Correlations of assessments of research group quality and productivity with group size, output numbers, and normalized impact[1]

Tim C.E. Engels, Nele Dexters, and Birgit Houben

tim.engels@ua.ac.be
Department of Research Affairs and Centre for R&D Monitoring (ECOOM), University of Antwerp, Middelheimlaan 1, 2020 Antwerp (Belgium);
Antwerp Maritime Academy, Noordkasteel-Oost 6, 2030 Antwerp (Belgium)

nele.dexters@ua.ac.be
Centre for R&D Monitoring (ECOOM), University of Antwerp, Middelheimlaan 1, 2020 Antwerp (Belgium)

birgit.houben@ua.ac.be
Department of Research Affairs, University of Antwerp, Middelheimlaan 1, 2020 Antwerp (Belgium)

Abstract
We present an analysis of the extent to which group size, output numbers and normalized impact of research groups correlate with assessments of the quality and the productivity of these groups. Data are taken from an on-going program of research assessments by international peer panels at the University of Antwerp, Belgium. For a set of 52 science research groups we find that assessments of quality are highly correlated with the size of research groups, but less so with the number of publications in high impact journals and not with the normalized citation impact of the articles published. Assessments of productivity appear to be highly correlated with the amount of funding acquired, the number of articles published as well as the number of PhDs defended.

[1] The authors thank everybody who has contributed to the success of the research assessments organized at the University of Antwerp. The assistance of the academic authorities in this process is greatly appreciated. The authors have written this article in their personal capacity.

Introduction

This paper presents results of an analysis of the extent to which size, output numbers and normalized impact of research groups correlate with assessments of the quality and the productivity of these groups. The question whether there exists an optimal size of research groups is of interest to bibliometricians from a fundamental as well as a practical viewpoint (Kretschmer, 1985; Richardson, 2011). Recently, Kenna & Berche (2011) have addressed this issue using data from the British Research Assessment Exercise (RAE) as well as from the French equivalent organized by the "Agence d'évaluation de la recherche et de l'enseignement" (AERES). They contend that for most disciplines a phase transition in the correlation of research group size with assessments of quality can be observed, illustrating the fact that from a certain size onwards adding additional researchers will not result in higher assessment scores. In a report to the British University Alliance, Adams & Thomson (2011) come to a similar conclusion. Furthermore, they hypothesize that quality is driving growth rather than the other way around. One important drawback of studies drawing on data from the British RAE is that these assessments take place at the departmental level (Hicks, 2012). Indeed, although universities need not necessarily submit entire departments to the RAE, the size of the 'units of assessments' resembles more that of departments than that of research groups in which day-to-day scientific interaction takes place. This is apparent especially when taking into account that 'size' in these studies is defined as the number of tenured and tenure track researchers only (Adams & Thomson, 2011). Because no country has implemented a performance-based university research funding system that assesses research performance at the level of research groups (Hicks, 2012), few studies regarding such assessments have been published. Research groups, however, are the most important work floor unit in research (van Raan, 2006a), thus rendering analysis at this micro level essential to understanding quality and productivity dynamics in science.

In Flanders, the northern part of Belgium, the Vrije Universiteit Brussel (VUB) decided in 1996 to introduce a continuous cycle of peer-review research evaluations per discipline (Rons, De Bruyn, & Cornelis, 2008). The University of Antwerp (UA) initially implemented periodic university-wide evaluations, and then in 2007 switched to a cyclic system of research evaluations per discipline with the research groups of the university as the elements of assessment (Spruyt & Engels, 2010). Following a comparative study of the commonly used research assessment systems (see, for two recent overviews, European Commission, 2010; Hansen, 2009), the University opted to closely follow the Dutch Standard Evaluation Protocol (SEP: VSNU, 2003; VSNU, 2009), in particular as far as assessment criteria and scale are concerned (see Data and Methodology). Moreover, the university decided to provide the peer panels with all relevant documentation regarding the performance of the research groups, including bibliometric indicators. The inclusion of bibliometric indicators in the documentation allowed for their validation by the research groups themselves before submission to the peer panels and made it clear to all involved that bibliometric indicators and peer review can be used in combination (Moed, 2007). In this paper we draw on this documentation, as well as the quality and productivity assessments of the research groups, to address the question of which research group characteristics best predict quality and productivity. In particular, we look at whether quality assessments correlate with the size of research groups, with the amount of publications in high impact journals, and with the normalized citation impact of the published articles of the groups. With regard to productivity, we analyse the extent to which productivity assessments are correlated to the amount of funding acquired, to the number of articles published, and to the number of PhDs defended.

Data and methodology

In this paper we report data on 52 science research groups of the University of Antwerp[i] that have been assessed in the period 2008-2011 as part of the evaluations of the departments of Computer Science (7 groups), Chemistry (12 groups), Pharmaceutical Sciences (11 groups), Physics (9 groups), Biology (9 groups), and Mathematics (4 groups). In all cases the official period under evaluation were the preceding calendar years t-8 to t-1. In total 158 professors and 973 contract researchers have been involved. For the assessment of each of these disciplines, a panel of between four and seven peers visited the university during two or three days. On each occasion, the assessors enter into a dialogue with the academic authorities, representatives of the departments and faculties to which the research groups belong, and, last but not least, with the spokespersons and professors of these research groups. This approach allows the panel members to gather additional information over and above the content of the self-evaluation report. The potential drawbacks of evaluations by committees of peers notwithstanding (Hansson, 2010; Langfeldt, 2004), this approach ensures active preparation and involvement of all parties and allows for collegial and interactive decision making by the panel members.

In accordance with the Dutch SEP protocol (VSNU, 2003), the peer panels assessed the quality and the productivity of each research group on a five point scale (including the possibility to assign halves):

- Excellent (5): Work that is at the forefront internationally, and which most likely will have an important and substantial impact in the field. The research group is considered an international leader.
- Very good (4): Work that is internationally competitive and is expected to make a significant contribution; nationally speaking at the forefront in the field. The research group is considered an international player and a national leader.
- Good (3): Work that is competitive at the national level and will probably make a valuable contribution in the international field. The research group is considered internationally visible and a national player.
- Satisfactory (2): Work that is solid but not exciting, will add to our understanding and is in principle worthy of support. The research group is nationally visible.
- Unsatisfactory (1): Work that is neither solid nor exciting, flawed in the scientific and or technical approach, etc.

With regard to quality, the panels were instructed that quality refers to (1) the originality and the innovativeness of the research, (2) the importance of the questions and problems addressed, (3) the co-ordination, focus and planning of the research, (4) the prominence of (the members of) the research group, and (5) the quality of the (key) publications of the research group. Not all of these aspects can be operationally defined in terms of quantitative data provided in the documentation to the peer panels.

In this study we use group size as a proxy for prominence of the group. Group size is defined as the number of full-time equivalents (FTE) available for research and includes non-tenured researchers in order to faithfully reflect the research capacity of the groups. As a proxy for the quality of publications, we use the number of publications in journals with a Journal Citation Reports (JCR) impact factor in the top 20% of the discipline (defined as subfields as in Glänzel

& Schubert, 2003). Thus, regardless of the total output of the group, we take the number of best quality publications, i.e. publications in the journals belonging to the top 20% of journals with the highest impact of the disciplines in which a group is active. A similar indicator for high quality publications, Q1, is used in the SIR World Report 2011 (SCImago research Group, 2011). By using the number of best quality publications, as opposed to a ratio or an indicator based on the average of the impact factors of the complete set of journals the group has published in, we avoid punishing high productivity (albeit that the impact factor is itself a central tendency statistic, see Leydesdorff & Bornmann, 2011). Lastly, normalized citation impact (i.e. the normalized mean citation rate or NMCR, see Glänzel, Thijs, Schubert, & Debackere, 2009) of the group's publications, defined as the ratio of citation counts of the papers of the group against the standards set by the specific subfields and therefore largely insensitive to the differences between the citation practices of the different science fields and subfields, is used as a proxy for the importance of the research. Whereas the group size and the average number of articles in top 20% journals are significantly correlated (Spearman's rho = 0.652), the NMCR is not correlated to the other variables (Spearman's rho = 0.093, and 0.193, for correlation between NMCR and group size, and the average number of articles in top 20% journals, respectively).

With regard to productivity, the panels agreed to score productivity taking into account (1) the number of PhD theses, (2) the number of scientific publications, (3) the national and international scientific activity of the research group, (4) the substance of the research funding attracted, and (5) the distribution of scientific activities within the research group. Thus, productivity is defined broader than productivity in terms of publications only. Again, not each of these aspects could be readily quantified. Hence we choose as variables the average number of PhDs per year defended, the average number of peer-reviewed journal articles per year, and the average amount of competitively acquired funding. Each of these variables is significantly correlated with the Spearman's rho values ranging from 0.607 to 0.694.

We present scatter plots to visualise the relationship between the scores quality and productivity on the one hand, and the variables discusses above on the other. All the data on research activity are highly skewed, with many instances of low values and a small number of instances of high values. Therefore, to allow visualization of the data, we have used a logarithmic x-axis (except of course in the case of the NMCR).

Results

Quality

Figure 1 presents the relationship of the group size, as expressed by the average research FTE per year with the quality assessment for each of the 52 research groups. Although the assessments pertain to different disciplines and have been carried out by six different panels, a strong correlation of quality with research group size emerges (Spearman's rho = 0.651). Indeed, groups consisting of over 10 research FTE receive a very good to excellent score (3.5 or more) in almost all cases. Moreover, all scores of 2.5 or lower have been received by (very) small groups. Nevertheless the quality of a number of small groups was assessed positively.

The number of observations per discipline is too small to allow firm conclusions, but visual inspection of Figure 1 suggests that for some disciplines (Chemistry, Pharmaceutical Sciences) critical mass is more important than others (especially Physics). This observation seems to be in line with the analysis of the British RAE data (Adams & Thomson, 2011; Kenna & Berche, 2011). However, a comparison of Figure 1 with the figures in reports on the British RAE data clearly demonstrate the difference in size between the meso level of departments and the micro level of research groups. Even though in our study all researchers are taken into account to determine the research FTE per group, most groups are very small compared to the size of the units of assessment included in the RAE. Still, the data show that at the micro level of the research group too, size and quality are related.

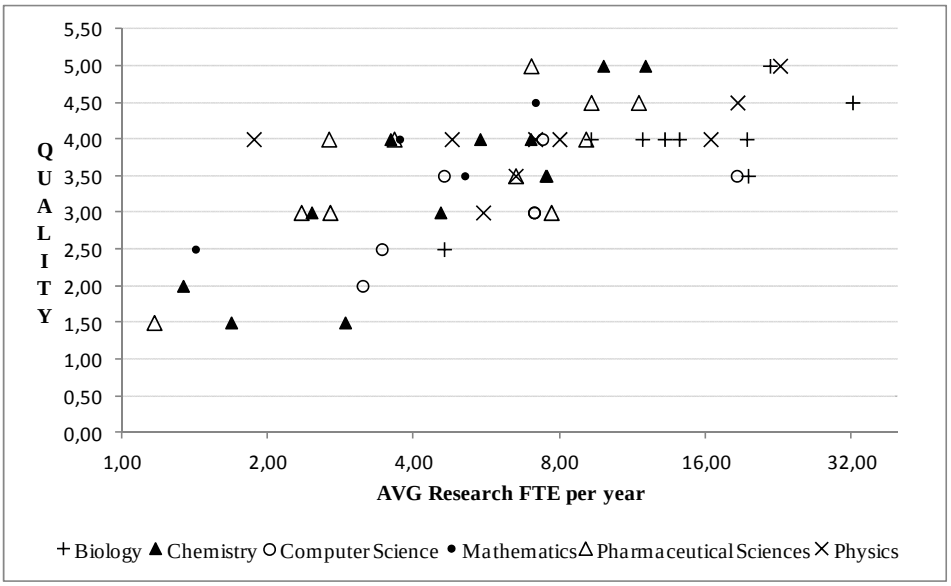

Figure 1: Average yearly research FTE in the research group vs. Quality (Spearman's rho = 0.651)

In Figure 2, the average number of articles in top 20% journals is plotted against the assessment of quality. Spearman's rho is 0.513, indicating a meaningful correlation between quality and the average amount of top 20% publications. In this case, each group is identified with the average amount of best quality publications they produce. The graph illustrates that groups with a lot of top publications score very good to excellent with respect to quality. The lack of publications in top journals, however, does not implicate a low quality score. In addition to the data shown on the graph, there are 2 groups (one in Mathematics and one in Computer Science) without publications in top 20% journals that are missing in the graph because of the logarithmic scale for the x-axis. These groups receive quality scores of 4 and 2 respectively, further illustrating the possibility of receiving a high score for quality without having papers published in top journals. Looking at the disciplinary distribution of the scores, it appears that especially in Computer Science the number of publications in top 20% journals is unrelated to quality, whereas in the disciplines of Biology, Pharmaceutical Sciences, and Physics a correlation is apparent. This also corresponds with the recommendations made by the panels, which did indeed stress the importance of publications in top outlets in the latter reviews.

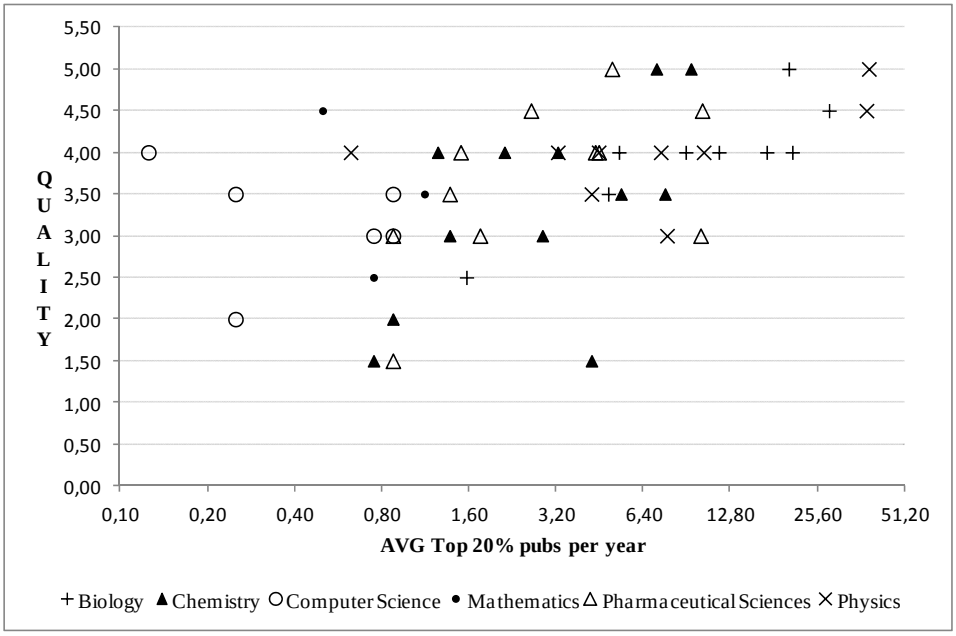

Figure 2: Average number of articles in top 20% journals per year vs. Quality (Spearman's rho = 0.513)

Figure 3 shows the relationship between the quality assessment and the normalized mean citation rate (NMCR) of the articles published by the group in the years t-8 to t-4 (a shorter time window is necessary in order to allow for accumulation of citations). The figure includes data for 38 research groups only, as for some of the groups, including the Computer Science groups, the number of publications to be included in the calculations of the NMCR was too small to allow for a reliable result (and hence was left out of the documentation).

Most groups have an NMCR that is bigger than 1, the world average. Although it is well known that citation indicators become less reliable at lower levels of aggregation, the absence of a significant correlation (Spearman's rho = 0.205) between the NMCR and the assessment of quality is remarkable. Others have reported meaningful correlations (Rons et al., 2008) – albeit to a varying degree depending on the research domain or the evaluation methodology (Aksnes & Taxt, 2004) – or at least the capacity of normalized citation indicators to distinguish between different quality classes of research groups (van Raan, 2006a). Close inspection of Figure 3 reveals that there are some groups with a NMCR that is below or on the world average, but still received a good quality assessment. High performance in terms of NMCR, i.e. NMCR > 1.6, does not appear as factor predictive of high quality (although the number of observations is low). Even more striking is the fact that the few really low scores for quality have been assigned to groups that had a good NMCR. These observations are in line with those made during the site visits, i.e. that the panel members pay little attention to normalized citation indicators and sometimes openly doubt their validity. Still, academics as well as administrators tend to take notice of (normalized) impact indicators, making them an indispensible element in the documentation presented to evaluation panels.

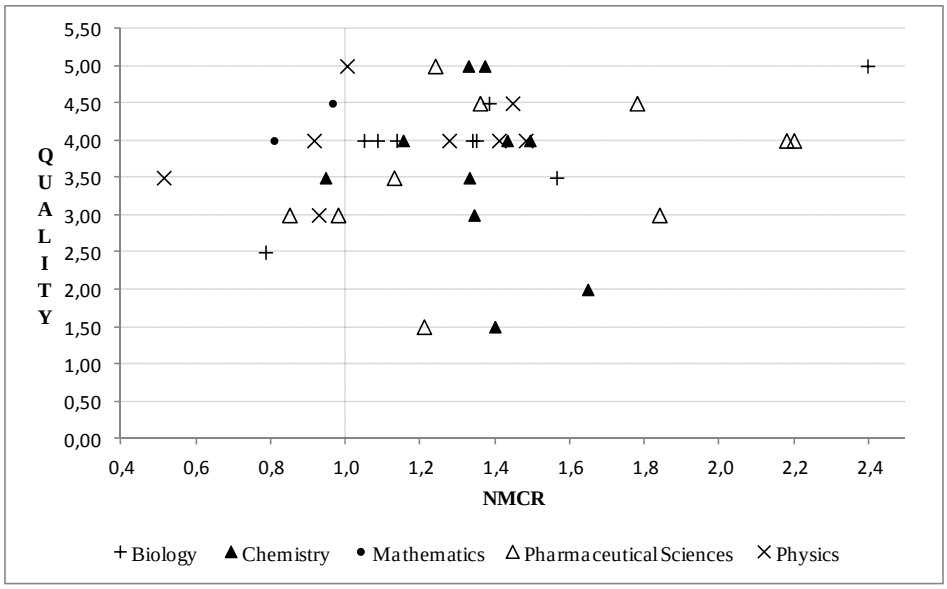

Figure 3: Normalized Mean Citation Rate vs. Quality
(Spearman's rho = 0.205)

Productivity
In each of the research assessments the panels also evaluated and scored the productivity of the research groups on a five point scale ranging from 1 (unsatisfactory) to 5 (excellent). Figures 4, 5, and 6 present the plots of the productivity assessments versus selected activity indicators.

Figure 4 plots the average number of peer reviewed articles per year versus the assessment of productivity. Spearman's rho is 0.619, indicating a strong correlation. The figure illustrates the difference in publication behaviour between Computer Science and Mathematics on the one hand, and the other disciplines on the other hand. The Computer Science and the Mathematics groups, although not necessarily smaller (cf. Figure 1), tend to publish much less than their counterparts in Biology and Physics. The scatter plot also shows nicely that the productivity score is more than just published material: groups that do not have that much output per year can still receive very good for productivity. Yet in order to receive an excellent score, a critical mass of about 20 papers per year seems indispensable.

In Figure 5 the average funding per year is plotted versus the assessment of productivity. Spearman's rho is 0.617, also indicating a strong correlation. It is obvious that the research groups that attract a high amount of funding mostly do score well for productivity. Research groups with a low amount of funding tend to receive a low productivity score. Yet a critical minimum is not apparent in Figure 5.

Lastly, Figure 6 plots the average number of PhDs defended against the assessment of productivity. Spearman's rho = 0.614. As a criterion of productivity, PhDs seems less self-evident than the previous two. Yet almost all panel members stressed the importance of the number of PhDs as an indicator of research intensity of a discipline and its research groups. Hence it is no surprise that the assessments of productivity do correlate well with the average number of PhDs defended per year. In fact, none of the groups with less than 1 PhD defence per year received a score higher than 'very good', whereas the groups that managed to produce at least 1,5 PhDs per year mostly received very high scores.

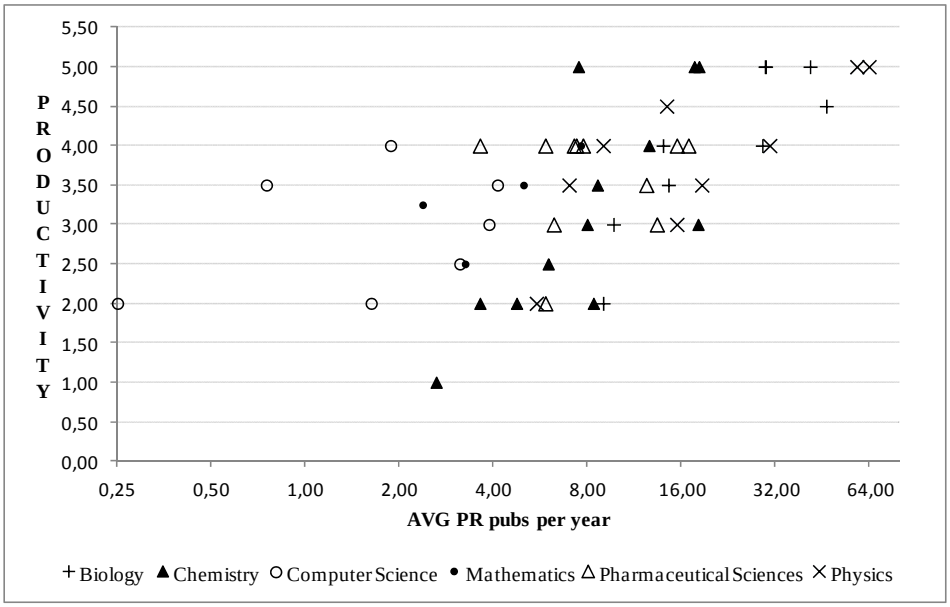

Figure 4: Average number of peer reviewed articles per year vs. Productivity (Spearman's rho = 0.619)

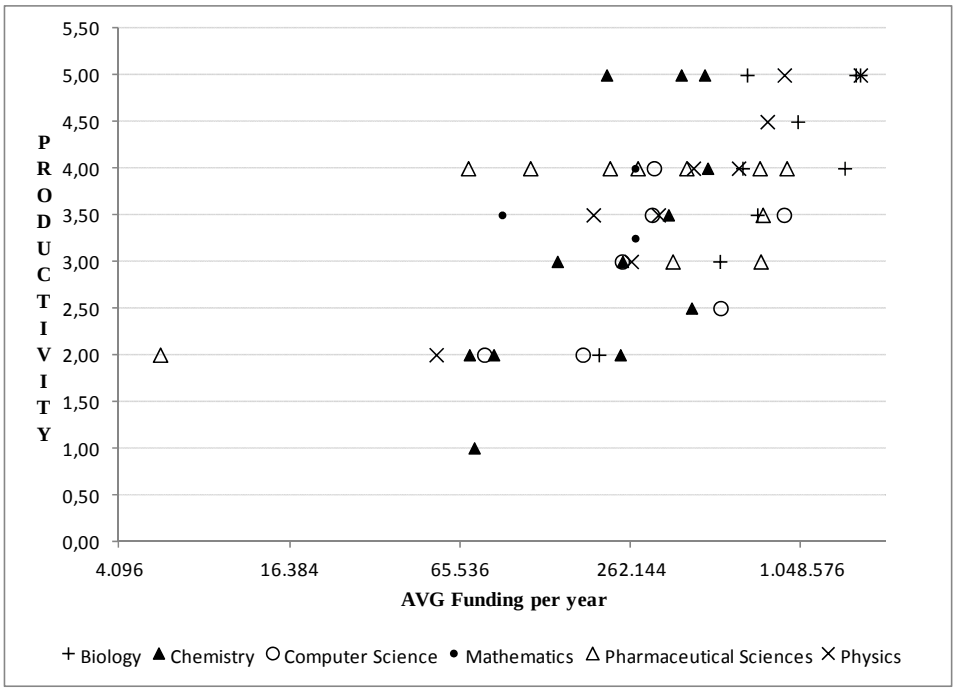

Figure 5: Average yearly amount of competitively acquired funding (in Euro) vs. Productivity (Spearman's rho = 0.617)

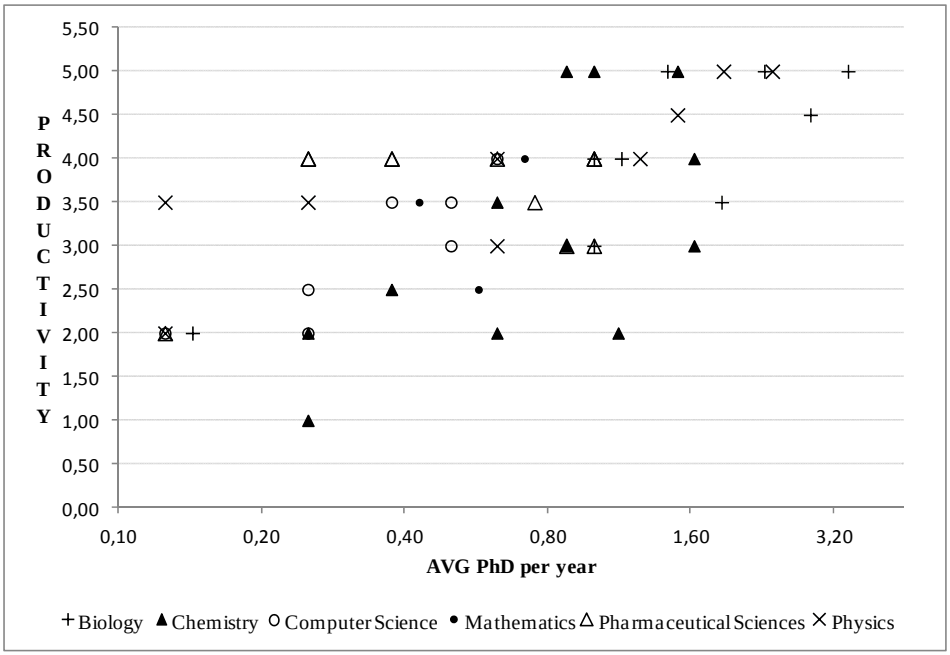

Figure 6: Average yearly number of PhDs defended under the supervision of a professor of the group vs. Productivity
(Spearman's rho = 0.614)

Discussion

For 52 science research groups in the sciences we present an analysis of the correlations of size, output numbers and impact indicators of the groups with the assessments of quality and productivity by independent international peer panels.

We find that the quality assessments are correlated with the size of the research groups and to a lesser extent with the number of publications in top 20% journals. Moreover, we observe no correlation of the quality assessments with the normalized mean citation rates. Although it is well known that citation indicators become less reliable at lower levels of aggregation, other researchers have reported correlations of quality assessments with normalized citation impact indicators (Rons et al., 2008; van Raan, 2006b). Two possible explanations for our surprising finding are the calculation of normalized impact as ratios of averages (vs. averages of ratios, see Larivière & Gingras, 2011) and/or the interdisciplinary orientation of some of the groups which may have resulted in inaccurate delimitation of the reference domain (Rons, 2012).

With regard to the productivity assessments we find consistently high correlations with each of the three selected output numbers, i.e. publications, funding and PhDs. This finding comes as no surprise, as panel members have consistently stressed the importance of international publishing, research income, and training of young researchers in their recommendations. Moreover, using RAE data, Butler and McAllister (2011), reported the importance of research income for research groups in chemistry, one of the core science disciplines.

Overall, our findings indicate the need for a certain critical mass at the level of research groups, i.e. the basic unit in which day-to-day research work takes place. This was also apparent from the panel members' comments, which regularly mentioned the need for clustering. Still, we should be careful in jumping to the conclusion that small groups are 'no good', as our data also show that some smaller groups, whether in terms of research FTE or output indicators, were assessed positively in terms of quality and/or productivity. In order to better understand our findings additional analysis taking into account the leadership and the (recent) growth of the research groups will be conducted.

Conclusion

The nature of research group quality and productivity is not well understood. In this paper we have presented data regarding 52 science research groups at the University of Antwerp. Each of the groups has been assessed in the frame of an on-going programme of research assessments by international peer panels that the university initiated in 2007. The results show that the assessments of quality are related to the size of the groups, indicating the need for critical mass in research groups. However, quality assessments correlated only weakly with publications in top journals and not with normalized citation impact. The assessments of productivity are correlated to the number of papers, the groups' research income, as well as the number of PhDs defended.

References

Adams, J. & Thomson, S. (2011). *Funding research excellence: research group size, critical mass & performance.* London: University Alliance.

Aksnes, D. W. & Taxt, R. E. (2004). Peer reviews and bibliometric indicators: a comparative study at a Norwegian university. *Research Evaluation, 13,* 33-41.

Butler, L. & McAllister, I. (2011). Evaluating university research performance metrics. *European Political Science, 10,* 44-58.

European Commission (2010). *Assessing Europe's university-based research.* Brussels: Expert Group on Assessment of University-Based Research, Directorate-General for Research, European Commission.

Glänzel, W. & Schubert, A. (2003). A new classification scheme of science fields and subfields designed for scientometric evaluation purposes. *Scientometrics, 56,* 357-367.

Glänzel, W., Thijs, B., Schubert, A., & Debackere, K. (2009). Subfield-specific normalized relative indicators and a new generation of relational charts: Methodological foundations illustrated on the assessment of institutional research performance. *Scientometrics, 78,* 165-188.

Hansen, H. F. (2009). *Research evaluation: Methods, practice, and experience.* Copenhagen: Danish Agency for Science, Technology and Innovation.

Hansson, F. (2010). Dialogue in or with peer review? Evaluating research organizations in order to promote organizational learning. *Science and Public Policy, 37,* 239-251.

Hicks, D. (2012). Performance-based university research funding systems. *Research Policy, 41,* 251-261.

Kenna, R. & Berche, B. (2011). Critical mass and the dependency of research quality on group size. *Scientometrics, 86,* 527-540.

Kretschmer, H. (1985). Cooperation structure, group size and productivity in research groups. *Scientometrics, 7,* 39-53.

Langfeldt, L. (2004). Expert panels evaluating research: Decision-making and sources of bias. *Research Evaluation, 13,* 51-62.
Larivière, V. & Gingras, Y. (2011). Averages of ratios vs. ratios of averages: An empirical analysis of four levels of aggregation. *Journal of Informetrics, 5,* 392-399.
Leydesdorff, L. & Bornmann, L. (2011). Integrated Impact Indicators *(I3)* compared with Impact Factors *(IFs)*: An alternative research design with policy implications. *Journal of the American Society for Information Science and Technology, 62,* 2133-2146.
Moed, H. F. (2007). The future of research evaluation rests with an intelligent combination of advanced metrics and transparent peer review. *Science and Public Policy, 34,* 575-584.
Richardson, M. (2011). Behind the data. Two's company: how scale affects research groups. *Research Trends,* 13-14.
Rons, N., De Bruyn, A., & Cornelis, J. (2008). Research evaluation per discipline: a peer-review method and its outcomes. *Research Evaluation, 17,* 45-57.
Rons, N. (2012). Partition-based field normalization: An approach to highly specialized publication records. *Journal of Informetrics, 6,* 1-10.
SCImago research Group (2011). *SCI World Report 2011* Granada: SCImago Research Group, University of Granada.
Spruyt, E. H. J. & Engels, T. C. E. (2010). Onderzoeksevaluaties als beleidsinstrument aan de Universiteit Antwerpen. *Thema: Tijdschrift voor Hoger Onderwijs en Management, 3,* 31-37.
van Raan, A. F. J. (2006a). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgements for 147 chemistry research groups. *Scientometrics, 67,* 491-502.
van Raan, A. F. J. (2006b). Statistical properties of bibliometric indicators: research group indicator distributions and correlations. *Journal of the American Society for Information Science and Technology, 57,* 408-430.
VSNU (2003). *Standard Evaluation Protocol 2003-2009 for public research organisations.* Utrecht/den Haag/Amsterdam: VSNU, NWO and KNAW.
VSNU (2009). *Standard Evaluation Protocol 2009-2015. Protocol for research assessment in the Netherlands.* Utrecht/den Haag/Amsterdam: VSNU, KNAW and NWO.

[i] The University of Antwerp (www.ua.ac.be), Flanders, Belgium, is a relatively young university that was founded in 2003 after the merger of three independent, already collaborating university institutions. The University of Antwerp has about 15000 students (including 1200 doctoral students). Research is carried out by 650 FTE academic staff members (tenured and tenure track academic staff as well as research and teaching assistants) and more than 1050 FTE researchers employed on externally funded research projects. The university comprises seven faculties: four in the social sciences and humanities (Arts & Philosophy, Political & Social Sciences, Law and Applied Economics) and three in the natural and biomedical sciences (Sciences, Medicine and Pharmaceutical, Biomedical & Veterinary Sciences). Each year, around 1200 students successfully complete a master degree program and 150 doctoral students obtain a PhD degree. Research at the university results in over 1500 publications included in the Web of Science.