increases, while for the class of highly cited articles, cited more than 40 times, it increases to 4.2%. Roughly speaking, the lower cited articles receive, on an average, the highest percentage of citations showing discrepancies. Next follow the highly cited articles (cited more than 40 times), while the moderately cited articles (cited between 4 and 40 times) show the lowest percentage of discrepancies.

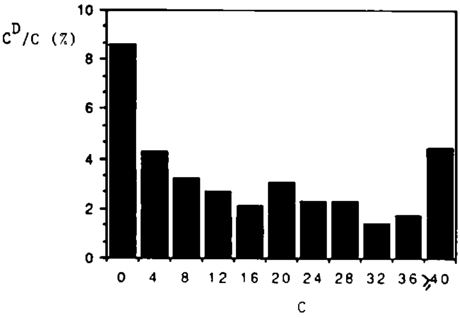

Fig. 2. Discrepancies related to citation frequency.
C: the number of 'correct' citations to all articles in the journal; C^D: the number of citations with a discrepancy in one datafield only (first author name, volume number or starting page number). The 'correct' citation scores of the target articles are grouped into classes. At the horizontal axis, only the lower bound of each class is indicated. Consequently, '0' indicates the class of scores between 0 and 3 'correct' citations.

4.4. On multiple discrepancies, and their sources

With 'multiple discrepancies' we indicate the phenomenon that there are *several* publications that cite a specific target article and that they all show the same discrepancy. We found that almost 83% of the discrepancies with respect to specific target articles occur only once, i.e. in one cited reference only. In 6.5% of the cases, a discrepancy is 'repeated' in three or more cited references, and in 0.6% it occurs more than ten times. Discrepancies of sub-type A1 are not included. For every 100 articles cited 'correctly' more than ten times, we had one article receiving hardly any correct citations, but cited more than ten times in references showing a discrepancy. So roughly speaking, due to discrepancies we missed 1% of the articles cited more than 10 times.

With respect to multiple discrepancies occurring at least three times, we checked manually whether the information in our target dataset is correct, by looking up a sample of the original articles in the journals. Out of 15 cases, we found one case in which the bibliographic descriptions in the printed text and in our target dataset did not correspond. A particular author had the initials 'PBMWM' according to the original text, and 'BMWM' in the corresponding source article in SCISEARCH online. This article was cited 14 times in references giving the correct initials.

We picked out three cases in which we checked a sample of the cited references. Thus, we compared 16 cited references in our dataset to the corresponding references, printed in the original article. The first case relates to a discrepancy in which an article in our target dataset has 'BASKIN DG' as a first author—this name is equal to the name in the original text and is therefore correct—while all citations in our dataset to this article contain 'BASKIN DA'. It appeared that the citation entered first in SCISEARCH indeed contained the (incorrect) name 'BASKIN DA' in the original text. However, a sample of five citing articles entered later all gave the correct name in the original text. As a consequence, the author names in our citations dataset are—at least in five cases—different from the name given in the original text. It seems to be a persistent error, since in the SCISEARCH database updated till June 1988, the article receives 45 citations showing the discrepancy, and no correct citations at all. We suggest that ISI has some internal procedures for unification of cited references, and that in this case the incorrect unification key was applied.

With respect to the other two cases, the information in our dataset of cited references was equal to that in the original text. One of these cases relates to an article published by H.A. Singer in 1983. This case will be discussed in Section 4.5 below.

4.5. Multiplication of erroneous citations: a case study

We analysed in detail citations to an article by H.A. Singer, published in PET in 1983, in volume 226, starting on page 690. We collected additional citation data for the (citing) year 1986. The article

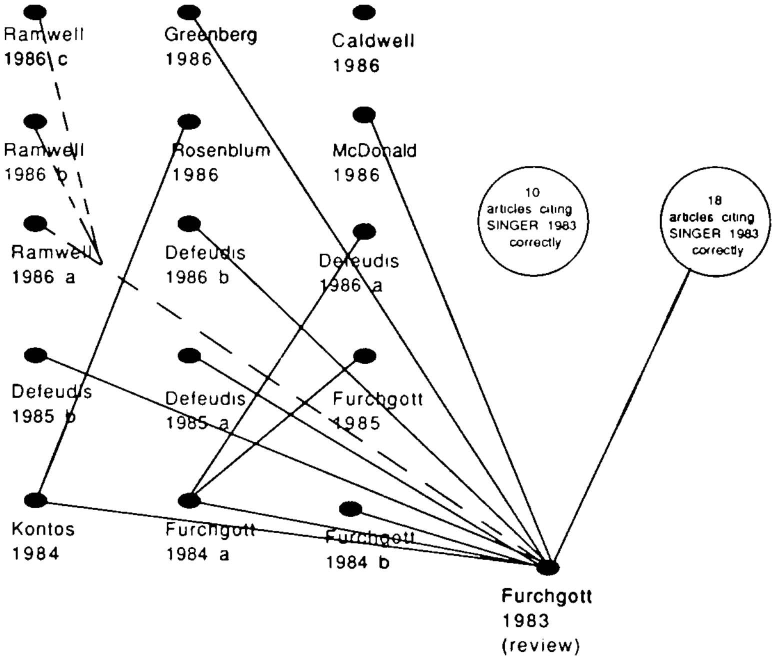

Fig. 3. Citation relationship between articles citing Singer 1983 erroneously.
Each dot indicates an article citing the specific 1983-publication by Singer erroneously. Solid lines between two dots indicate citation relationships between the corresponding articles. For instance, Kontos 1984 cites Furchgott 1983. The dashed lines between each of the three 1986-articles by Ramwell and the 1983-article by Furchgott indicate that the three Ramwell articles cite publications authored by Furchgott, though *not* the ones given in this figure. The articles citing Singer 1983 correctly are represented by two circles. The circle at the right-hand side contains 18 articles that do cite Furchgott's review from 1983, while the circle at the left indicates ten articles that do not cite this review.

is cited 44 times during the period 1983–1986. Twenty-eight citations give bibliographic information identical to that of our target, while in 16 citations the cited volume number is 227 instead of 226. We found that the information on the target and on four cited references in our datasets corresponds to the printed text of the original articles. We therefore assume that all errors were made in the original text.

In Fig. 3, articles citing our target erroneously are represented as dots. For each erroneously citing article we indicated the publication year and the citing author. For most publications this is the first author. However, if an author appears in more than one citing article as first author or co-author, we indicated his name even if he is not the first author. The articles citing Singer 1983 correctly, are represented in two circles. Of these articles we indicated only whether they cite Furchgott 1983 (right circle) or not (left circle).

We conjectured that the multiplication process of errors is caused by the fact that authors in compiling reference lists may copy references from other articles. In order to provide evidence in favour of this hypothesis, we examined whether we were able to reconstruct the propagation of the error on the basis of citation relationships, co-authorships or institutional cooperations.

Analysing Fig. 3 from this perspective, we observe that the first article citing our target erroneously is a review article by Furchgott, published in 1983. This seems to be a useful paper, since it receives more than 300 citations in the entire SCISEARCH database during 1983–1986. Of the 15 articles citing Singer 1983 erroneously, eight actually cite this review.

Focussing on the seven articles that do not cite the 1983 review paper by Furchgott, two of these (Furchgott 1985 and Defeudis 1986B) are authored by researchers who have published articles that do cite this review. The three articles by Ramwell cite several other articles by Furchgott—though not the ones that cite our target erroneously—and therefore can be expected to be aware of the particular review as well. For this reason, in Fig. 3 the three articles by Ramwell are connected to the Furchgott-review by dashed rather than solid lines. Rosenblum 1986 cites Kontos 1984, who in turn cites Furchgott's 1983-review. In fact, the publications by Rosenblum and Kontos emanate from the same department. With respect to Caldwell 1986, we could not detect any relationship in terms of citations, co-authorships or institutional affiliations with the articles citing our target erroneously.

Our results suggest that the authors citing our target erroneously have copied the particular reference from an important review paper containing the error for the first time. There is evidence that almost all authors had this particular review on their desks when they wrote their articles. We did not *prove* that the particular authors copied erroneous references from other articles. But in our view they may have done so.

With respect to the process of multiplication of errors, two other observations should be made. First, it appears that not all authors citing Furchgott's 1983-review, cite our target erroneously. In fact, of the 28 articles citing correctly, 18 also cite the review paper. So, if any copying of reference lists occurs, it does occur in a limited number of cases.

The second observation is that the multiplication process seems to vanish after a few years. In an additional analysis we found that, in the period 1987 – June 1988, only three erroneous citations are given while the number of correct citations amounts to 17. In fact, we observed that authors apparently become aware of the error and correct it in later publications; e.g. Ramwell, who gave three erroneous citations in 1986, but also published in 1986 an article that contains the correct citation. It is entered in SCISEARCH later than the other three articles.

5. General discussion and conclusions

As indicated in Section 4.4, we checked for 29 articles (targets or cited references) whether the bibliographic information in our dataset from SCISEARCH corresponds to the information in the original printed articles. Of course our sample is very small, and we should be careful in drawing general conclusions with respect to the sources of the discrepancies in our datasets. In the 'BASKIN' case, we suggested that the discrepancies were due to irregularities in some internal unification process at ISI. However, we are not able to give any accurate estimate of the frequency with which such irregularities occur. Apart from this case, we found only one discrepancy in which ISI has probably made an error in key-punching the information into the computer. Therefore, it seems justified to assume that the major part of the discrepancies in our dataset are due to errors or variations in cited references that are present in the original text.

In Section 4.1 we found that for our five journals aggregated, the percentage of citations in our datasets showing a discrepancy in at least one datafield, relative to the number of citations not showing any discrepancy, amounts to 9.4%. Citations to 'unpublished' papers are included and account for 2%. One can ask how large the percentage of citations showing no discrepancy is, relative to the total number of 'intended' citations to our targets. However, this total number of citations can hardly be determined exactly. In fact, in Section 4.2 we showed that a large number of citations in our dataset were missing due to a discrepancy between publication years. If we do not consider these systematic omissions, we may still have missed citations to our targets, because of other types of errors or unexpected variations in the cited journal title and publication year. On the other hand, our datasets of cited references may contain citations to articles other than our targets. If these two factors cancel out, and the total number of citations in our datasets is the correct number, then we find that the number of citations not showing any discrepancy constitute 91% of the total number of citations, while the citations with discrepancies in one datafield only (author name, volume or page number) account for almost 5%.

We observed large variations among the journals. These were partly explained by specific characteristics of target articles in those journals. However, large variations still remained unexplained. Therefore our overall percentages given above indicate only the order of magnitude.

Roughly speaking, for every 100 received citations not showing any discrepancy, two citations contain a minor variation in either author name (A9), volume number (V3) or starting page number (P4), while the other discrepancies that may occur in principle with respect to each target article, contribute another two citations. Consequently, the minimum level for discrepancies amounts to four, regardless of whether the target article has specific characteristics that attract more discrepancies, such as containing an 'L' in the page number field or being published in a combined volume.

Taking into account the number of times target articles are cited 'correctly', we estimate on the basis of Fig. 2, that this minimum level will be seven for articles cited 0–3 times, two or three for articles cited between 8 and 40 times, and again four for articles cited more than 40 times.

If we consider the main discrepancies due to the specific characteristics of targets as identified in our analysis, we find for instance that for every 100 correct citations to articles of the type 'BEDDELL CR, V225, P17', two additional citations show discrepancies since the author name contains two subsequent identical consonants (type A8) *and* more than one initial (A3), and the volume number is of type V2. If the page number is 'L17' instead of '17', another two citations should be added. Thus, the total number of discrepancies may amount to six or eight, and may be even higher than ten for lowly or highly cited targets of these types. It is assumed that the various types of discrepancies occur independently one from another.

It should be noted that the levels indicated above relate to discrepancies in one datafield only, and therefore represent a lower bound, since discrepancies in two or more datafields will raise these levels.

In the case study presented in Section 4.5, we found evidence that copying references from other articles may be a cause of the observed multiplication of errors in cited references. We did not *prove* that the particular authors copied erroneous references from other articles. However, our results suggest that they may have done so. It should be noted that we studied only one case. Characteristic in this case is that the first error was made in an important (highly cited) review article. It is interesting to note that some authors publish more than one article containing the error. So, not only copying references from other authors, but also from one's own articles is at stake here. The error occurs frequently during the first three years, and then seems to vanish. More cases should be studied in order to obtain a more complete picture of this phenomenon.

In Section 4.3 we found evidence that authors are less accurate in citing articles with a low impact—measured by citation counts—than those with a high impact. Considering the articles with high impact, authors possibly tend to think that the important documents in a field can be indicated less accurately since all readers know such documents very well and can be expected to have a copy on their desks.

Summarizing, we present the following checklist of technical points that one should keep in mind in performing citation analysis accurately:

a. Author names should be abbreviated. The length of the last name should not exceed 15 characters, and the maximum number of initials is three.
b. Severe inaccuracies may arise due to discrepancies between publication years of cited reference and target. We expect that the probability for such discrepancies to occur is higher for articles in special journal volumes such as proceedings, and in volumes or issues published at the beginning or at the end of a year.
c. Target articles published in a combined volume constitute a 'risk group' since often the 'wrong' volume number is cited, or ISI may process cited references to such targets in a strange way.

Other risk groups are:

d. Target articles in letter sections, containing an 'L' in the page number field. In a citation, the 'L' may be omitted or may be substituted by '1'. We expect that a similar discrepancy may arise for articles in sectionalized journals, containing an 'A' or 'B' in the volume number field, but we had no sectionalized journals in our sample.
e. Authors with two or more initials. Second or third initial of the cited author may be missing.
f. Author names containing two subsequent identical consonants or three subsequent consonants. One cites BEDELL instead of BEDDELL, or SHOEMAKER instead of SCHOEMAKER.
g. Author names starting with 'van', 'de', 'vondem', 'Mac' etc.; for instance one cites GILST WHV instead of VANGILST WH.
h. Volume or page numbers of three digits with two subsequent identical digits. One cites 227 in stead of 277.

The implications of our findings for the *reliability* of citation analysis of articles on the level of scientific journals, research groups or institutes, and entire countries will be discussed in a future publication.

Acknowledgement

Part of the research described in this paper was financially supported by Elsevier Science Publishers. The authors thank Robert Braam for the computerized drawing of the figures presented in this paper.

References

[1] H.F. Moed, The use of online databases for bibliometric analysis, in: L. Egghe and R. Rousseau, ed., *Informetrics 87/88* (Elsevier Science Publishers, Amsterdam, 1988) 133–146.