The use of bibliometric data for the measurement of university research performance *

H.F. MOED, W.J.M. BURGER, J.G. FRANKFORT and A.F.J. VAN RAAN

Research Policy and Science Studies Unit, Bureau Universiteit, University of Leiden, P.O. Box 9500, 2300 RA Leiden, The Netherlands

Final version received November 1984

In this paper we present the results of a study on the potentialities of "bibliometric" (publication and citation) data as tools for university research policy. In this study bibliometric indicators were calculated for all research groups in the Faculty of Medicine and the Faculty of Mathematics and Natural Sciences at the University of Leiden. Bibliometric results were discussed with a number of researchers from the two faculties involved.

Our main conclusion is that the use of bibliometric data for evaluation purposes carries a number of problems, both with respect to data collection and handling, and with respect to the interpretation of bibliometric results. However, most of these problems can be overcome. When used properly, bibliometric indicators can provide a "monitoring device" for university research-management and science policy. They enable research policy-makers to ask relevant questions of researchers on their scientific performance, in order to find explanations of the bibliometric results in terms of factors relevant to policy.

1. Introduction

The central issue of this study is the examination of the potentialities of quantitative, literature-based (i.e. bibliometric) indicators as tools for university research-policy. As a subject of extensive investigation we analysed the research performance of two large faculties [1] of the University of Leiden (Faculty of Medicine and the Faculty of Mathematics and Natural Sciences) for the period 1970–80.

The study presented in this paper [2] was actuated by the necessity for a large-scale project evaluation resulting from a drastic change in the allocation system at the University of Leiden. The academic staff allocation system for teaching *and* research at Leiden was originally based almost entirely on student numbers. In fact, in this old allocation system there was a 50 : 50 ratio between staff time spent in teaching and in research. As a direct consequence, disciplines with an increasing number of students also had an increasing research capacity. A few years ago the allocation system at Leiden was changed radically to a research project grant system. Two major considerations led to this change. First, the need was felt to finance scientific research more explicitly on the basis of scientific quality. Second, it was felt that the research capacity of a discipline should be protected against the consequences of a continuous decline of student numbers (see Van Raan and Frankfort [1]). Essentially, in the new system a separate research financing channel has been created. This was achieved by a very considerable reduction of a proportionality factor which in the

* The research on which this paper is based was supported by a grant of the Netherlands Ministry of Education and Sciences, Directorate-General for Science Policy, The Hague. The statements in this paper are the responsibility of the authors.

[1] In the Netherlands the "Faculty" is the largest organizational unit of education and research within a University. A faculty (e.g. Mathematics and Natural Sciences) is often divided in "subfaculties" (e.g. Mathematics, Physics, Biology, Chemistry). Within a (sub) faculty, the departments are the next "lower" organizational units (e.g. Physical Chemistry, Theoretical Physics). Departments are divided into several "research groups".

[2] An earlier, detailed presentation of our work is given in a report by Moed et al. [3]. This paper is a revised, comprehensive version of this report.

Research Policy 14 (1985) 131–149
North-Holland

earlier allocation system coupled the research financing (in terms of academic staff) with the teaching capacity based on teaching load calculations. The resources which thus became available could be "earned back" with current research projects if the departments concerned could demonstrate the quality of these projects. Therefore, in the last few years university departments have been stimulated to develop research performance criteria and, subsequently, to apply these criteria in a sort of self-evaluation, in order to avoid a considerable decrease of research support. The University Executive Board superintends these evaluation procedures and makes the budgetary decisions, taking the results of the evaluations into account.

Recently, the Netherlands government changed the national system for university financing into a system very similar to the Leiden system. Both the Leiden system and this recent policy of the national government made the need felt for more objective, quantitative research performance indicators. The project presented in this report (started in 1981) arose directly from this need.

In this paper we are concerned with two important aspects of research performance: output and impact. Output refers to the extent to which the research creates a body of scientific results. Impact is defined as the actual influence of the research output on surrounding research activities. In this work we discuss the relation between "quality" and "impact" (an extensive discussion is given in Moed et al. [3]).

The output indicators used in this study are essentially based on the numbers of publications within the international scientific literature. Impact indicators were constructed on the basis of the number of times these publications were cited during a certain period by other articles published in the international scientific literature. In contrast to other recent studies (e.g. Martin and Irvine [2]), we chose the research group as the level of aggregation, since a research group [3] usually constitutes the "natural" unit of research activity (at least in the two faculties involved).

Publication and citation data were obtained from the Institute for Scientific Information (ISI, Philadelphia). The *Science Citation Index* (SCI) covers several thousand scientific journals, which constitute the core of the international scientific serial literature for many fields within the natural and life sciences. In addition, the SCI contains non-journal material, such as published proceedings, multi-authored books, monographs and thematic collections of papers. Two important indexes are relevant for this investigation. First, the Source Index, which contains bibliographic descriptions of all articles published in the journals or books processed for the SCI. These journals (or books) are called source journals (or books). The articles published in them are called source articles. Secondly, the Citation Index, that lists all the references (i.e. citations) given in the source articles in a given year.

In this study we operated on all publications from the SCI Source Index that, according to their addresses, originated from the University of Leiden and that were published between 1970 and 1980. In addition, we obtained bibliographic data on all source articles published between 1970–80 that cited any of the Leiden publications. Thus we operated on some 6700 publications and 42,000 citations, the results of about 4000 academic man-year research activities (12,5000 man-year total personnel) and about Hfl 1,000,000,000 (400 million dollars) support from public resources. An extensive software package (Moed et al. [3]) was developed to carry out numerous tasks with respect to the data handling, which is in fact the combination of two large data clusters (bibliometric data and university data).

The project started in 1981. Approximately two man-year equivalents were needed to collect and handle all the data involved (the development of the software package included), and to produce the final graphical presentations of the bibliometric results for all research groups with respect to the period 1970–80. The costs of the bibliometric data provided by the Institute for scientific Information amounted to 15,200 US-dollars.

In March 1984 the University Executive Board, following the advice of the two faculties involved, decided to update the bibliometric data gathered in this project and to use the developed bibliometric indicators as a try-out in the forthcoming evaluation of all research projects in the two faculties. We have planned future publications on the outcome of this experiment.

[3] The research group is the "lowest" organizational level of researchers within a department. In practice, it is defined by specific long-term research projects.

We structured this paper around the following central questions:

(1) What do we mean by impact and quality? What are useful types of bibliometric indicators in connection with university research policy? (section 2).

(2) Which indicators were used in this study? (section 3).

(3) What are the specific problems with respect to collecting, handling and interpreting bibliometric and university data? We discuss problems related to the completeness of bibliometric data (section 4.1) and of university data (section 4.2). Moreover we discuss the limitations of the *Science Citation Index* as a data base for output and impact evaluations of research groups in the various fields of science (section 4.3). In addition, a number of disturbing factors are mentioned (section 4.4.). Finally, the problem of "statistics" is posed in section 4.5.

(4) How do university researchers feel about the results of the analyses carried out in this study? What are their major comments? Do these bibliometric results and interpretations provide a meaningful basis for the discussion of research of groups? Are the bibliometric analyses carried out in this project useful tools for a university research policy? (section 5).

2. The concepts of "impact" and "quality"

2.1. Short-term and long-term impact

We assume that scientific publications in a certain field during a certain period reflect the research front of that particular field. By looking at the number of times a research group's publications are cited, we can gain insight into its impact at the research front. A distinction should be made between short-term and long-term impact. Looking at impact over a long period offers the possibility of relating impact to "durability". It can be determined whether, and to what degree a research group has made a more permanent contribution to scientific advance. Although the various fields produce numerous publications, only a small number of them are included in the "basic knowledge" accepted in the field, and subsequently rendered in reviews, books etc. The study by Chang [4] is an example of long-term evaluations. He determined the impact of articles over a very long period of about 40 years. The question is not so much "is research on magnetic resonance in the Netherlands more or less successful than before", but rather "what contribution did Dutch researchers make to the development of this field between 1940 and 1976". The long-term impact bring to light to what extent groups have been able to triumph over rival groups at the research front.

Short-term impact refers to the impact of researchers at the research front a few years after the publication of research results. Citation counts over relatively short periods of time do not primarily reflect the extent to which a group has made a "permanent contribution" to a field, these counts in the first place indicate factors such as the extent to which the group exerts itself at the research front, whether it forms part of the research community, and the extent to which the group and its publications are known among colleagues and play a part in scientific discussions at the research front. Thus short-term impact, operationalized in citation counts, should be related to the visibility of research groups at the research front and can be ranked with other visibility indicators such as international contacts, awards, invitations to take part in important conferences, etc.

Moreover, not only the short-term impact of research groups, operationalized in citation counts, may be determined. It is also possible to determine short-term impact of journals at the research front. In fact, mainly on the basis of short-term impact the ISI selects the journals which it includes in the ISI data collection. In general a research group publishes in various journals, a set of journals (in this study called "a journal packet"). The short-term impact of this journal packet can be determined, and this enables us to assess the extent to which a research group publishes in more or less prestigious journals. Furthermore, the impact of a group can be compared to the impact of its journal packet and on the basis of this, it is assumed in this study that one can determine a group's impact at the international level.

An important question now is the relationship between short-term and long-term impact. We assume that there is a research front in every scientific field. At this front scientists develop theories about the structure of reality and these theories are confronted with each other through experimental

research. In the end certain theories will triumph – temporarily or otherwise –and be added to the basic knowledge in the field. The short-term impact indicates how groups maintain themselves at the research front, the long-term impact indicates to what extent they eventually succeed in scoring "triumphs". We further assume that, in view of the abundance of communication possibilities, at present it is very unlikely that researchers are not visible at the research front, but nevertheless turn out to acquire high impact at later stage, that is long-term impact. Moreover, we feel that researchers, particularly in cases where research is funded by the community, are obliged to make their work (internationally) known and to play an active part at the research front. It goes without saying that a high short-term impact does not guarantee a high long-term impact: many theories will founder and will therefore not acquire long-term impact. However, if a group appears to have a high short-term impact over a long period of time, it might be possible that the older publications of the group generate a considerable long-term impact as well. We plan further research on the relationship between short-term and long-term impact in the follow-up of this project. Moreover, we shall try to operationalize more precisely the time period to be considered in short- and long-term impact evaluations.

Another important question refers to the relevance of short-term and long-term impact evaluations for a university research policy. Long-term impact can only be determined after a considerable period of time, and when the time is ripe to assess this kind of impact, it is often no longer useful as a policy indicator, for one thing, because the groups concerned may no longer exist or may now be devoted to other subjects, and secondly, because it is unclear whether research groups that once made a "permanent" contribution will continue to do so. But what is more, one cannot require research groups to do work which acquires long-term impact. On the other hand, they indeed can be required to take part in the scientific discussions in their field. Consequently, in view of the scant understanding of the relationship between short-term and long-term impact, long-term impact evaluations of research groups can at present hardly be useful for university policy decisions. This entails that in the present study, which aims at finding possibilities of evaluating academic research on behalf of academic research policy, short-term impact is of primary importance.

2.2. Quality

Up to now we have hardly mentioned the notion "quality" in connection with scientific research. This is rather remarkable in a study which deals with the evaluation of scientific research. However, we have deliberately avoided the use of the notion quality, since it is virtually impossible to operationalize this general concept. Quality may refer to a variety of values. With regard to scientific research, we can distinguish between cognitive quality, methodological quality, and esthetic quality. Cognitive quality is related to the importance of the specific content of scientific ideas. Therefore, this type of quality is assessed only on the basis of pure scientific considerations. Methodological quality is related to the accuracy of methods and techniques and is assessed with the help of rules and criteria current in a particular scientific field. Esthetic quality deals with the degree of attractiveness of mathematical formulations, models, etc. The assessment of this type of quality is a highly subjective affair, it is usually based on the relationship between the simplicity of a formulation and its explanatory value.

Of course, there exists a considerable overlap with regard to the values mentioned and, in addition, it might be possible to think of other related pure scientific values. We consider the above aspects of "quality" to be related to "quality in a more restricted sense" or "basic quality". A judgement of this quality is based on criteria intrinsic to scientific research. Therefore, only colleague researchers ("peers") can decide about this basic quality of research projects. However, for research to have impact, first of all it is necessary that colleague researchers do have the opportunity to form an opinion on the basic quality of that research. Then it becomes possible that this research, when it indeed has a certain basic quality, does make an impact. This means that basic quality is a necessary, but not a sufficient condition to make an impact. According to this view, one aspect of successful research performance is that researchers are active in presenting their research findings to colleague researchers. In fact, we consider this activity as an aspect of scientific quality in a more broader sense. Scientific quality thus defined includes basic quality as well as the extent

to which researchers successfully perform "public relations" activities. We think that our impact indicators are indicators of scientific quality in this sense. We hope that this notion of scientific quality clarifies discussions about the assessment of research performance.

3. Description of the bibliometric indicators used in this study

3.1. Trend analysis

For each research group we calculated, for each publication year during the decade 1970–80, bibli-

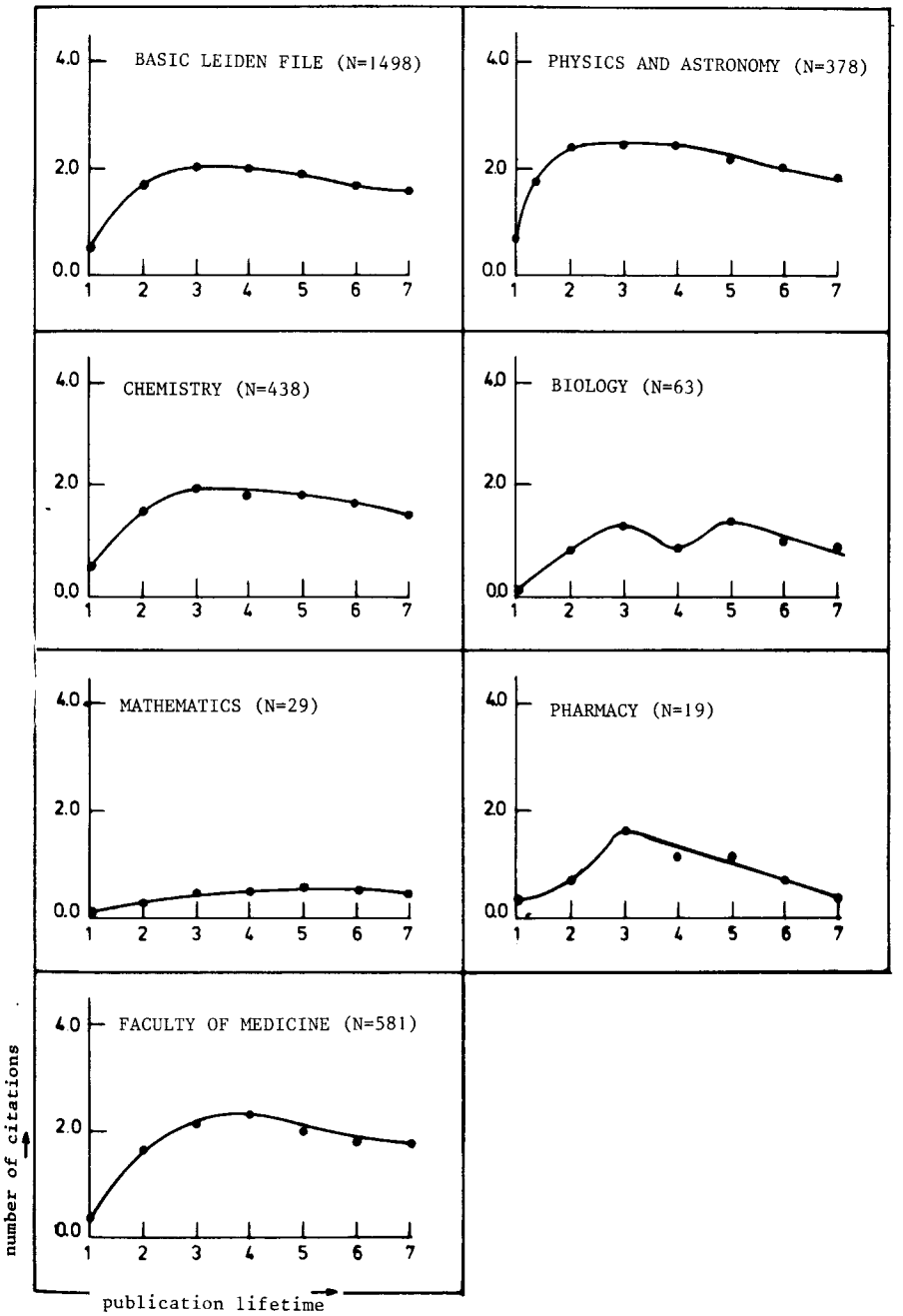

Fig. 1. Curves representing the average number of citations received by publications as a function of their age. Publications published in the period 1970–74a are aggregated (excluding meeting abstracts) for the (sub)faculty indicated. The number of publications involved is given in parentheses.

ometric indicators, primarily based on annual numbers of publications and annual numbers of short-term citations received by these publications, and analysed *trends* for these indicators.

We first present our basic assumptions in this analysis:

(1) If the number of annual publications of a research group (output indicator) increases (decreases), we regard this as an indication that the output (scientific production) of the group increases (decreases);

(2) If the number of (short-term) citations increases, we regard this as an indication that the impact of the work of a research group is increasing. If the values of these indicators decrease during the decade, we view this as an indication that the impact of the group is declining, and that the group is getting out of touch with the research front;

(3) If the number of (short-term) citations is increasing, but the number of citations-per-publication is decreasing simultaneously (which means that the number of short-term citations increases less sharply than the number of publications of a research group), we take it that this phenomenon indicates that the research group concerned is reaching some saturation level: such a group continues to publish more and more articles, but the impact of its performance as a whole does not increase proportionally.

We consider a three-year citation-counting period as a standard period. On the one hand, we preferred to keep the counting period as short as possible, since we wish to evaluate (short-term) impact of the most recent publications. On the other hand, the period should not be too short, since it takes some years for a publication to have any impact at all.

An analysis of all publications in the Leiden Basic File brought to light that an average publication receives its maximum number of citations in the third year of its life-time. For this reason, we took the first three years of a publication life-time as a standard period from which citations are counted. Although an average publication in the Basic Leiden File receives its maximum number of citations in the third year, we found that this is not always true on a lower aggregation level: fig. 1 nicely illustrates this finding, see for example the results for Mathematics, Biology and Medicine.

We also carried out citation counting during the first two years of a publication's life-time. We counted this in order to check the results from the standard case: if a research group has a real increase in (short term) impact, one would expect a positive trend in the number of citations counted in both ways. If the number of citations received by publications during the first two years of their life-time increases, while the number of citations counted during the first three years does not (or vice versa), this could be an indication that some disturbing factors may be involved which should be analysed in more detail. This phenomenon – that is, the fact that the results of both counting modes do not converge – was observed several times.

3.2. *Trend figures*

The construction of the bibliometric indicators for the trend analysis is illustrated in fig. 2. Each such figure comprises the results of one specific research group.

The left-hand panel of fig. 2 gives: the number of publications for each year (research-output indicator) (solid line); the number of citations received by these publications during the first *three* years of their life-time [4], excluding in-house [5] cita-

[4] The year of publication is the first year of a publication life-time. For example, 1982-citations of a 1980-publication are citations in the third year.

[5] The concept of "self citations" or "in-house citations" can be defined in various ways. The aim of these concepts is to identify citations by authors who work within the domain of direct influence of publishing authors. Clearly each publication has its specific domain. How should such a domain be defined?

In this project we used a rather broad conception of "in-house citations": the domain ("house") consists of all authors (both first and co-authors) of all publications in the Basic Leiden File. Generally speaking, these authors are researchers appointed for at least some years at Leiden University. As well, there are researchers who are not affiliated to Leiden University, who nevertheless have published together with Leiden researchers. It follows that if such in-house citations are excluded, in fact only citations are counted from source articles of which the first author (we remained that from *citing* articles only the name of the *first* author is present in our Citation File) has never published a paper with a Leiden address during the period 1970–80. In this way we are able to evaluate the impact Leiden publications have outside Leiden. The overall percentage of in-house citations amounts to approximately 31 percent.

tions (publication-impact indicator) (dashed and dotted line); and the ratio of the above two indicators, or: the number of citations per publication (citations-per-publication ratio) (dashed line).

The middle panel of fig. 2 gives the same indicators as in the left panel, but rather than calculating the values per annum, the year-average values for *four-year* publication blocks are calculated.

The right panel of fig. 2 also gives the same type of indicators, but this time for a *three*-year publication block and a *two*-year citation counting scheme.

3.3. Level analysis

A trend analysis alone is not sufficient to obtain a complete picture of the impact or output of a group, since in principle one cannot on the basis of trends alone distinguish between an increase from "very low" to "low" and an increase from "average" to "high". Therefore, trend analyses are supplemented with analyses that provide indicators of the output or impact *level* achieved by a research group. Probably the best way to obtain such indicators is to rank Leiden groups with other international groups working in the same field on the basis of their bibliometric scores. We did not calculate indicators with respect to other groups outside Leiden for two reasons. The first reason is that it would have been an enormous effort to identify for each Leiden group other groups working in the same field and to gather and handle both bibliometric and non-bibliometric (e.g. institutional) data for these non-Leiden groups.

Second, even if we could have succeeded in calculating bibliometric indicators and in obtaining rankings of groups for each field, it would be quite difficult to interpret these rankings without background knowledge on the non-Leiden groups. Members of these non-Leiden groups should be able to participate in some way in our project. Again, it would bean enormous effort to organize such participations of so many non-Leiden groups. Therefore, we tackled the level problem in another way. First, citation and publication counts of research groups within one (sub)faculty were compared. Second, citation counts to publications of a group are compared with the "expected" number of citations, i.e. average citation counts to all publications in the journals in which the group itself has published. We assume that this comparison provides a first indication of the international impact level of the Leiden group.

In order to construct an indicator for the "expected" impact, we used publication and citation

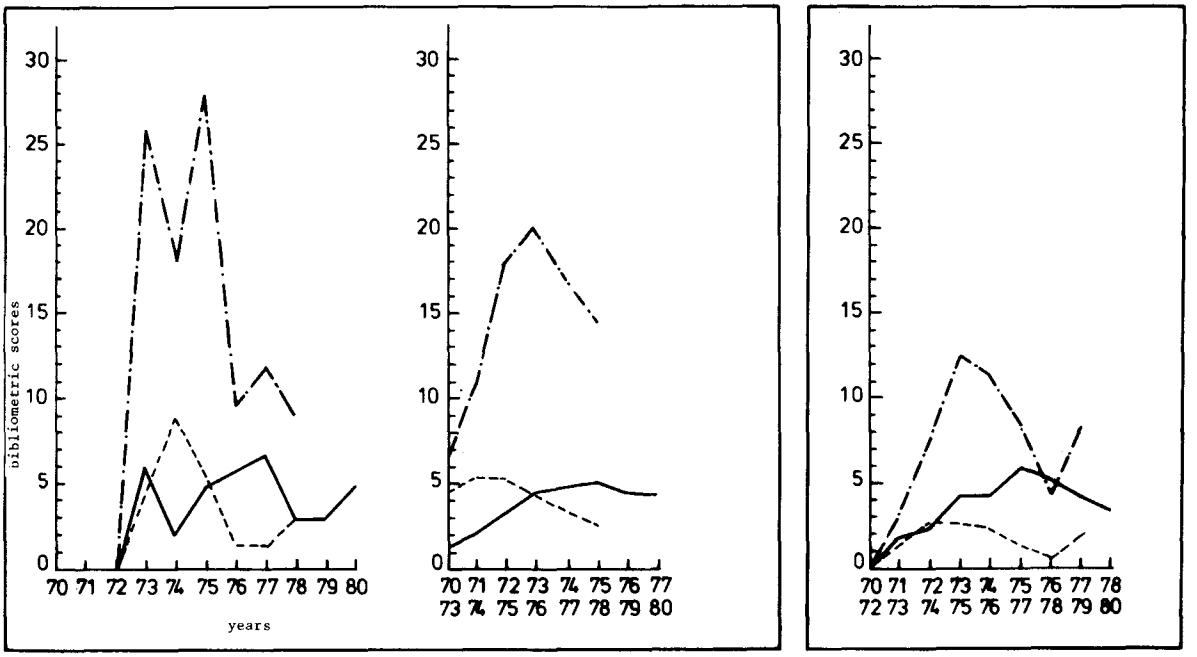

Fig. 2. Example of the trend analysis for an arbitrary university research group (for detailed explanation see section 3.2).

data per journal from the Journal Citation Reports. For instance, in the 1979 Journal Citation Reports we find for each source journal:

(a) The number of articles (excluding meeting abstracts), published in the journal during the year 1977;

(b) the number of tims these articles have been cited in the year 1979 by other source-journal articles. The counted citations are article–article links, so if one article refers to another article several times, this is counted as one.

The ratio (b)/(a) represents the number of times an article from a journal is cited by other journal articles in the third year of its life-time. For instance, in the third year of their life-time the 1977 publications in *Physica A* receive on average 1.5 citations from other journal articles. For *Phys. Rev. B* (*Physical Review B*) this average number of citations per publication amounts to 3.4. We call this ratio the Journal Citation Score (JCS) of a journal.

A research group however usually publishes articles in a number of different journals with different JCS values. For each research group, we calculate a weighted-average JCS value (written as $\overline{\text{JCS}}$) of the journal "packet" in which the group has published. The weighting factors are equal to the number of publications in each different journal. As an example we take a research group which published 25 articles in 1977: 11 articles in *Physica A* (JCS = 1.5) and 14 articles in *Phys. Rev. B* (JCS = 3.4). We then calculate the weighted-average $\overline{\text{JCS}}$:

$$\overline{\text{JCS}} = \frac{(11 \times 1.5) + (14 \times 3.4)}{11 + 14} = 2.6$$

This $\overline{\text{JCS}}$ value represents the average number of citations received by these 25 publications published in the two journals in the third year of their existence. We use this $\overline{\text{JCS}}$ indicator as an "expected impact" indicator based on the journal packet in which a research group published. For each results group or department, we calculated $\overline{\text{JCS}}$ values with respect to two periods: 1971–74 and 1975–78.

We use $\overline{\text{JCS}}$ values to tackle the level problem as discussed at the beginning of this section, i.e. to obtain a reference for the impact *level* of a research group (since it represents an a priori expected number of citations per publication for any research group publishing in those journals forming the particular journal packet of that group).

It follows that by comparing the actual number of citations per publication of a research group with the expected number based on the $\overline{\text{JCS}}$ values, one can obtain an indication of whether this actual number of citations per publications is low or high. Moreover, in this type of level analysis, differences in citation practices for the various fields are taken into account (see section 4.4) since citation practices and average citation levels within a field will be reflected, at least partially, in the citation scores of journals that cover the field.

There is, however, a striking problem. Within a specific research field we can find journals with low, and journals with high impact. Suppose that there are two groups working in the same field. Both groups have a citation-per-publication ratio more or less equal to the $\overline{\text{JCS}}$ of the journal packet in which they publish. The first research group, however, publishes in high-impact journals, and the second in low-impact journals. In the comparison presented in this section, i.e. for each group a comparison with their "own" journal packet, both groups will present "equal results", although the first group has a higher citation-per-publication ratio than the second. On the basis of this, we conclude that the results of this type of level analysis ("expected impact") alone, are not sufficient to obtain a "complete" picture of the impact of groups. They should be complemented by an indication of the impact level of the journals as such. Such an indication can be obtained by comparing $\overline{\text{JCS}}$ values with the JCS value of specific journals that are generally recognized as either prestigious or obscure in the research field concerned.

3.4. Level figures

The construction of the bibliometric indicators for the level analysis is illustrated in fig. 3.

The left-hand panel of fig. 3 shows the comparison of "actual" with "expected impact. Only citations received by publications in the *third* year of their life-time are counted, *in*cluding in-house citations. The indicator for "expected" impact is the $\overline{\text{JCS}}$-value for the journal packet of the research group concerned, as discussed in section 3.3. Books are excluded as sources of publications

and citations. The solid line indicates the "actual" number of citations per publication; the dashed line indicates the "expected" number of citations per publication ($\overline{JCS}$).

The middle panel of fig. 3 shows a level analysis giving research *output* normalized on input of the research group concerned. We present a histogram which gives the distribution of the publications-per-researcher number for each research group of the subfaculty concerned. This indicator is calculated with respect to the most recent four-year period: 1977–80. The dark rectangle in the figure marks the position of the particular research group (the ordinate gives the number of research groups in a particular (sub)faculty with a publication-per-researcher score as given on the abscissa).

The right-hand panel of fig. 3 shows a level analysis giving publication *impact* normalized on input of the research group concerned. We present a histogram which gives the distribution of citations-per-researcher numbers for each research group within the (sub)faculty concerned. Citations received by publications (excluding in-house citations) during the first three years of their life-time are counted. This indicator is calculated with respect to the four-year period 1975–78. The dark rectangle in the figure marks the position of the particular research group (the ordinate gives the number of research groups in a particular (sub)-faculty with a citations-per-researcher score as given on the abscissa).

In this way we obtained a large number of intriguing trend and level analysis results of research groups in both faculties (see also Moed et al. [3, pp. 37–54]). The various different types of outcome are illustrated in section 5 by means of a selection of the research groups analysed, along with our interpretations and comments of the interviewed researchers.

4. Specific problems of data collection, data handling and interpretation

4.1. Completeness of bibliometric data

Obtaining complete publication and citation data was by no means an easy task. Several omissions, some due to programmatic or operational errors at ISI, were detected by us and completed at our request by ISI. The number of missing data amounted in the first instance to about 10 percent of the total number of data involved. However, the nature of some of these errors was such that for several individual departments or research groups the bibliometric data were highly incomplete. The following may serve as an example. Due to some programmatic error, all citations to all publications of a number of journals were missing. These missing citations were about 3 percent of the total number of citations. However, for a number of specific research groups, namely those who published in these journals, 50 percent of the citations were missing by this error.

A complicating factor is that from an analysis of the Leiden bibliometric data it appears that most of the publications of a research group receive only a few citations, while a few particular publications can be highly cited. If one misses such a highly-cited article, citation counts can be most incomplete. For example the Netherlands National Survey Committee on Biochemistry [5]

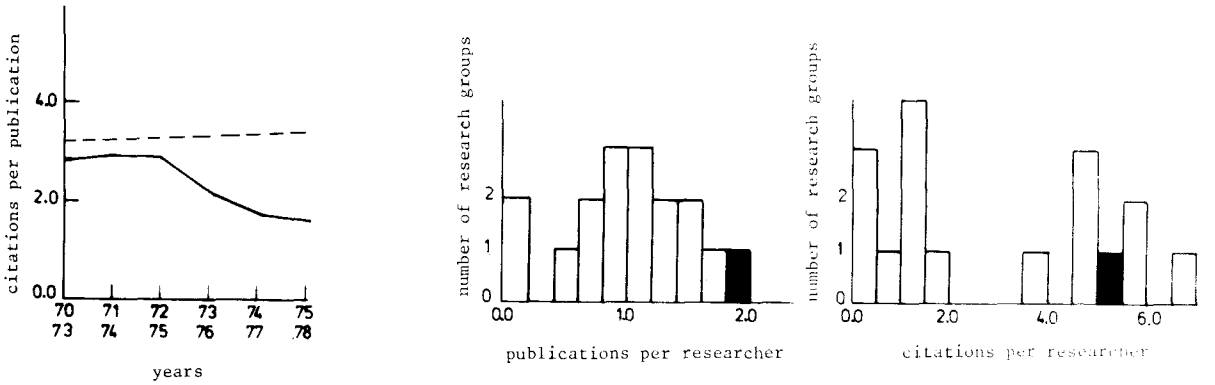

Fig. 3. Example of the level analysis for an arbitrary university research group (for detailed explanation see section 3.4).

evaluated, as part of an extensive investigation on research performance of Dutch biochemistry, the impact of one of the Leiden biochemistry research groups, and appeared to have missed *one* publication that was cited as many times as the total number of citations that the Committee had collected for this group.

We would therefore like to emphasize the following problem. In regard to the relevance of impact and output analyses for university research policy, the research group seems to be the most adequate research unit. However, these units are small (they contain two to ten researchers), and produce relatively few publications and citations. Consequently, bibliometric indicators are based on low numbers. Small errors or a few omissions can lead to dramatic differences in results and interpretations. It follows that one should make all possible effort to obtain sufficiently complete bibliometric data. If a higher level of aggregation is chosen, for instance large departments, clusters of departments, or (sub)faculties as a whole, thus operating on larger numbers, say a hundred publications and a few hundred citations, completeness becomes less important. However, bibliometric analyses on such high levels of aggregation are less relevant to research policy at a university level.

Finally we achieved in cooperation with ISI a completeness percentage of 99 with respect to publications with a Leiden address from the SCI Source Index. The search and selection techniques applied in the project hardly allow for a higher completeness percentage. We propose this 99 percent as a standard rule with respect to the completeness of the publication data in impact evaluation studies of (small) university research groups.

A general problem with the selection of citations to given publications is that researchers in their list of references can be given bibliographic descriptions that do not exactly correspond to the bibliographic data of the article they intend to cite. For instance a slight variation in the first author name occurs, or an erroneous volume or page number is given. Citation analysis results of the National Survey Committee on Biochemistry [5] however indicate that variations or errors in (first) author names occur much more frequently than errors in volume and page number, or year.

The citations in our data base were selected by matching journal title, year, volume, page number and the first three characters of the first author name of Leiden publications with the same data of the cited references (i.e. articles cited in any SCI source publication) in the Citation Index. The match algorithm did *not* compare the full name of a first Leiden author with a first author name in the cited references. As a consequence, omissions due to variations in first author name probably did not occur. However, in cases of errors in journal title, year, volume and page numbers, citations were not selected. In order to obtain some estimate of the percentage of citations lost due to these errors, citations to a relatively small sample of 35 publications were checked by carefully inspecting the printed volumes of the Citation Index. The difference between the number of citations in our file and the number found in the printed volumes was approximately 5 percent.

It appears to us that a completeness percentage of 95 with respect to citation data is sufficient for an evaluation of small research groups, provided that no "systematic" errors are made of the type illustrated in the example given at the beginning of this section. Therefore, citations should be selected by matching only journal title, volume and page number, and year. In addition, a number of manipulations on citation data should be carried out in order to examine whether systematic errors were made when these data were collected. Since the necessary matching and manipulations can actually only be carried out with the help of computer algorithms, all data should be on computer tapes.

4.2. *University data*

It should be emphasized that extensive university data are indispensable for bibliometric analysis on research group or department level. In order to properly assign publications (and citations) to the research groups from which they originated, one needs complete data on research groups, i.e. either the names of all researchers who belong or have belonged to the group, or – even better – complete lists of publications per research group. Therefore we conclude that a very crucial part of this type of research performance analysis lies in a sophisticated combination of these two data clusters (bibliometric data and university data). For a detailed discussion we refer to Moed et al. [3].

4.3. Problems in using the SCI data base as a tool for past performance analysis

In order to obtain an insight to the appropriate of SCI source journals and SCI source books as an adequate tool for bibliometric output and impact analysis, we carefully studied the publications of each (sub)faculty listed in annual University research surveys, and determined the percentage of publications in SCI source journals or books (benchmark year 1979). The following results were obtained.

For the subfaculties Chemistry and Physics & Astronomy, 80 percent or more of all listed publications are published in SCI source journals and books. The remaining publications are mainly contributions to meetings and symposia. For the subfaculties Pharmacy and Biology, and for a sample of four departments in the Faculty of Medicine, the percentage of publications in SCI source journals and books is considerably lower: 33, 52 and some 45 percent respectively. The main reason that these percentages are much lower than those for the other two subfaculties, is that some 25 percent of all listed publications of the subfaculty Biology are written in English, though published in journals not processed for the SCI. These are mainly publications of the Department of Taxonomy.

For the subfaculty Mathematics the percentage of publications in SCI source journals or books is the lowest of all subfaculties: 24 percent. The main reason for this is that almost 40 percent of all listed publications are research reports of the subfaculty or of the Mathematical Centre, a national research institute in Amsterdam. However, these are all published in English.

Summarizing, we find that with respect to research groups in the subfaculties Chemistry and Physics & Astronomy, calculation of output and impact indicators based on SCI data present no serious problems. Nearly all journals in which they publish are covered.

With respect to research groups in the subfaculties Pharmacy and Biology, and the Faculty of Medicine the output in Dutch is not covered by the SCI source journals or books. In our opinion this is not a serious problem. Since these articles are written in Dutch, they should not be considered as directed towards the international scientific community. From the point of view of university research policy, output and impact on the international research front are aspects in their own right, and this study deals with these aspects. It should be kept in mind that impact on the national level cannot be adequately indicated on the basis of SCI data. The fact that 26 percent of the listed Biology publications are written in English though published in journals not processed for the SCI, is of more concern. Taking the language in which they are written as a criterion, these publications are directed towards the international research community, yet the SCI Source Index does not cover them. A field like Taxonomy is probably not adequately covered by the SCI. Consequently, it is doubtful whether one can obtain reliable impact and output results based on SCI data alone.

Finally, the fact that almost 40 percent of all listed publications of the subfaculty Mathematics are published in research reports, not included in the SCI source journals or books, constitutes a problem as well. Again, taking the language in which they are written as a criterion, they are directed towards the international scientific community, however the reports are not published in journals. Possibly the role of the serial (journal) literature for the communication of research findings differs from discipline to discipline. Further research into this problem is needed.

4.4. Some important disturbing factors

Dealing with bibliometric impact indicators, i.e. variables determined (partly) by the impact level of a group, we should keep in mind that these indicators are also influenced by disturbing factors that have little to do with impact as such (see also Martin and Irvine [2]). In this study, we analysed disturbing factors related to specific citation practices within fields, and factors related to (changes in) the coverage of the SCI data base.

The main results of these analyses:

(1) Citation practices appear to differ significantly from *field to field*. In some fields researchers tend to cite recent (for instance two years old) articles more frequently than in others. For example, an article in the *Journal of Differential Equations* (Mathematics) contains on an average one citation to two-year old journal articles. Yet, an article in the *European Journal of Biochem-*

istry contains an average of four citations to two-year old journal articles. This result suggests that the probability of a two-year old article being cited differs significantly within the two fields mentioned in this example. One should expect – and one actually observes – much higher short-term citation levels in Biochemistry than in Mathematics.

We found that differences in citation counts for groups in different fields cannot be interpreted merely in terms of impact. In other words, one cannot establish relative impact by directly comparing citation counts for groups in one (sub)faculty, or even in one department. Disturbing factors due to differences in citation practices can to a larger extent affect the numerical values of the citation-based indicators.

(2) Citation practices within fields can also change *during the decade*. In fact an analysis of the SCI data base between 1970 and 1980 brings to light that a journal article contains, on average, an increasing number of citations to 0–2-year-old articles during the decade. Hence, the increase of the average number of citations that 0–2-year-old Leiden articles receive during the decade should be directly proportional to this. We assume that this increase does not reflect an increase of the impact of the Leiden publications.

(3) Considerable disturbances result from the fact that ISI has *included more* source journals during the decade and that, as from 1977, non-journal material (books) are also included. The following may serve as an example. Due to the inclusion of books in 1977, the number of Leiden publications increased by about 10 percent, while some groups profit more than others. Moreover, the number of citations received by Leiden articles is affected as well. For some groups, the number of citations doubled, while for others it did not change at all.

We conclude that because of possible changes in citation practices or changes in the SCI source journal and books, one should be very cautious about interpreting trends in the numerical values of bibliometric indicators as trends in output and impact. It requires considerable effort and knowledge to identify the effects of these disturbing factors, especially at the level of research groups. It must be noted though that we constructed several bibliometric indicators that neutralize some of the disturbing factors mentioned above. In particular the ratio of "actual" and "expected" impact (see section 3.3). We recommend further research on the distribution of the SCI source books over the various research fields and on differences with respect to communication practices between fields.

4.5. The problem of "statistics"

A fundamental problem in the bibliometric analysis presented here is: which differences between bibliometric scores (both in trend and level analysis) can be considered as "significant" with a certain probability, and which differences should be ascribed to mere chance? To our opinion, this problem can be attacked by further empirical investigation and by developing a theory on scientific performance (impact) of research groups, in which the concept of "impact of the *oeuvre* of a research group" (i.e. an ensemble of related publications during a period of, say, four years) is related to a probability distribution of citations amongst articles constituting this *oeuvre*.

We consider the work of Chang [5] as a first step in this direction. Chang considers citing as a stochastic process. Articles are assumed to attract citations due to their impact, but also due to many accidental effects. According to Chang the stochastic character of citing can be described adequately by a Poisson probability distribution. He has developed a mathematical model that enables him to determine which differences between citation scores of articles should be ascribed to mere chance – with a certain probability – and which differences are due to differences between the impact of the articles. However, Chang considers the impact of individual publications only, while we are interested in the impact of the complete *oeuvre* of a research group. In this study we have not yet worked through the above ideas.

In any case, one should be careful in interpreting small effects in trends or differences between bibliometric scores as significant changes or differences with respect to impact. At this moment we are completing the updating (as discussed in section 1) of our bibliometric data. The strongly increased number of data per research group puts us in a much better position to tackle the problem of statistics along the lines indicated above.

5. Results

5.1. Presentation of bibliometric results for some specific research groups: Comments and interpretations

In this section we present, as an example, the results of our bibliometric past performance analysis for six selected research groups. Since the results and interpretations of many of the 120 Leiden research groups analysed in this study were rather similar, and could be classified in a limited number of "typical cases", presentation of those research groups characterizing these typical cases gives a clear picture of the possible outcomes of our methodology.

We discuss the bibliometric results in the following scheme. First, (A), the presentation of the trend analyses as discussed in sections 3.1. and 3.2. Second, (B), the presentation of the level analysis (sections 3.3. and 3.4). Third, (C), we give (briefly) our interpretation of the bibliometric results and (D) the comments of the interviewed peers. Finally (E) we comment on the statements made by the interviewed peers. In these "concluding notes" we primarily focus on questions such as to what ex-

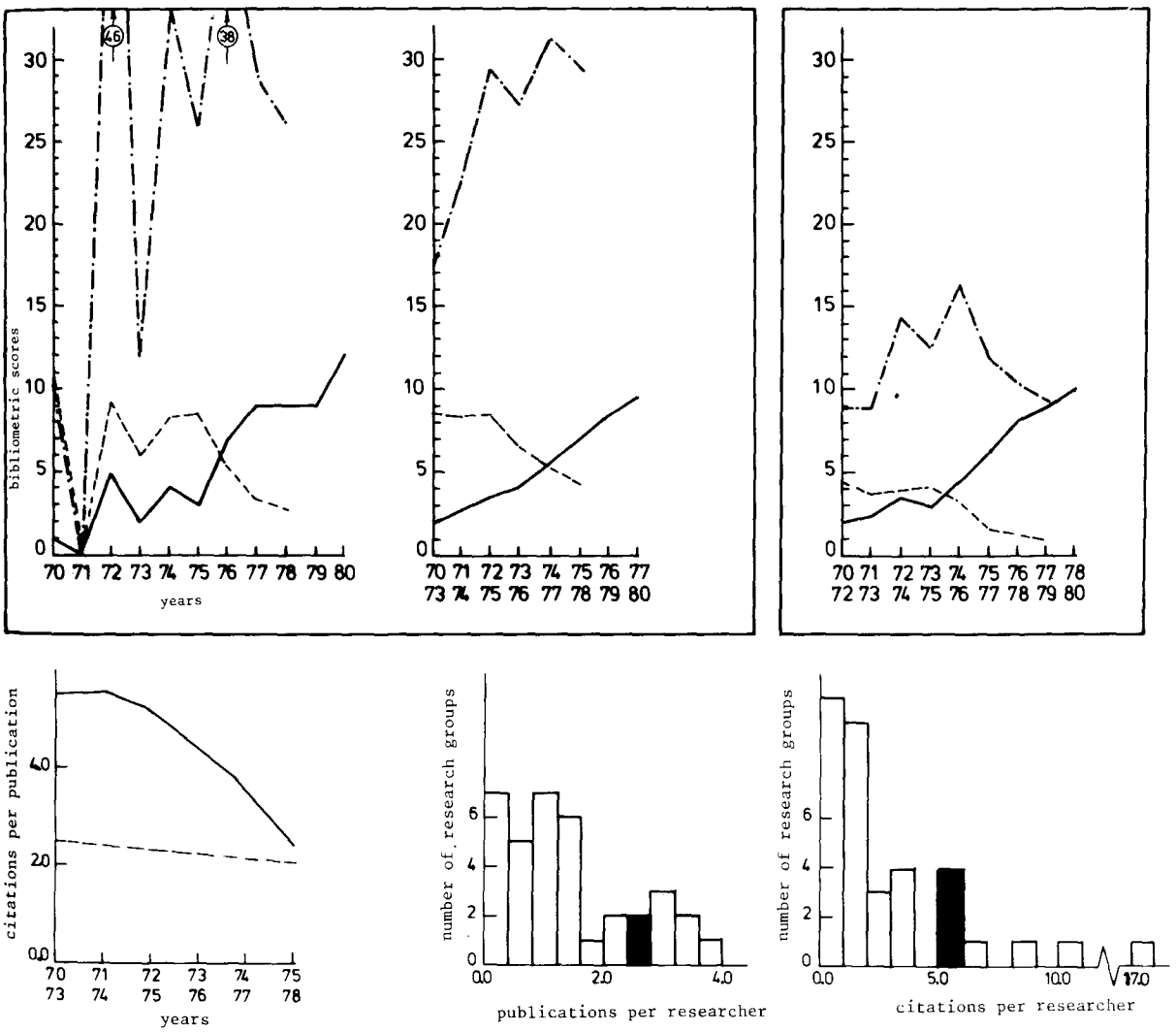

Fig. 4. Bibliometric analysis of a non-clinical research group in the Faculty of Medicine. *Upper*, trend analysis (solid line: publications; dots and dashes: citations; dashed line: citations per publication; for detailed explanation see section 3.2). *Lower*, leve analysis (left panel solid line: "actual" impact, dashed line: "expected" impact; for further detailed explanation see section 3.4).

tent do our findings agree, or disagree, with the perception of the interviewed peers; what are the pitfalls, limitations and interpretational problems connected with the constructed indicators; and do our bibliometric results – apart from these problems – provide a meaningful basis on which to evaluate past performance of research groups.

I. A non-clinical research group in the Faculty of Medicine

A. Trend analysis

The results of the trend analysis for this research group are presented in fig. 4a.

B. Level analysis

The various types of level analysis for this research group are shown in fig. 4b.

C. Our interpretation of the bibliometric results

The impact of this research group decreases from a high to an average level.

D. Comments of the interviewed researchers

All the researchers were surprised by our findings. However, they did not simply disagree with our interpretation but tried to find explanations. One of them put forward the following explanation: during the decade concerned, there were relational problems between the two leading senior

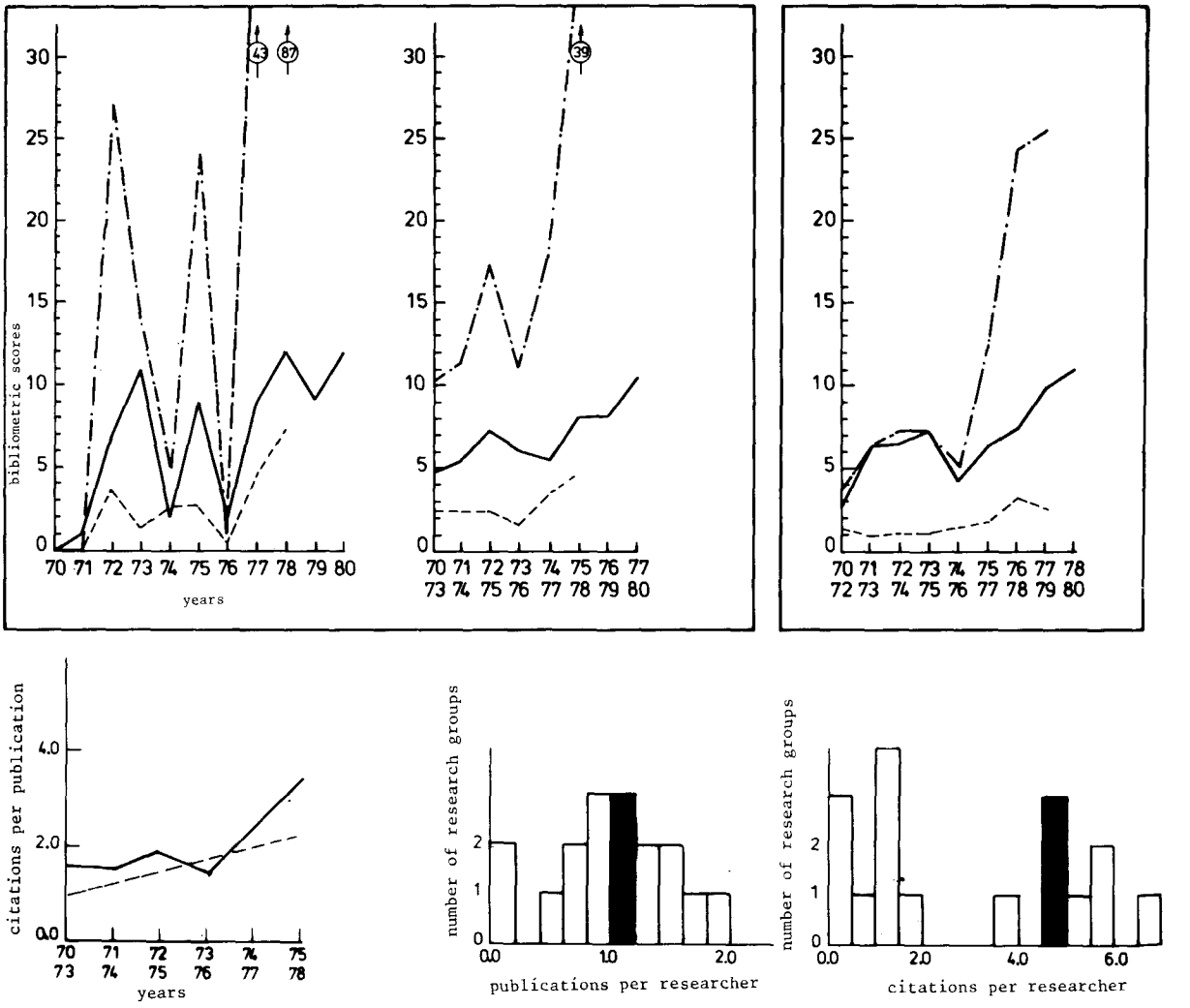

Fig. 5. Bibliometric analysis of a resarch group in the subfaculty Physics & Astronomy. *Upper*, trend analysis; *Lower*, level analysis. For explanation of symbols see legend to fig. 4.

staff members. This conflict resulted in the departure of one of them. According to him, these problems had a major impact on the research performance of this department. However, another researcher came up with a different explanation: he thought that the group performs important research, but the leading professor is up against strong opposition from colleague researchers. Notably, members of editorial boards try to ignore the work and prevent it from being published.

E. Concluding notes

In analysis the bibliometric results of this department, some relevant information came to light which would otherwise have remainded obscure.

II. A research group in the subfaculty Physics and Astronomy (Faculty of Mathematics and Natural Sciences)

A. Trend analysis

The results of the trend analysis are presented in fig. 5a.

B. Level analysis

The results of the various types of level analysis are presented in fig. 5b.

C. Our interpretation of the bibliometric results

Both output and impact of this research group increased. The size of the group remained constant during the decade. There is a significant change of journals. During the period 1971–74, the group mainly publishes in *Physica* and *Physics Letters A*, journals with a rather low impact. During 1975–78, it started to publish frequently in *Journal of Physics F, Solid State Communications* and *Physical Review B*; these journals have a high impact. The citation-per-publication ratio is much higher than the expected value based on the average Journal Citation Score – which is already rather high. The impact of this research group obviously has increased to a very high level. According to our interpretation, the group is internationally recognized as a top group.

D. Comments of the interviewed researchers

They agree fully with our interpretation. At the beginning of the decade a new professor was appointed. He proved to be a source of inspiration. The group consists of active researchers, who have many international contacts and they are recognized by fellow researchers in the field.

E. Concluding notes

None.

III. A research group in the Department of Biochemistry (Faculty of Mathematics and Natural Sciences, Subfaculty Chemistry)

A. Trend analysis

The results of the trend analysis are presented in fig. 6a.

B. Level analysis

The results for the various types of level analysis are presented in fig. 6b.

C. Our interpretation of the bibliometric results

The output of this research group increased during the decade and reached a level which is a little higher than the average. The number of citations received by publications during the first three years of their life-time, reached a maximum in 1975 and decreased sharply after that year. The impact of this group passed a peak and decreased in the second half of the decade to a rather low level.

D. Comments of the interviewed researchers

The researchers neither agreed nor disagreed with our interpretation. One of them immediately stated that he was not able to assess the past performance of this research group. The other researcher felt that our interpretation could be correct, but he mentioned two other possible interpretations:

- the quality of the work has not been constant during the decade. One should expect fluctuating figures, but it is incorrect to say that the impact of the group passed a peak.
- In the second part of the decade the group focused on research subjects which are less "popular" in the field. This may lead to a declining short-term impact, but again it is incorrect to conclude that the impact of the group passed a peak.

E. Concluding notes

Without detailed knowledge on the *causes* of the observed impact decline, it is impossible to

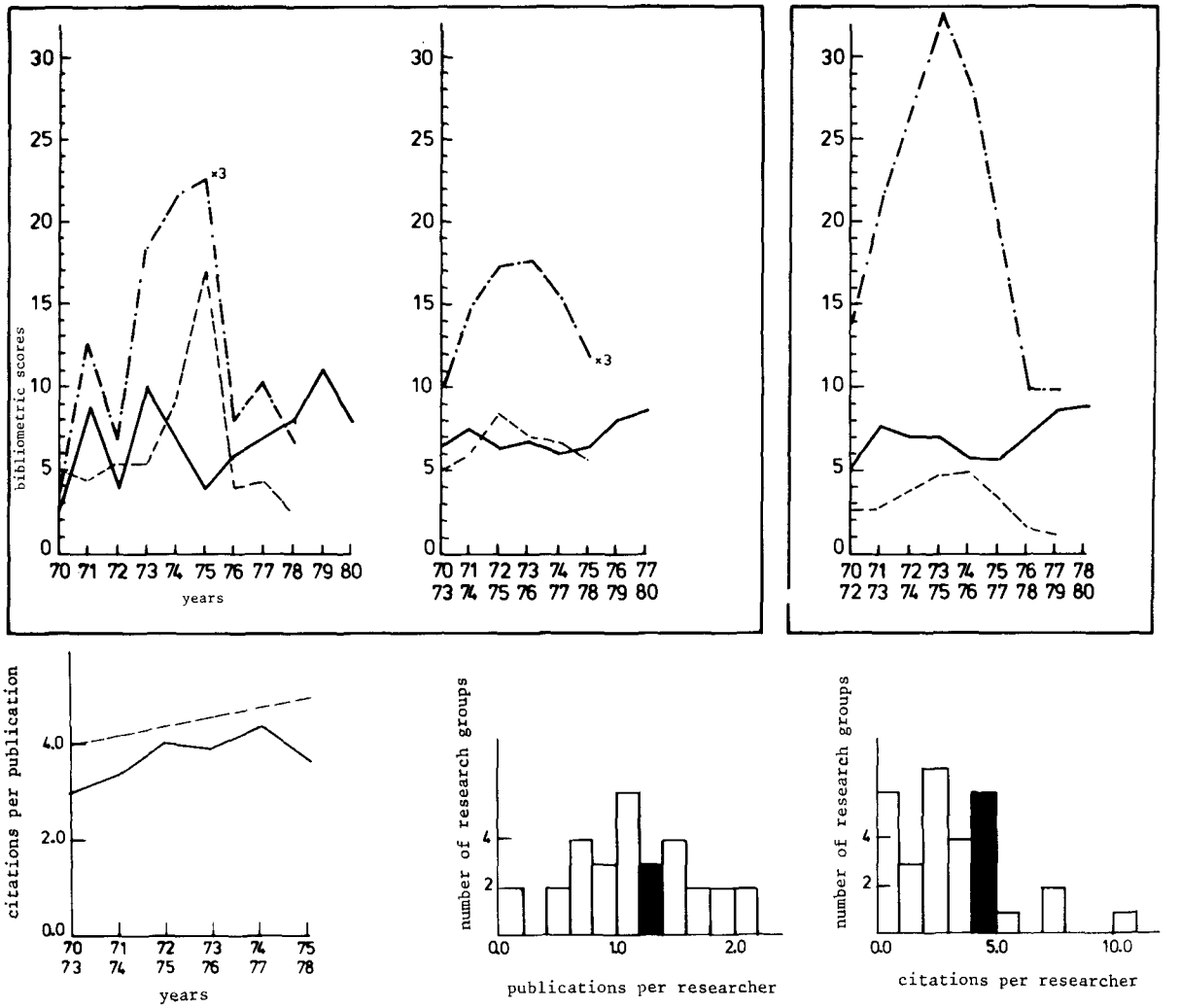

Fig. 6. Bibliometric analysis of a research group in the subfaculty Chemistry. *Upper*, trend analysis; *Lower*, level analysis. For explanation of the symbols see legend to fig. 4.

predict whether this declining trend will continue. We feel that the bibliometric results for this research group form a valid reason to pay close attention to this group's future research performance.

5.2. *General comments of interviewed researchers*

Twelve researchers of recognized international scientific status in the (sub)faculties involved were interviewed. These interviews were meant as an "acceptance test", i.e. to examine whether our bibliometric results and interpretations differed or agreed with the ideas of scientists in the field; whether we had overlooked hidden pitfalls and other problems with respect to the bibliometric indicators, and whether the bibliometric results and interpretations provide a meaningful basis for the discussion of past research performance of research groups. During the interviews, the researchers saw our results for the first time. They expressed their personal view. So their comments should not necessarily be considered as representative of the view of any (sub)faculty body or for the "scientific community" within the various (sub)faculties. They made comments that were related to all the issues described above; comments dealt with matters such as: what the constructed indicators actually indicate; the type of indicators chosen; the completeness of the bibliometric data; the

coverage of the SCI data base and to disturbing factors that should be kept in mind in interpreting bibliometric results.

To our opinion the most interesting general result of these interviews is the following. In many cases the interviewed researchers did not reject our bibliometric results and interpretations, but they tried to find explanations for the observed output and impact of the research group. For example, they related a given low-impact level of a particular research group to the low quality of the research performed by this group. Or they related the low impact to the fact that the researchers within the group are not fond of the limelight, or do not present their findings in an effective way. They mentioned causes that refer to particular events or circumstances within the groups, such as relational problems between staff members, changes in the permanent staff, or the fact that the group does not have a "second best man". In some cases, they thought the observed low impact of groups in experimental fields was related to the fact that they had fallen behind on a technological level. In addition, causes were mentioned that refer to characteristics of research fields, such as strong competition and complete lack of consensus, resulting in opposition from members of editorial boards of significant journals. So during the interviews, many factors relevant to policy were brought to light that otherwise would have remained undiscussed. The figures formed a sound basis for these discussions as they gave a concise indication of research performance over a relatively long period up to recent years.

We conclude that the type of bibliometric analyses carried out in this study can be a useful tool for a university research policy. It constitutes a "monitoring device" with which research groups can be followed over a long time up to quite recent years. In principle, bibliometric analyses provide a meaningful basis for discussion of the research performance of research groups with scientists in the field, and possibly also with members of the monitored research groups themselves. The analyses enable research policy-makers to ask relevant questions about research groups; to some extent the analyses provide an historical picture to those who are not familiar with the research performed by a group or with the field in which it works. According to us, bibliometric analyses can in principle provide a basis for a dialogue between research policy-makers and the researchers of various university groups.

It should be emphasized here that we feel there is no straightforward or simple relationship between the results of bibliometric analyses and the nature of future policy decisions. In order to make proper policy decisions, insight should be obtained about the factors or causes underlying the observed impact and output, and about the extent to which these factors can be expected to persist. Bibliometric analyses are not a substitute for background knowledge about these factors. They can in fact be used to bring these factors into the open through discussions with the researchers in the groups being analysed.

6. Conclusions

The central issue of this paper is the examination of the potentialities of bibliometric indicators as tools for university research policy. We focused on indicators based on the number of times publications are cited in the international scientific literature.

We argued that citation counts indicate "impact" rather than quality. Impact is defined [2,3] as actual influence on surrounding research activities. It is assumed that publications must have a certain basic quality in order to generate impact. However, other factors can determine impact as well, factors like the state of the art of the scientific field concerned, the visibility of journals or the extent to which researchers carry out public relations tasks. Impact is a relevant aspect of research performance since from a research policy viewpoint, one should not only require that researchers produce results of some scientific quality, but also that they make their results known to colleagues.

Moreover, we made a distinction between short and long-term impact. Short-term impact refers to the impact of researchers at the research front up to a few years after publication of their research results. Looking at impact over a long period offers the possibility of relating impact to "durability". This long-term influence of research can only be determined after a (very) long time. However, this period is often too long for university science policy, which is concerned with evaluation of recent research. Therefore, in this study re-

search performance analysis is confined to short-term impact. This impact is measured by counting citations received by publications during the first years of their life-time. We are now investigating the relation between short-term and long-term impact.

We chose the research group as the level of aggregation since a research group usually constitutes the "natural" unit of research activity (at least in the two Leiden faculties involved in this study).

This study offers an instrument for bibliometric research group evaluation: a *trend* analysis as a past performance evaluation over a period of one decade, and a *level* analysis to determine the relative score on a (sub)faculty and on an international scale. Results of this method have been discussed with researchers in the various fields under study. On the basis of these interviews we can conclude that when used properly, this instrument can be a "monitoring device" for research management and science policy. It enables research policy-makers to ask relevant questions in order to find an explanation of the bibliometric results in terms of policy relevant factors. This offers them the possibility of obtaining relevant information needed to make justified policy decisions.

Aside from the above general conclusions, we found that the following specific points are essential in working on bibliometric analyses of university research groups:

(1) The *completeness* of bibliometric data presents itself as a crucial problem. We proposed a completeness percentage of 99 with respect to publication data in impact evaluations of (small) university research groups. We argued that such a high percentage is necessary, because the distribution of citations over publications appears to be so skew, that for an impact analysis on a low aggregation level, missing only one publication and its citations can cause dramatic distortions of the results. With respect to citation data a completeness percentage of 95 percent seems both sufficient and technically achievable. Therefore, citations should be selected by "matching" only journal title, year, volume and page numbers, and *not* full (first) author names. Moreover, a number of checks on the citation data should by carried out in order to examine whether systematic errors were made when these data were collected.

(2) In order to *assign* publications and citations to research groups, one needs complete data for those groups: names of all researchers participating in the group, or even better: complete lists of articles published by the groups.

(3) The SCI data base may not always be an *adequate* tool to assess the performance of research groups in each particular field. We found quite large discrepancies between lists of publications given in University Research Reports and in our Leiden/SCI file, especially for fields like Mathematics and Taxonomy.

(4) Bibliometric indicators can be *disturbed* by a number of factors. First of all, citation practices can differ from field to field. We showed for example that in Biochemistry recent (two-years old) articles were, on an average, much more frequently cited than two-years old articles in Mathematics. As a consequence one should expect – and one actually observes – much higher short-term citation levels in Biochemistry than in Mathematics.

The immediate consequence of differences in citation practices, is that a comparison of citation-based performance indicators between disciplines is completely invalid. Even *within* disciplines, a comparison between subfields is not possible without an exhaustive investigation of the particular situation.

Second, we discussed the fact that citation practicies can also change in the course of time. We found that an SCI source article contains, on average, an increasing number of 0–2-years-old references during the period 1970–80. As a consequence, ones should expect – and again actually observes – an increase of the citations-per-publication ratio for all Leiden publications during 1970–80, with citations counted only during the first three years of a publication's life-time. Moreover, considerable disturbances in bibliometric trend analyses can arise from changes of the coverage of the SCI data base in the course of time.

(5) About the problem of *significance* of differences between bibliometric scores, we argued that this problem can be tackled only by working on more data per research group (which will be realized in our work after completing the updating procedure) and by developing a "theory" on the impact of a research group's complete scientific

oeuvre during a certain period of time, related to the distribution of citations amongst the individual publications constituting this *oeuvre*.

The bibliometric indicators discussed in this paper will be used as a "try out" in the forthcoming evaluation of all research projects in the Faculty of Medicine and the Faculty of Mathematics and Natural Sciences. This evaluation will be finished by the beginning of 1985. We shall report on our experiences with this "bibliometric try-out" in future publications.

References

[1] A.F.J. van Raan and J.G. Frankfort, An Approach to University Science Policy: A New Research Funding System, *International Journal of Institutional Management in Higher Education* 4 (1980) 155–163.

[2] B.R. Martin and J. Irvine, Assessing Basic Research: Some Partical Indicators of Scientific Progress in Radio Astronomy, *Research Policy* 12 (1983) 61–90.

[3] H.F. Moed, W.J.M. Burger, J.G. Frankfort and A.F.J. van Raan, *On the Measurement of Research Performance: the Use of Bibliometric Indicators* (Research Policy Unit of the University of Leiden, Leiden, 1983).

[4] K.H. Chang, *Evaluation and survey of a Subfield of Physics: Magnetic Resonance and Relaxation Studies in The Netherlands*, FOM Report 37175 (FOM, Utrecht, 1975).

[5] *Over Leven*, Report of the Verkenningscommissie Biochemie (National Survey Committee on Biochemistry), Staatsuitgeverij, The Hague, 1982 (in Dutch).