Jointly published by Akadémiai Kiadó, Budapest
and Springer, Dordrecht

Scientometrics, Vol. 76, No. 1 (2008) 3–21
DOI: 10.1007/s11192-007-1900-z

The evaluation of university departments and their scientists:

Some general considerations with reference to exemplary bibliometric publication and citation analyses for a Department of Psychology

GÜNTER KRAMPEN

Department of Psychology and Institute for Psychology Information (ZPID – Leibniz Institute),
University of Trier, Trier (Germany)

In reference to an exemplary bibliometric publication and citation analysis for a University Department of Psychology, some general conceptual and methodological considerations on the evaluation of university departments and their scientists are presented. Data refer to publication and citation-by-others analyses (PsycINFO, PSYNDEX, SSCI, and SCI) for 36 professorial and non-professorial scientists from the tenure staff of the department under study, as well as confidential interviews on self- and colleagues-perceptions with seven of the sample under study. The results point at (1) skewed (Pareto-) distributions of all bibliometric variables demanding non-parametrical statistical analyses, (2) three personally identical outliers which must be excluded from some statistical analyses, (3) rather low rank-order correlations of publication and citation frequencies having approximately 15% common variance, (4) only weak interdependences of bibliometric variables with age, occupational experience, gender, academic status, and engagement in basic versus applied research, (5) the empirical appropriateness and utility of a normative typological model for the evaluation of scientists' research productivity and impact, which is based on cross-classifications with reference to the number of publications and the frequency of citations by other authors, and (6) low interrater reliabilities and validity of *ad hoc* evaluations within the departments' staff. Conclusions refer to the utility of bibliometric data for external peer reviewing and for feedback within scientific departments, in order to make colleague-perceptions more reliable and valid.

Received May 9, 2007

Address for correspondence:
GÜNTER KRAMPEN
University of Trier, Department of Psychology and ZPID, D-54286 Trier, Germany
E-mail: krampen@uni-trier.de

0138–9130/US $ 20.00

Introduction

Scientific work is not only complex, but also very diverse, because it not only includes research and teaching (at least in universities) but publication activities, active and passive participation at conferences and scientific meetings, personnel management, academic self-administration, grant and fund raising, public relations etc. too. Therefore, personnel evaluation in university departments can not use single evaluation criteria. An empirical survey carried out in a representative sample of 265 German professors of psychology, working at university departments and research institutions, pointed – exemplarily for this academic discipline – at a very high consensus in subjective positive (42-times) or very positive (59-times) evaluations of 101 out of a total of 117 potential evaluation criteria for their own academic productivity [KRAMPEN & MONTADA, 2002, CHAPTER 1]. These evaluation criteria refer to a large-scale spectrum of qualitative (e.g., scientific originality, great deal of care over sampling, expenditure of empirical studies, creativity in data interpretation, engagement and commitment in teaching) as well as quantitative indicators (e.g., number of publications, citations by others, single and first authorships, editions, conference lectures, poster presentations, research grants, academic prizes and awards) of scientists' productivity. Of special interest is the fact that the clear majority of the sample (78%) knows the social norms (109 out of the total of 117 evaluation criteria) in their scientific community very good and conforms subjectively in their individual evaluations to them – i.e., German professors of psychology have very well integrated social orientations (referring to the terminology of [BREZNITZ, 1967]) regarding evaluation criteria of their own academic work.

Affected by these results is the controversy about qualitative and quantitative evaluation criteria, which often is discussed emphatically as an either-or debate (see, e.g., [CAMPANARIO, 1998; GUSTAFSON, 1975; RUSH & AL., 1996]). Neither an "either-or" nor a completely balanced "as well as" is the appropriate solution to this debate, because most, perhaps all, quantitative evaluation criteria are controlled and filtered by (qualitative) peer reviews (see Figure 1, modified adaptation from [KRAMPEN & MONTADA, 2002, CHAPTER 3]). Consequently, peer-reviews are prior to bibliometric evaluation criteria: In models of science and scientists' evaluation peer reviews are primarily, and scientometrics are secondarily, which, however, show significant feedback-loops to the peer-review procedures and their quality (see Figure 1), as well as to peer reviewing itself. Peer reviewing and evaluations are always influenced and co-determined by scientometric results (explicitly) or reviewers' considerations (implicitly) – explicitly and empirically founded on scientometric results *or* implicitly by the, at least partly random, i.e., subjective knowledge and/or stereotypes of reviewers.

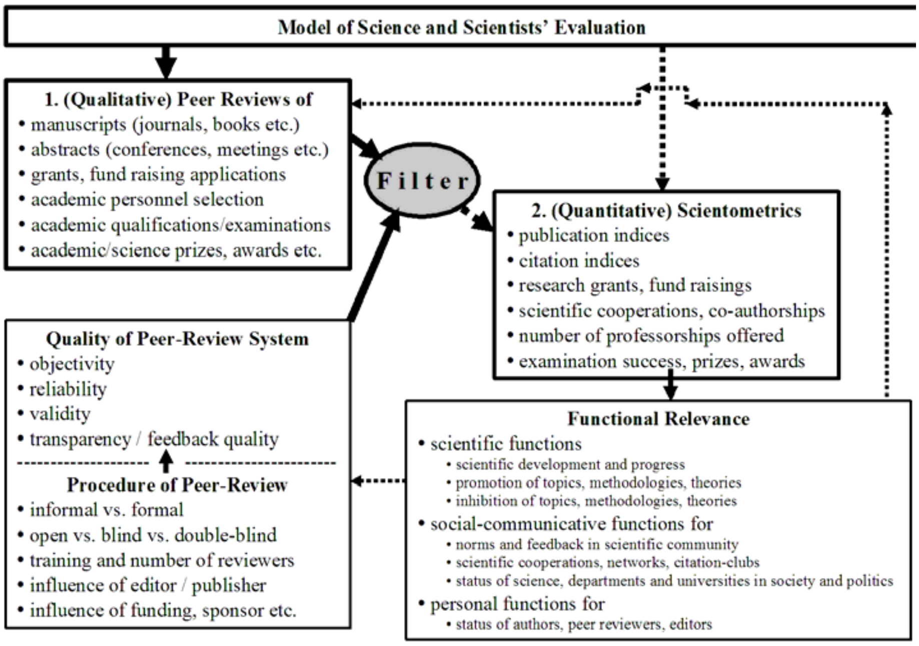
Model of Science and Scientists' Evaluation
1. (Qualitative) Peer Reviews of
• manuscripts (journals, books etc.)
• abstracts (conferences, meetings etc.)
• grants, fund raising applications
• academic personnel selection
• academic qualifications/examinations
• academic/science prizes, awards etc.
Filter
2. (Quantitative) Scientometrics
• publication indices
• citation indices
• research grants, fund raisings
• scientific cooperations, co-authorships
• number of professorships offered
• examination success, prizes, awards
Quality of Peer-Review System
• objectivity
• reliability
• validity
• transparency / feedback quality
Procedure of Peer-Review
• informal vs. formal
• open vs. blind vs. double-blind
• training and number of reviewers
• influence of editor / publisher
• influence of funding, sponsor etc.
Functional Relevance
• scientific functions
• scientific development and progress
• promotion of topics, methodologies, theories
• inhibition of topics, methodologies, theories
• social-communicative functions for
• norms and feedback in scientific community
• scientific cooperations, networks, citation-clubs
• status of science, departments and universities in society and politics
• personal functions for
• status of authors, peer reviewers, editors

Figure 1. Status and interrelations of peer reviews and bibliometric publication and citation indices in science evaluation (modified adaptation from [KRAMPEN & MONTADA, 2002, FIGURE 3.1, P. 51])

Because of its better objectivity, explicit foundation of peer reviewing and qualitative evaluations on bibliometric and other scientometric indices must be preferred. These arguments hold up at least for all peer-reviewed manuscripts, abstracts, grant applications, academic staff selections, science awards and academic examinations. They are the only ones being taken serious in modern sciences. To complete, functional relevance of bibliometric publication and citation analyses (within scientometrics) substantiated by the filters of the peer-review system is differentiated in Figure 1 for the domains of its scientific, social-communicative and personal functions, as well as their functional feedback to peer reviewing.

The model presented for science evaluation is true for the evaluation of single scientists, as well as for the evaluation of research units and university departments. Compensatory effects between scientists within a department are possible (e.g., some department members working hard on research, others on teaching or academic self-administration or fund raising, etc.), but not yet empirically confirmed, because different indicators of productivity – i.e., number of scientific publications, number of supervised first degree dissertations and of supervised PhDs – are correlated strongly (see, e.g., [DANIEL, 1986; KRAMPEN & MONTADA, 2002, CHAPTER 5]). Indeed and rather trivial, we have to accept inter- and intra-individual differences in scientific productivity as well. Therefore at all, evaluations of individual, institutional (on the level of research units, departments, and universities), and national differences in scientific productivity and output are necessary and useful. This has come into the focus of science research, science politics and the public in the last decades (see, e.g., [COLE & COLE, 1971; ENDLER & AL., 1978; GRAY, 1983; KRAMPEN & MONTADA, 2002; MAY 1997]).

There are many evaluation criteria. Their operationalization is more or less difficult. Their interrelations are partly known, partly unknow, and – being the rule – only some are selected in empirical evaluations and rankings. At least from a superficial point of view (see, e.g., [KRAMPEN & AL., 2007; SCHUI & KRAMPEN, 2006]) bibliometric publication and citation indices are easily and therefore frequently applied with reference to literature data-bases (such as PsycINFO with Anglo-American focus and PSYNDEX with its focus on psychological literature from the German-speaking countries, etc.) and citation databases (such as Social Science Citation Index, SSCI, and Science Citation Index, SCI). However, a lot of problems concerning such scientometric evaluation criteria are being questioned today. Some of them are investigated here empirically and exemplarily with reference to the tenure-track staff of a Department of Psychology at a University in the German-speaking countries.

Our research questions address to the evaluation of inter-individual differences in publication- and citation-quota *between* the scientists *within* the selected university department. Besides statistical parameters of publication and citation indices their frequency distributions, inter-correlations and some of their correlates (i.e., age, individual publication career, gender, academic status, engagement in basic versus applied research and teaching) are empirically studied. Of special interest are the frequency distributions of the bibliometric evaluation criteria within the department's staff: Do the distributions confirm more or less to the normal distribution (similar to most other achievement indicators) or are they skewed? Another focus of our research refers to the inter-correlations of publication-indices and citation-indices (only citations by others will be considered): Are there strong or rather weak correlations? Does controlling for age or occupational experience – i.e., individual publication years – has impact on such interrelations?

At the same time the typological scientists' model suggested by SCHUI & KRAMPEN [2006] will be tested firstly. Starting points of this normative model for the evaluation of scientists' research productivity and impact are the frequency distributions (1) of the number of publications and (2) of the number of citations by other authors within the academic staff of an institution. With reference to statistical parameters of the means or medians of both variables a 2 x 2 contingency table can be constructed (see Figure 2).

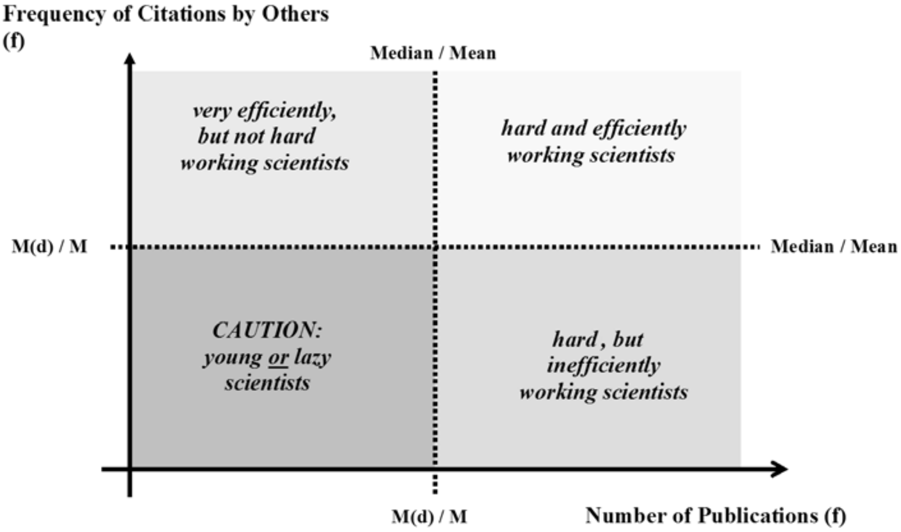
Frequency of Citations by Others (f)
Median / Mean
very efficiently, but not hard working scientists
hard and efficiently working scientists
M(d) / M
Median / Mean
CAUTION: young or lazy scientists
hard , but inefficiently working scientists
M(d) / M
Number of Publications (f)

Figure 2. Prototypes of scientists based on cross-classifications of the number of scientific publications and of the frequency of their citations by other authors (adapted from [SCHUI & KRAMPEN, 2006, FIGURE 3, P. 14])

The resulting four cells of the table (representing different types of scientists) prototypically can be characterized as

1. *very efficiently, but not hard working scientists,* having low numbers of own publications, which, however, are cited very frequently by other authors;
2. *hard and efficiently working scientists,* publishing a lot of papers and being cited frequently by other authors;
3. *hard, but inefficiently working scientists,* publishing a lot of papers, but being rather rarely or not cited by others at all;
4. *lazy or (!) young scientists* having published none or only few papers and being not or rarely cited by other authors. Caution: This type (or group of scientists) is ambiguous and – in contrast to the three other types – can only be interpreted individually with reference to age or duration of occupational experience of the individual scientist. I.e., older scientists with long careers can be characterized as "lazy", young scientists starting their academic career must be picked out, and the future will tell whether they become the efficiently, hard, inefficiently or lazily working type scientist.

Additionally, some of the bibliometric results will be compared to subjective evaluations of colleagues gathered in a small sample of the department staff. Co-variations of the bibliometric results and subjective cognitive evaluations of the productivity and engagement of all colleagues were studied with reference to 10 ratings of colleagues' productivity and engagement in research and publication, teaching, academic self-administration, public relations, fund raising as well as helpfulness and consideration. Contemporaneously, data were gathered with an exemplary other- and self-identification test to get some information about the validity of self-perceptions and the perceptions of one's own colleagues in comparison to objective publication and citation indices.

Methods

Methods of data collection

Bibliometric analyses on the numbers of papers, books, and book chapters published by every tenured scientist of the university department were run in PsycINFO (psychological literature database with Anglo-American focus including English, but only selected German journal publications from the German-speaking countries) and in PSYNDEX (focus on psychological literature from the German-speaking countries). Hits were counted – firstly – for *all* publications documented for each scientist (PsycINFO: 1890-2006; PSYNDEX: 1977–2006) and – secondarily, to reduce age effects – separately for the last seven publications years (2000–2006).

Citation analyses refer to Social Science Citation Index (SSCI) and Science Citation Index (SCI) excluding doublets and self-citations from the count. Hence, for every individual member of the department's staff, only citations by other authors in journals are taken into account. Time of all citation and publication analyses was March 2007.

Besides these quantitative, objective bibliometric analyses, data were gathered in private, *confidential semi-structured interviews* with a random selection of 20% of the department's staff ($n = 7$). Controlling for professorial versus non-professorial status, gender, age, and focus on basic versus applied research seven scientists were asked in the *first part of the interview* for subjective ratings for all of their 35 colleagues regarding to their

1. publication activities in the last seven years,
2. citation-by-others frequencies in the last seven years,
3. engagement in academic self-administration (with the instruction to consider academic status and duration of staff membership),
4. engagement for the department in total,
5. engagement for the image and ranking of the department,

6. scientific contributions and the creativity as well as originality of them,
7. scientific impact on psychology and its appreciation in the scientific community,
8. teaching quality,
9. engagement and commitment in teaching,
10. helpfulness and consideration as colleague.

All ratings refer to Likert-scales ranging from “not at all / none” (= 1 point) to “very much / a great deal” (= 10 points). With the exception of self-evaluations on these scales, 350 written ratings (10 scales x 35 colleagues) were given within the private interviews giving the possibility to raise queries and to reflect the subjective evaluations. Because all participants were psychologists, experienced with such ratings, time of data collection in this first part of the interviews did not exceed 30 minutes.

The *second part of the confidential interview* started with the presentation of a list of the numbers of publications (documented in PsycINFO and in PSYNDEX) and citations by other authors (documented in SSCI/SCI) for nine randomly selected colleagues and for the interview participant her- or himself. Firstly, with reference to these publication and citation frequencies, self-identification was asked for. Secondarily, the nine colleagues should be individually identified on the basis of the information given about publication and citation-by-others frequencies. Because of the confidentially and anonymity of the entire survey no feedback about the correctness of these judgements was given.

Sample

The sample refers to the academic staff (with contracts for an unlimited period; tenure track) of a Department for Psychology at a university in the German-speaking countries. Because confidentiality and anonymity of result presentation was promised, no names are communicated here and sample description has to be kept in general. The department under study is a larger one and achieved positions in the upper third of rankings published in the mass media in the last 10 years. Sample size is $N = 36$ scientists of whom 23 are members of the professorial staff (including emeritus professors from the last 10 years and assistant professors; subgroup of “professors” in the following) and 13 are members of the non-professorial staff with a PhD (without postgraduates and without fixed-term PhDs; subgroup of “PhDs” in the following) – all having a tenure. The age of the sample studied ranges from the mid 30’s to the mid 70’s. More than one fifth of the staff is female. The years of the first scientific publications range from the late 60’s to 2005. A slight majority of the sample is engaged in basic research and teaching, the others are engaged in applied psychology research and teaching.

Results

Publication and citation frequencies and their distributions

Statistical characteristics of all bibliometric indicators under study are given in Table 1 for the whole sample. Rather huge numerical values in the standard deviations and ranges of all bibliometric variables point at distinctive inter-individual differences within the department's staff. In addition, skewness and kurtosis of all frequency distributions are large, and *Kolmogoroff–Smirnov Tests for Normality* (with Lilliefors-limits; [LILLIEFORS, 1967]) reach statistical significance, disproving the hypothesis of normal distribution for all five bibliometric variables within the sample. Therefore, non-parametric statistical tests and descriptive parameters must be used in the following in all statistical analyses.

Considering the large ranges, the *medians (M_d)* presented in Table 1 show that the 36 scientists under study on average (1) have published in total 26 (PSYNDEX) and 15 (PsycINFO) papers respectively, (2) nine (PSYNDEX) and four (PsycINFO) papers respectively of these are documented for the last seven publications years (2000–2006) in the databases, and (3) the publications of the scientists are cited between 2000 and 2006 on average 23-times by other authors.

Non-parametric Wilcoxon-Matched-Pairs Signed Tests confirm significant differences between the numbers of publications documented in PSYNDEX and PsycINFO ($Z = 4.44$, $p < 0.01$ for 2000–20006 publications; $Z = 4.98$, $p < 0.01$ for all publications documented in the databases). This can be explained easily: PsycINFO includes only English- and some selected German-language journal publications of psychologists from the German-speaking countries while PSYNDEX includes all psychological publications form the German-speaking countries, i.e., books, chapters, dissertation thesis, and more German journal papers too.

Numbers of *publications documented per publication year* in PSYNDEX ($M_d = 2$; range: 0.3–9.8) and PsycINFO ($M_d = 1$; range: 0.0–3.8) are significantly different as well ($Z = 4.99$, $p < 0.01$). The median of the *number of citations by other authors per publication year* is $M_d = 3$ (range: 0–23). More interesting are the *numbers of citations by others per publication documented in PSYNDEX versus PsycINFO:* On average each paper documented in PSYNDEX is cited by other authors 2-times (range: 0–36), each paper documented in PsycINFO 6-times (range: 0–108), a difference which is statistically significant as well ($Z = 4.64$, $p < 0.01$). Consequently, internationally more visible publications (documented in PsycINFO with its Anglo-American focus) are more frequently cited (three time more indicated by SSCI and SCI) than those documented in a literature database with a non-English focus and a lesser international visibility accordingly.

Table 1. Statistical parameters, intercorrelations and correlates (Kendall's Tau, τ) of the bibliometric publication- and citation-indices ($N = 36$)

Statistical parameter	Number of publications *(f)* documented in				Citations by others *(f)* in SSCI und SCI 2000–2006
	PSYNDEX 2000–2006	PSYNDEX 1977–2006	PsycINFO 2000–2006	PsycINFO 1890–2006	
Mean *(M)*	49.9	12.3	24.6	6.2	72.1
Standard deviation *(SD)*	60.4	13.8	27.9	7.1	125.3
Median (M_d)	26	9	15	4	23
Range (min-max)	1–283	1–67	0–110	0–36	0–648
Skewness	2.26	2.60	1.60	2.34	3.40
Kurtosis	5.87	7.62	2.09	7.74	13.30
KSA-Test (Lilliefors)	0.25**	0.23**	0.18**	0.20**	0.24**
Intercorrelations (Kendall's τ)[a]					
PSYNDEX 1977–2006	1.00	0.51**	0.73**	0.45**	0.49**
PSYNDEX 2000–2006	0.53**	1.00	0.43**	0.61**	0.37**
PsycINFO 1890–2006	0.69**	0.43**	1.00	0.56**	0.57**
PsycINFO 2000–2006	0.44**	0.60**	0.56**	1.00	0.50**
Citations 2000–2006	0.45**	0.48**	0.54**	0.50**	1.00
Correlates (Kendall's τ)					
Age	0.51**	0.14	0.44**	0.17	0.28*
Publication years	0.54**	0.11	0.51**	0.17	0.30*
Sex[b]	0.24	0.08	0.17	0.07	0.24
Academic status[b]	0.32*	0.27	0.42**	0.35*	0.36**
Basic/applied research[b]	0.03	0.00	–0.12	–0.29**	–0.15
Citations per publication in PSYNDEX (total)	0.06	0.10	0.20	0.31**	0.57**
Citations per publication in PsycINFO (total)	0.02	0.11	–0.14	0.12	0.48**

**p < .01; *p < .05;

[a] Above the main diagonal rank-order correlations Kendall's Tau, below the main diagonal partial rank-order correlations (controlling for age);

[b] Biserial rank-order correlations (female, PhD, basic research = 1; male, Professor, applied research = 2).

Frequency distributions of the five bibliometric variables under study differ significantly from normal distributions (see Table 1). All the distributions are skewed, showing the pattern of *Pareto-distributions* presented graphically in Figures 3 to 5. Skewness is marked, showing absolute peaks left for highest frequencies of rather low numbers (1) of publications documented in PsycINFO and PSYNDEX in total (see Figure 3) and (2) of papers documented in the last seven publications years (2000–2006; see Figure 4), as well as (3) of citations by other authors (see Figure 5). Long stretches across the middle ordinate-values to the right side of the ordinates point at outliers whose identifications result in an astonishing pattern: The three most extreme outliers in Figure 5 (with very high citation frequencies) are the same scientists having published extreme high numbers of papers (Figures 3 and 4). Four to six other scientists represent the next group with not such extreme, but also rather high numbers of publications and citations – again identical people (see Figures 3 to 5).

Hence, there is a homogeneous *first group of three extreme outliers* (n = 3) which must be excluded in correlation analyses. The four to six scientists of the second group ($4 \le n \le 6$) are no (bibliometric) outliers, but show homogeneously rather high publication performances with frequent citations by other authors. All other scientists of the department's staff represent a large ($27 \le n \le 29$), very heterogeneous third bibliometric group with very different individual locations in Figures 3, 4 and 5, depending on what – the number of publications in PsycINFO versus PSYNDEX or the number of citations by others – is focussed.

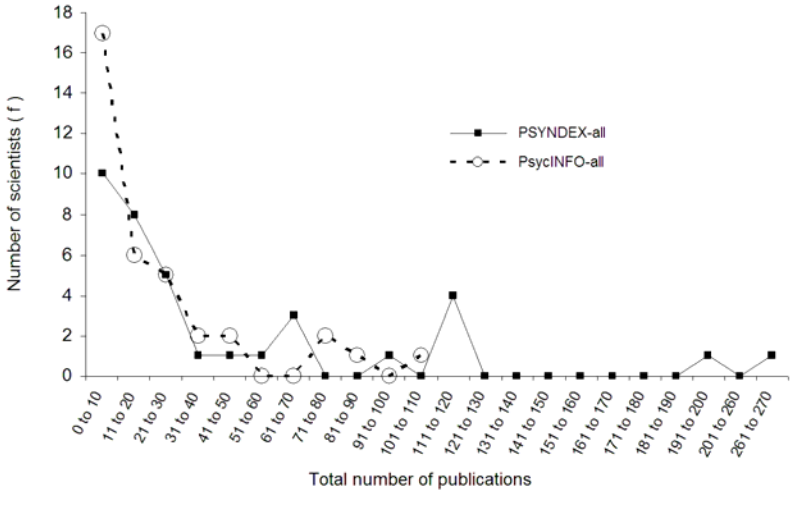

Figure 3. Frequency distributions of the number of all publications documented in PSYNDEX and PsycINFO (= 36)

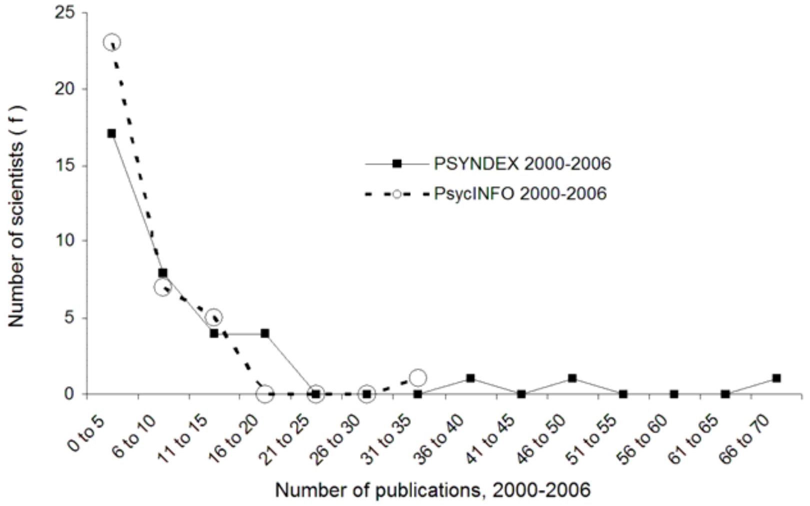

Figure 4. Frequency distributions of the number of publications in 2000–2006 documented in PSYNDEX and PsycINFO (N = 36)

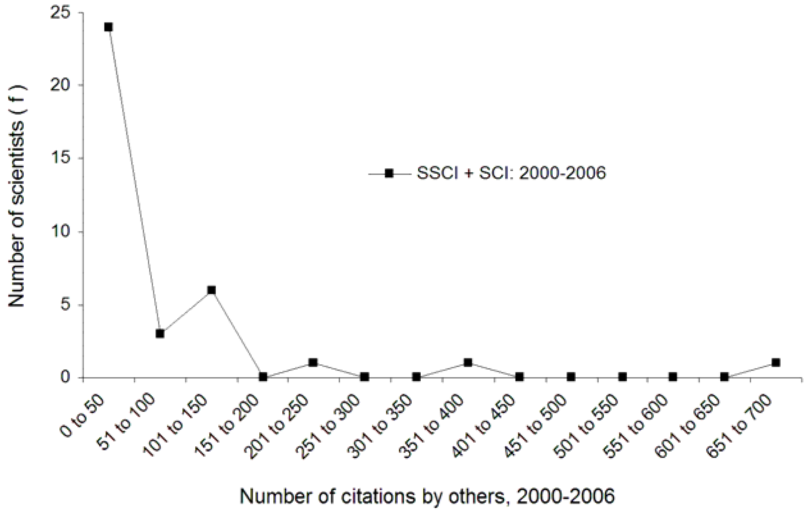

Figure 5. Frequency distrubutions of the number of citations by other authors in SSCI and SCI 2000–2006 (N = 36)

Interdependences of publication and citation frequencies

Rank-order correlation coefficients of the five bibliometric variables are presented twice in the middle of Table 1: Firstly, above the main diagonal in terms of Kendall's Tau, and secondarily, below the main diagonal in terms of partial rank-order correlations controlling for age. Because of no significant differences ($p > 0.10$) between the coefficients above versus below the main diagonal it is concluded that age as well as occupational experience (number of individual publication years is correlated with age to $r = 0.92$, $p < 0.01$) does *not* moderate the ordinal interdependences of the rank-orders of publication frequencies and citation frequencies. All correlations are statistically significant uniting a common variance of not more than 14%–32% (see Table 1).

Hence, there are significant interdependences between publication activities and the frequencies of being cited by other authors. However, the relationships are not as marked as expected and often assumed. This is exemplarily illustrated in Figure 6 by the plot of the correlation between the number of publications documented in PsycINFO (2000–2006) and the number of citations by other authors in SSCI/SCI (2000–2006) excluding the three outliers described above. Kendall's Tau drops after outlier exclusion to $\tau = 0.37$ ($p < 0.05$), and the plot shows a rather heterogeneous pattern.

Excursus: Because of general methodological interest and for the sake of the prevention of rather frequent statistical mistakes, shortly the outlier exclusion effects will be described. While Kendall's Tau drops from $\tau = 0.50$ ($p < 0.01$; $N = 36$) to $\tau = 0.37$ ($p < 0.05$; $n = 33$) after the exclusion of the three outliers described above, the (statistically wrong) computation of the parametric Pearson's correlation coefficient shows the stepwise effects even more distinctively: for the whole sample Pearson's correlation coefficient is $r = 0.79$ ($p < 0.01$; $N = 36$); exclusion of the first, most extreme outlier results in $r = 0.52$ ($p < 0.01$; $n = 35$); additional exclusion of the next outlier results in $r = 0.50$ ($p < 0.01$; $n = 34$); and the exclusion of all three outliers results in $r = 0.37$ ($p < 0.05$; $n = 33$). These are very different "results" for the relationships of publication and citation frequencies. Common variance of both variables varies between 14% and 62% making up a difference which is not only statistically but also practically significant. The error in statistical analyses can easily be avoided by (1) descriptive outlier analysis and outlier exclusion and (2) systematic tests of frequency distribution patterns resulting in the correct decision for parametric versus non-parametric statistics. If this is neglected, rather huge overestimations for the interrelationships of publication and citation frequencies are the result, which *de facto* are rather low, showing a common variance of 14% only. <*End of Excursus*>

Prototypes of scientists' research productivity

Figure 6 refers to the graphical illustration of the empirical test of the normative *evaluation typology for scientists' productivity and impact in research* described in the introductory part of this paper (see Figure 2) too. Using the medians *(M_d;* see Table 1) of the numbers of publications documented in PsycINFO (2000–2006) and of the frequencies of citations by other authors (SSCI/SCI 2000–2006) Figure 6 is segmented into four cells characterizing typologically scientists' productivity and impact in research as very efficiently (but not hard) working, hard and efficiently working, hard- but inefficiently working, and young or lazy types (see Figure 2). The three extreme outliers discussed above are disregarded in Figure 6, because their inclusion would lead to a strongly biased, rather empty expansion of the plot to the above right side (the second cell right above) coarsening the data presented. Numerical results for the tenure track staff of the Department of Psychology under study are:

- *n = 3 very efficiently, but not hard working scientists (8% of the tenure staff)* with only few publications which, however, are cited very frequently by other authors. Inspections of their papers and citations show that for each scientist there is only one paper being responsible for the high citation frequencies. These papers refer to literature overviews, which have been published under co-authorship together with an international visible senior scientist. Age and occupational experience of these three scientists refer once to a young scientist with two publication years only and twice to older, experienced scientists with many publication years, but only a few publications. However, one of these was cited very frequently.
- *n = 12 hard and efficiently working scientists* having published a lot of papers which are cited by other authors frequently. Adding our three outliers (see above), this group includes 15 scientists, making up *42% of the department's tenure staff*. This group is heterogeneous in age and occupational experience.
- *N = 4 hard, but inefficiently working scientists (11% of the tenure staff)* having published rather many papers which are, however, never or rather rarely cited by other authors. The proportion of younger and experienced scientists is equal in this group.
- *N = 14 scientists are in the ambiguous group (39% of the staff)* taking together young scientists with low experience and older scientists with long occupational experience mixed, because they have low numbers of publications and citations by other authors in common. Inspections of scientists' data in this group easily clarifies the groups' heterogeneity: Only *four of these scientists are young, having only a few publication years (11% of the tenure staff),* and the majority is experienced to be characterised here as being *lazily and inefficiently in research (n = 10; 28% of the tenure staff).*

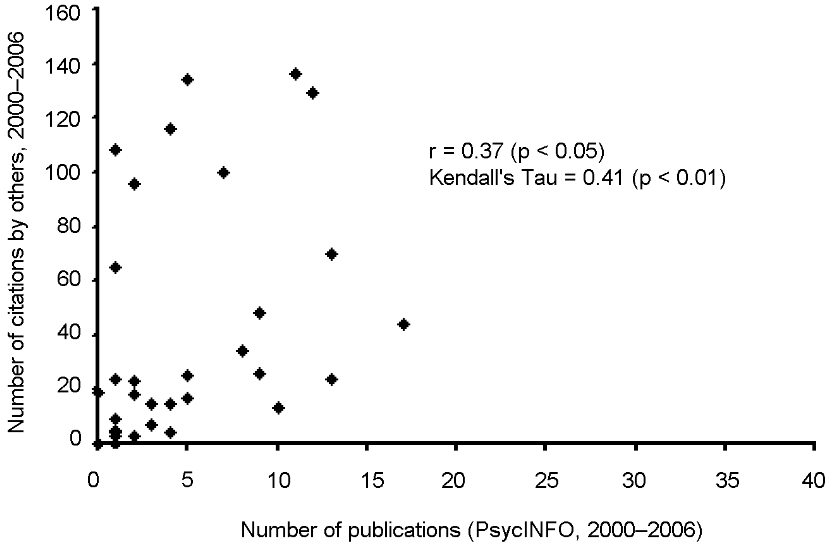

Figure 6. Correlation between publications documented in PsycINFO and citations by other authors in SCI/SSCI for publication years 2000–2006 (after elimitation of 3 outliers, N=33)

To summarize, the pattern of the department's tenure staff research productivity is heterogeneous: Approximately 40% is hard and efficiently working, 30% is lazy (at least in research), 10% is hard, but working inefficiently, another 10% is young and only the future will tell about their professional development in research, and the last approximately 10% are somewhat like fortune-hunters, having one cooperation with an international visible senior scientist, resulting in one high frequently cited paper. However, it must be considered that this evaluation typology refers to research productivity only. Until now, other professional evaluation criteria (engagement in teaching, academic self-administration, grants, etc.) are neglected. Additonally, it must be noted that the evaluation of scientists' research productivity refers to an in-group-comparison within one and only one department's tenure staff only, excluding comparisons with other university departments (with other reference norms, i.e., medians in publication and citation frequencies).

Correlates of and group differences in bibliometric variables

Rank-order correlation coefficients of the five bibliometric indicators under study to some socio-demographic and occupational variables are presented in the lower part of Table 1. As expected, *scientists' age and occupational experience* (i.e., number of publication years) are significantly correlated to the number of all publications documented in PsycINFO and PSYNDEX, as well as to the number of citations by other authors. However, the correlation coefficients of both variables to the numbers of papers documented in PsycINFO and PSYNDEX for the last seven publication years are not significant, pointing again at the independence of age and occupational experience with more recent bibliometric indicators of scientific productivity and impact.

There are no statistical significant (biserial) rank-order correlations between *scientists' gender* and any of the bibliometric variables (see Table 1) which is supported by insignificant group differences in the *Mann-Whitney U-Test* ($U > 75$, $z < 1.72$; $p > 0.10$). *Academic status* (i.e., tenure professorship versus tenure non-professorial PhD) is correlated significantly to the number of publications and citation frequency not exceeding a common variance of 18%. Engagement in *basic versus applied psychological research* is significantly, but weakly correlated only to the number of papers documented for the last seven publication years in PsycINFO, indicating basic researchers recently have had published in English journals somewhat more frequently (see Table 1).

In addition, the results of analyses of variance, not presented here in detail, show that *variances within groups* (grouping variables: basic versus applied researchers, professorial versus non-professorial status, female versus male) are – in terms of mean squares *(MS)* – for the most bibliometric variables greater than the variances between groups. This confirms earlier results [DANIEL, 1986; KRAMPEN & MONTADA, 2002] on large inter-individual differences within groups of scientists (either within university departments or within academic or socio-demographic groups) exceeding the differences between such groups. To say it provocatively, within each group of scientists there are always some researching, publishing and being cited a lot, and some others researching, publishing and being cited not at all or rather rarely. Until now, there are no empirical findings supporting a "bell-wether"-hypothesis promoting "go-with-effects" within groups of scientists.

Frequencies of *citations by other authors per publication* documented in PSYNDEX and in PsycINFO are significantly correlated to the absolute number of SSCI/SCI-citations (see Table 1). With one exception (number of papers documented in PsycINFO 2000–2006), however, this is not the case for the numbers of publications pointing again (see above) at a rather low relationship between publication activities and citation-by-others frequencies (which is overestimated frequently).

Self-identification and colleague-identification on the basis of bibliometric parameters

Within the confidential interviews with a small sub-group of seven randomly – but controlling for academic status, gender, age, and basic versus applied research fields –

selected scientists the anonymous presentation of the bibliometric publication and citation frequencies of nine randomly selected colleagues and of the interview partner her- or himself, resulted in very good self-identifications and very bad identifications of colleagues: With reference to publication numbers documented in PsycINFO and PSYNDEX, as well as citation frequencies (SSCI/SCI), all seven scientists identified her- or himself correctly in the list including 10 anonymously presented individual data sets. Thus, *self-identification* succeeded in 100% without problems. However, all interview partners missed rather high numbers of own publications not documented in PsycINFO (but in PSYNDEX), and some of them – i.e., researchers working in the fields of bio- and neuropsychology as well as general psychology – wondered about the fact that very short papers (like conference proceedings, abstracts) are neither documented in PSYNDEX nor in PsycINFO.

Second task within the interviews was the identification of nine colleagues from the departments' tenure staff with help of the bibliometric data. Out of the 63 completed judgements – resulting from seven scientists with nine assignments each – only seven were correct. Hence, *perceptions of colleagues' publication and citation-by-others status* in comparison to objective publication- and citation-indices resulted in a disaster with only 11% correct answers.

Subjective evaluation of one owns colleagues and its validity

Furthermore, confidential interview-data refer to 10 evaluation ratings for all 35 colleagues of the departments' tenure staff (10 ratings x 35 colleagues x 7 raters). However, these data-analyses were lavishly, leading to disappointing results presented here only in an overview. While average rank-order correlation coefficients (after z-transformations) of subjectively rated *"publication activities in the last seven years"* of colleagues and the bibliometric "truth" resulted in significant, but rather low values (PSYNDEX 2000–2006: $r = 0.25$; PsycINFO 2000–2006: $r = 0.19$; $p < 0.01$), none of the nine other ratings correlated – on average – statistically significant with any of the five bibliometric variables. Thus, the five bibliometric indicators under study are not in accordance with scientists' subjective evaluative representations of their colleagues in the own department.

Additionally, the results give no indication to any compensatory effects between staff members with reference to different evaluation criteria. The inter-correlation matrix for the ten ratings of all colleagues gathered is heterogeneously without outstanding (high) correlation coefficients, none of them exceeding $r = |0.35|$ and none of them falling below $r = –0.18$, pointing by this at no significant compensatory effects. All this may be a result of the low interrater-reliabilities.

In fact, *agreement of the seven interview partner on the evaluations of their colleagues* is very low: On average, interrater-reliability for all of the ratings and all of the rated colleagues is $r = 0.09$ ($p > 0.10$), pointing in total at no consensus in the evaluation of colleagues with reference to 10 different evaluation criteria. More detailed data-analyses show some variations in interrater-reliability between the 10 ratings and significant agreements on four of them: Higher and significant agreements result on the ratings of "publications activities" ($r = 0.23$; $p < 0.01$), "engagement for the department in total" ($r = 0.24$; $p < 0.01$), "helpfulness and consideration as colleague" ($r = 0.36$; $p < 0.01$), and "engagement in academic self-administration" ($r = 0.44$; $p < 0.01$). These evaluation criteria are somewhat more visibly which is especially true for the last mentioned: At the lowest, the individual data-analysis level there was one very high interrater-reliability ($r = 0.84$; $p < 0.01$). It refers to the "engagement in academic self-administration" of one and only one colleague, holding outstanding administrative positions in the university and in the department for many years.

Discussion and conclusions

Besides the rather heterogeneous picture described bibliometrically for the Department of Psychology under study, there are some more general aspects of results of the evaluation study presented here exemplarily. Firstly, it must be noted that the distributions of all bibliometric indicators on publication and citation frequencies under study within a departments' staff are extremely skewed (Pareto-distributions), requiring non-parametrical statistical analyses.

Secondly, for all bibliometric variables outlier analyses must be done, which must lead – at least in some of the data analyses – to outlier exclusions. In the study presented, there were three personally identical outliers in all of the five bibliometric variables, whose inclusion in data analyses would have resulted in strong biases. This was demonstrated for the correlations of publication numbers and citation-by-other-authors frequencies. Use of adequate rank-order correlation analyses and outlier exclusion showed that this relationship is frequently overestimated. Common variance of both variables is approximately 15%, at maximum 25% putting the hypothesis of strong relationships as being a scientometric myth in question. Correlation analyses controlling for age and occupational experience (i.e., number of individual publication years) led to very similar patterns of results. Both variables show significant bivariat co-variations with the total numbers of publications documented in literature databases, but not with the numbers of publications documented for the last (seven) publication years. Indeed, differences between groups of scientists within the department under study are rather weak with reference to grouping variables like age, gender, occupational experience, academic status, and engagement in basic versus applied research. The fact that variances of bibliometric variables within these groups are greater than the variances between these groups, confirms earlier results [DANIEL, 1986; KRAMPEN & MONTADA, 2002] on large inter-individual differences within groups of scientists, exceeding largely the differences between such groups.

The first empirical test of the evaluation typology for scientists' research activities presented by SCHUI & KRAMPEN [2006] based on cross-classifications of scientists' publication- and citation-by-others frequencies succeeded. Using the medians of both bibliometric variables four prototypes of (1) hard versus not hard and (2) efficiently versus non-efficiently working scientists within the department under study were empirically reconstructed: Approximately 40% of the department's staff is hard and efficiently working, 30% is lazy (at least in research), 10% is hard, but inefficiently working, another 10% is young and only the future will tell about their professional development in research, and the last approximately 10% is somewhat like fortune-hunters having one cooperation with an international visible senior scientist resulting in one high frequently cited paper.

These proportions are very specific for the Department of Psychology under study, giving a rather ambiguous picture, because of the high proportion of nearly one third of the staff's members as being lazy in research. Recently published rather gross evaluation analyses for German professors [KAMENZ & WEHRLE, 2007] resulted in an estimation of 5% as being lazy in *all* duties, 45% as being lazy at least in some of their duties, and 50% working in accordance with their duties or more. These estimations include other duties and evaluation criteria (i.e., engagement in teaching and student supervision, academic self-administration, personnel management, grant and fund raising, public relations, active and passive conference participations, etc.; see, e.g., [KRAMPEN & MONTADA, 2002], which were neglected in the presented bibliometric analyses. However, the bibliometric and typological methods of the analysis presented here exemplarily for a Department of Psychology can become part of a broader evaluation methodology including other evaluation criteria. First empirical tests confirm its applicability and utility, and the results may be used as a meaningful and effective feedback for the staff of scientific departments.

Certainly, the bibliometric approach implemented in the present study needs enlargements and completions by other evaluation criteria. Herewith, the hypothesis could be tested empirically that engagement and productivity of the scientists within one department in different fields of work (i.e., teaching, research, administration, personnel management, grants, etc.) may be compensatory. However, neither existing results [DANIEL, 1982; KRAMPEN & MONTADA, 2002] nor the results of the confidential interviews presented here confirm this hypothesis. On the contrary, interview data presented point at a low validity of the subjective evaluations of the colleagues within ones' own department. Together with the low interrater-reliabilities of colleagues evaluations, the results convey a picture of stereotyped and strongly biased perceptions of the staff and its members. The conclusion must be that *ad hoc* evaluations within the staff of a department are neither reliable nor valid – this approach should be forgotten in evaluation research and applications, because the results do not proof anything. However, feedback of objective bibliometric data can help to make such social perceptions within scientific departments more reliable and valid. The same is true for the underpinning and empirical validation as well as crosscheck of external peer-reviewing by objective, professionally compiled bibliometric data.

References

BREZNITZ, S. (1967), Confidence estimation of group norm as a function of subjective conformity, *Psychonomic Science,* 7 : 399–400.

CAMPANARIO, J. M. (1998), Peer review for journals as it stands today – Part 1 and 2, *Science Communication,* 19 : 181–211 and 277–306.

COLE, J., COLE, S. (1971), Measuring the quality of sociological research: Problems in the use of the Science Citation Index, *American Sociologist,* 6 : 23–29.

DANIEL, H.-D. (1986), Die Vermessung der Forschung (The measurement of research). In: H. METHNER (Ed.), *Psychologie in Betrieb und Verwaltung (Psychology in Oganisations and Administrations).* Deutscher Psychologen Verlag, Bonn (Germany), pp. 208–218.

ENDLER, N. S., RUSHTON, J. P., ROEDIGER, H. L. (1978), Productivity and scholary impact (citations) of British, Canadian, and U.S. Departments of Psychology (1975), *American Psychologist,* 33 : 1064–1082.

GRAY, P. H. (1983), Using science citation analysis to evaluate administrative accountability for salary variance, *American Psychologist,* 38 : 116–117.

GUSTAFSON, T. (1975), The controversy over peer review, *Science,* 190 : 1060–1066.

KAMENZ, U., WEHRLE, M. (2007). *Professor Untat (Professor Not-work),* Econ Verlag, Berlin, Germany.

KRAMPEN, G., MONTADA, L. (2002), *Wissenschaftsforschung in der Psychologie (Science Research in Psychology),* Hogrefe, Göttingen, Germany.

KRAMPEN, G., BECKER, R., WAHNER, U., MONTADA, L. (2007), On the validity of citation counting in science evaluation, *Scientometrics,* 71 : 191–202.

LILLIEFORS, H. W. (1967), On the Kolmogoroff-Smirnov test for normality with mean and variance unknown, *Journal of the American Statistical Association,* 62 : 399–402.

MAY, R. M. (1997), The scientific wealth of nations, *Science,* 275 : 793–796.

RUSH, A. J., GULLION, C. M., PRIEN, R. F. (1996), A curbstone to applications for National Institute of Mental Health grant support, *Psychopharmacological Bulletin,* 32 : 311–320.

SCHUI, G., KRAMPEN, G. (2006), Bibliometrische Indikatoren als Evaluationskriterien: Möglichkeiten und Grenzen (Bibliometrical indicators as evaluation criteria: Possibilities and limits). In: G. KRAMPEN, H. ZAYER (Eds), *Didaktik und Evaluation in der Psychologie (Teaching Methods and Evaluation in Psychology).* Hogrefe, Göttingen (Germany), pp. 11–26.