Indicators for Research Performance in the Humanities? The Scholars' View on Research Quality and Indicators[1]

Michael Ochsner*, Sven E. Hug**

*ochsner@gess.ethz.ch
Department of Humanities, Social and Political Sciences, ETH Zurich, Muhlegasse 21, 8001 Zurich (Switzerland) and FORS, University of Lausanne, Géopolis, 1015 Lausanne (Switzerland)

** sven.hug@gess.ethz.ch
Department of Humanities, Social and Political Sciences, ETH Zurich, Muhlegasse 21, 8001 Zurich (Switzerland) and Evaluation Office, University of Zurich Muhlegasse 21, 8001 Zurich (Switzerland)

INTRODUCTION

In this paper, we present indicators for research quality in the humanities collected in our previous work (Ochsner, Hug, & Daniel, 2012). We focus on how these indicators are accepted by humanities scholars. We also investigate differences between different subgroups of the humanities scholars we surveyed with regard to their preferences for such indicators.

We address the theme of the conference ('peripheries frontiers and beyond') regarding two notions of (scientometric) periphery: First, we investigate indicators for research quality in the humanities, a field where there is a lack on knowledge on how to assess or even measure research quality, in fact, there is a quite broad consensus that (evaluative) bibliometrics cannot be readily applied in the humanities (Hicks, 2004; Nederhof, 2006). Second, we fully cover three humanities disciplines at Swiss universities and member universities of the League of European Research Universities (LERU). Scholars are a neglected stakeholder when it comes to the design of research assessment procedures or the selection of research indicators. However, they are directly affected, they know best what research quality in their field is and what impact the use of certain indicators could have on their research practices.

The paper is structured as follows: first, we present the background for selecting indicators for research quality. This is followed by a description of our analysis methods and the presentation of the results. We finally discuss the results with regard to their use in research evaluation.

INDICATORS FOR RESEARCH PERFORMANCE LINKED TO QUALITY CRITERIA

Humanities scholars have many objections against research evaluation, especially against quantification of research performance. This is at least partly due to the fact that there is a missing link between indicators for research performance and research quality (Ochsner et al., 2012). Scientometricians also note that research indicators are only loosely tied to quality definitions (Brooks, 2005; Donovan, 2008). Such weak or missing links between indicators

[1] This work was supported by the Rectors' Conference of the Swiss Universities (CRUS) within the framework of the SUK B-05 Innovation and Cooperation Project 'Mesurer les performances de la recherche' (Measuring Research Performance) as part of the cooperative initiative of the Universities of Zurich and Basel entitled 'Developing and Testing Research Quality Criteria in the Humanities, with an Emphasis on Literature Studies and Art History'. Matching funds for the initiative were provided by the University of Zurich.

and quality make it difficult for the assessed scholars to understand what is being measured. Therefore, the reluctance of humanities scholars to accept a quantitative representation of research quality is not surprising. At the same time, if the measurement is not or only loosely tied to the object that is to be measured (i.e. research quality), unintended effects become more likely. A sound measurement approach can replace the missing links between indicators and the concept(s). This means that before one can measure a concept with indicators, the concept needs to be clearly defined (Lazarsfeld & Barton, 1951, p. 155). Borsboom, Mellenberg, and van Heerden (2004, p. 1067) formulate the need of the definition of the concept in the following way: '[The issue is not] first to measure and then to find out what it is that is being measured but rather that the process must run the other way'. In a project on research quality and assessment in the humanities, we applied this approach by defining our concept ('research quality') by explicating quality criteria (Hug, Ochsner, & Daniel, 2014). In a next step, every quality criterion is specified and defined explicitly by one or more aspects (i.e. the analytical definition). Then, each aspect is operationalized by one or more indicators that specify how the aspect can be observed, quantified or measured (i.e. operational definition). Of course it is possible that for some aspects no indicators can be found, thus such an aspect cannot be measured by indicators.

Using Repertory Grid interviews (Ochsner, Hug, & Daniel, 2013) and a Delphi survey (Hug, Ochsner, & Daniel, 2013), we found 19 criteria for research quality in the humanities, specified by 70 aspects. We then identified aspects that reach a consensus among the humanities scholars. These aspects can be used to assess research quality in the three disciplines we studied (German and English literature studies and art history). In a next step, we collected indicators for research quality from the literature and directly from humanities scholars during the Repertory Grid interviews and the Delphi survey. This resulted in a long list of indicators, some very specific, some very vague. We grouped them into 62 indicator groups and linked them to the quality aspects they can potentially measure (for a complete list, see Ochsner, et al., 2012). Humanities scholars then rated the indicators according to their utility in measuring the corresponding quality aspects. In this paper, we will investigate differences in preferences for indicators between subgroups of our population.

METHOD

We designed a questionnaire to rate the indicator groups linked to quality aspects.[2] The scholars had to rate the indicator groups according to a statement on a 6-point scale (1 'strongly disagree' to 6 'strongly agree'). The statement consisted of two parts: A generic part ('The following quantitative statements provide peers with good indications of whether I...') followed by an aspect (e.g. 'I participate in a scholarly discourse regarding my field') of a criterion (e.g., 'scholarly exchange'). The scholars were presented the indicator groups that can potentially measure the given aspect and they had to rate each indicator group assigned to this aspect according to the statement. Because there were some discipline-specific aspects (i.e., aspects that reached consensus only in one or two disciplines), the questionnaires differed between the three disciplines. In German literature studies (GLS), the scholars had to rate 86 items consisting of 59 unique indicator groups assigned to 19 aspects (some indicator groups can be assigned to more than one aspect). In English literature studies (ELS), the respondents had to rate 85 items consisting of 45 unique indicator groups, and in art history

[2] For example, scholars were asked to indicate their (dis-)agreement with indicators as follows: The following quantitative statements provide peers with good indications of whether I participate in a scholarly discourse regarding my field: (a) number and weighting of publications for a disciplinary audience, (b) number of sources from my discipline I quote in my publications, (c) number, weightings and durations of editorships in my discipline etc.

(AH), the scholars had to rate 74 items consisting of 44 unique indicator groups assigned to 15 aspects. The questionnaire was administered in German and in English. Invitations were sent to all scholars holding at least a PhD working in one of the three disciplines at a Swiss university or at a member university of the League of European Research University (LERU). All in all 664 invitations were sent out. The field period lasted from October 2011 to January 2012.

We analyse the data using descriptive statistics such as means and medians to describe the acceptance of the indicators to measure the relevant quality aspects in the three disciplines. In this paper, we focus only on those indicators that have been rated in all disciplines (i.e. indicators measuring quality aspects reaching consensus in all disciplines) because the goal is to compare between different subgroups of the sample, including discipline. We also identify indicator groups that reach consensus. We define consensus the same way as we defined it concerning quality aspects, i.e. the median is above 4 (50% of the scholars rated the indicator group with a 5 at least) and the 10^{th} percentile is above 3 (not more than 10% of the scholars reject the indicator group). Because we did not use a random sample but a population survey, we cannot use inferential statistics. Therefore, we use bootstrap resampling (with 1000 replications) to estimate the stability of the results (95%-'stability intervals', see, e.g., Schneider & van Leeuwen, 2014) and standardized effect sizes to analyse differences in means across subgroups.

RESULTS

In total, 133 out of 664 questionnaires have been returned which corresponds to an overall response rate of twenty per cent. Among the respondents were 48 scholars of GLS, 43 scholars of ELS, and 42 scholars of AH (corresponding to response rates of 23%, 22%, and 17% respectively). Fifty-two respondents were members of Swiss universities and 81 respondents were members of LERU universities (corresponding to response rates of 33% and 16% respectively). Fifty-six women and 77 men (corresponding to response rates of 21% and 19% respectively) participated in the survey. Because the questionnaires differed between disciplines, an analysis of all indicator groups can only carried out by discipline (which does make sense as we are looking for indicators that adequately inform on quality criteria in a discipline). Most indicator groups were accepted by a majority of our respondents if analysed per discipline (acceptance being defined as a median higher than '4'). In GLS, 93% of the items reached that threshold, in ELS, 91% and in AH 97% respectively. However, also a minority that is not to be neglected clearly disagreed with many indicators: only 10 indicator groups (12%) reached consensus in GLS, one indicator group (1%) in ELS, and 16 indicator groups (22%) in AH. For a more information on the results of the whole questionnaire, see Ochsner, Hug, & Daniel, 2014).

In this paper, we focus on those 39 items that have been part of all three questionnaires in order to investigate whether there are differences between different subgroups of our sample. The 39 items consist of 34 unique indicator groups assigned to 8 quality aspects specifying 7 quality criteria. Of the 39 items, three indicator groups (8%) were rejected by a majority. However, only two indicator groups (5%) reached consensus over all respondents (see table 1). If we look at mean differences between disciplines, we see that most differences are small to moderate (*Cohen's d*<0.8). However, we find also that ELS scholars rated the indicators quite lower than GLS and AH scholars (8 items with a *Cohen's d*>0.5 in ELS vs. GLS, 18 in ELS vs. AH) and AH scholars rated some items higher than GLS scholars (4 items with a *Cohen's d*>0.5). Regarding gender, there are no big differences in means, only 8 items exhibit

a *Cohen's d* between 0.2 and 0.3, which can be considered small. We find some differences, however, between tenured and non-tenured scholars: tenured are more in favour of the indicator group 'initiation/foundation' (number and weighting of what the person has initiated or founded, e.g. book series, institutions, journals etc.), no matter whether

Table 1: Overall Mean, percentage of negative ratings (bootstrapped 95% stability intervals in parentheses), and Cohen's d of subgroups for indicator groups.

Variable	Mean	% of negative ratings	Cohen's d GLS vs ELS	Cohen's d AH vs GLS	Cohen's d ELS vs AH	Cohen's d Gender	Cohen's d Tenure
Publications: disciplinary exchange	4.95 (4.77-5.14)	0.06 (0.02-0.10)	0.22	0.27	0.27	-0.04	0.02
References: disciplinary exchange	3.86 (3.62-4.10)	0.30 (0.22-0.38)	-0.15	0.65	0.65	0.02	0.09
Presentations: disciplinary exchange	4.55 (4.35-4.75)	0.14 (0.08-0.19)	0.24	0.00	0.00	0.11	0.04
Editorship: disciplinary exchange	4.35 (4.13-4.56)	0.21 (0.14-0.28)	0.26	-0.33	-0.33	-0.10	0.16
Organized events disciplinary exchange	4.44 (4.24-4.65)	0.17 (0.10-0.23)	0.51	0.02	0.02	0.03	-0.16
Collaborations: disciplinary exchange	4.35 (4.13-4.58)	0.19 (0.12-0.25)	0.83	-0.06	-0.06	-0.04	-0.12
Personal contacts: disciplinary exchange	3.92 (3.68-4.17)	0.30 (0.22-0.38)	0.25	0.05	0.05	0.29	0.15
Review Activities: disciplinary exchange	4.20 (3.99-4.41)	0.20 (0.13-0.27)	0.20	0.12	0.12	-0.06	0.37
Academic associations: disciplinary exchange	3.98 (3.78-4.18)	0.28 (0.20-0.36)	0.09	0.04	0.04	0.27	0.14
Panels: disciplinary exchange	4.14 (3.94-4.35)	0.23 (0.15-0.30)	0.22	0.07	0.07	0.10	0.22
Survey: renewal of interpretations of the past	3.81 (3.59-4.04)	0.35 (0.27-0.43)	0.19	0.85	0.85	-0.12	-0.05
Citations: impact on research community	3.74 (3.49-3.98)	0.35 (0.27-0.43)	-0.05	0.05	0.05	0.08	-0.01
Acknowledgements: impact on research community	3.20 (2.94-3.45)	0.54 (0.46-0.62)	-0.42	0.19	0.19	0.15	0.24
Success of junior researchers: impact on research community	4.27 (4.07-4.48)	0.21 (0.14-0.28)	0.15	0.21	0.21	0.07	0.27
Started initiatives: impact on research community	4.02 (3.78-4.25)	0.30 (0.22-0.38)	-0.09	0.07	0.07	0.09	0.54
Editorship: impact on research community	4.04 (3.81-4.26)	0.25 (0.17-0.32)	0.46	-0.24	-0.24	0.00	0.25
Opportunities for junior researchers: openness to other persons	4.47 (4.25-4.69)	0.18 (0.12-0.25)	0.16	0.40	0.40	-0.20	-0.02
Assessed oppenness: openness to other persons	4.81 (4.62-5.00)	0.08 (0.04-0.13)	0.27	0.21	0.21	-0.15	-0.03
Heterogeneity of junior researchers: openness to other persons	4.20 (3.97-4.42)	0.26 (0.18-0.33)	0.20	0.41	0.41	-0.07	-0.33

Table 1: continued

Variable	Mean	% of negative ratings	Cohen's d GLS vs ELS	Cohen's d AH vs GLS	Cohen's d ELS vs AH	Cohen's d Gender	Cohen's d Tenure
Assistance: openness to other persons	3.62 (3.39-3.85)	0.41 (0.32-0.49)	0.19	0.23	0.23	0.24	-0.15
Course accessibility: openness to other persons	4.23 (4.00-4.47)	0.27 (0.19-0.35)	0.52	0.14	0.14	0.11	-0.34
Availability of publications: openness to other persons	3.95 (3.71-4.20)	0.35 (0.27-0.42)	0.06	0.39	0.39	0.11	-0.11
Sources: rich experience with sources	4.67 (4.43-4.90)	0.19 (0.12-0.25)	0.11	0.23	0.23	0.33	0.18
Documentation activities: rich experience with sources	3.98 (3.74-4.23)	0.33 (0.25-0.41)	0.54	0.24	0.24	-0.05	-0.08
Output of documentation activities: rich experience with sources	4.14 (3.91-4.36)	0.27 (0.20-0.34)	0.54	0.21	0.21	-0.06	-0.09
Research time: knowledge based on own research	4.32 (4.04-4.59)	0.25 (0.18-0.32)	0.40	0.02	0.02	-0.09	-0.14
Personal library: knowledge based on own research	3.14 (2.85-3.43)	0.57 (0.48-0.66)	-0.19	0.52	0.52	0.02	0.25
References: knowledge based on own research	3.81 (3.56-4.07)	0.38 (0.30-0.46)	0.22	0.29	0.29	0.22	0.03
Monographs: knowledge based on own research	4.26 (4.04-4.49)	0.22 (0.15-0.29)	0.13	0.25	0.25	-0.02	0.11
Attractivity to junior researchers: arouse passion for research	4.59 (4.41-4.78)	0.13 (0.07-0.18)	0.04	0.21	0.21	-0.06	0.00
Qualification of junior researchers: arouse passion for research	4.53 (4.33-4.72)	0.17 (0.10-0.23)	0.21	0.33	0.33	0.04	0.06
Success of junior researchers: arouse passion for research	4.35 (4.13-4.58)	0.23 (0.16-0.31)	0.20	0.24	0.24	-0.04	0.07
Teaching awards: arouse passion for research	3.66 (3.44-3.88)	0.40 (0.31-0.48)	0.20	-0.23	-0.23	0.14	0.12
Acknowledgments: arouse passion for research	3.19 (2.96-3.42)	0.56 (0.47-0.64)	-0.32	0.18	0.18	0.27	0.25
Survey enthusiasm teaching: arouse passion for research	4.15 (3.94-4.36)	0.26 (0.18-0.33)	0.30	0.01	0.01	0.21	-0.27
Survey enthusiasm public: arouse passion for research	3.38 (3.15-3.62)	0.50 (0.41-0.58)	0.40	0.27	0.27	-0.16	-0.18
Started initiatives: pointing out important research for the future	3.84 (3.61-4.08)	0.32 (0.24-0.40)	-0.31	0.07	0.07	0.17	0.45
Strategies: pointing out important research for the future	4.00 (3.77-4.23)	0.34 (0.26-0.42)	0.28	0.12	0.12	-0.04	0.13
Utilizing sources: pointing out important research for the future	4.03 (3.81-4.25)	0.29 (0.21-0.36)	0.59	0.18	0.18	-0.17	-0.19

it measures the quality aspect 'impact on research community' or 'vision of the future'. It is striking that among the items with a *Cohen's d* above 0.2, those which presuppose a strong network (reviews, board memberships, acknowledgements, success of young scholars,

initiations/foundations, editorships, personal library) are rated more favourably by tenured scholars while those which refer to teaching, collaboration and social competence (heterogeneity of students and staff, openness and accessibility to courses, survey of students and junior researchers on arousing passion for the subject) are rated more favourably by non-tenured scholars.

To be effective, indicators applied in research evaluation need to be accepted by those who are directly affected by them (i.e. the scholars). Therefore, indicators with a high degree of acceptance, i.e. a consensus, need to be identified. In our definition, an indicator reaches consensus if less than 10% of the scholars rate it negatively (i.e., with 1, 2 or 3) and if, at the same time, 50% or more rate it with 5 or 6. Only two out of the 39 items reach consensus over all respondents (*publications* measuring 'disciplinary exchange' and *assessed openness* measuring 'openness to other persons'). With regard to consensus, we find some differences between disciplines. While in AH 8 items and in GLS 4 items reach consensus, none of the items reaches consensus in ELS. Only two items reach consensus in two disciplines: not surprisingly the ones that reach consensus over all respondents. These are also the two items attracting the smallest proportion of ELS scholars rating them negatively (16% for both items). If we look at the stability intervals derived from bootstrap resampling, however, no item has a stability interval that does not include values above 10% (of persons choosing a negative rating), one item just scratching the threshold with the upper level of the stability interval being only slightly higher than 10% (*publications* measuring 'disciplinary exchange'). With regard to disciplines, two items have stability intervals that do not exceed the 10% threshold in AH (in fact, for these two items, not a single respondent in AH chose a negative value: *publications* measuring 'disciplinary exchange' and *assessed openness* measuring 'openness to other persons'), in GLS one item shows a stable consensus (*publications* measuring 'disciplinary exchange') and in ELS none.

If we compare the ratings by gender, we find only one item that reaches consensus among men and women (*publications* measuring 'disciplinary exchange') and two more reaching consensus among men (*assessed openness* measuring 'openness to other persons' and *attractivity to young researchers* measuring 'arouse passion for research'). However, no item reaches a stable consensus, yet two items that only reach consensus among men miss the 10% threshold only slightly, thus could be considered as stable (10.05% and 10.3% resp., again the *publications* and *assessed openness* we know from the other comparisons). Similarly, there are not many differences between tenured and non-tenured scholars: the *publications* measuring 'disciplinary exchange' reach consensus among both tenured and non-tenured scholars (the upper limits of the stability intervals being 13% for tenured and 10.3% for non-tenured, i.e. quite stable consensus). All other items do not reach consensus except *assessed openness* measuring 'openness to other person' that reaches consensus only among tenured scholars (the upper level of the stability interval being 11%).

CONCLUSION

This paper examines the scholars' acceptance of indicators to measure research quality in the humanities. In our previous research, we already found that humanities scholars are open regarding research evaluation using quality criteria that relate to their own notions of quality but are reluctant to accept indicators or a quantitative approach towards research evaluation (Ochsner et al., 2014). We also found that there is a mismatch of quality criteria between research evaluators and humanities scholars (Hug et al., 2013). In this paper, we investigated whether there are differences in preferences for research indicators among subgroups of

humanities scholars. While we found small differences between the three disciplines we studied, we did find only few differences between other subgroups (gender and tenure). Our analysis shows that while most indicators would be accepted by at least 50% of our respondents[3] in an informed peer review process, almost all indicators face rejection from a large minority (regarding the indicators in this study on average 28%). The only indicators that reached consensus in all three disciplines were *publications* measuring 'disciplinary exchange' and *assessed openness* measuring 'openness to other persons'. From this, it follows that a purely quantitative approach to research evaluation is rejected by a vast majority of our respondents.

We can conclude that the use of quantitative information in the evaluation of humanities research is possible if some restrictions are considered. The indicators must be linked to the humanities scholars' quality notions. They also have to be accepted by the scholars in order not to interfere in a destructive way with their research practices. While many scholars agree to some indicators measuring certain quality criteria during an informed peer review process, our results suggest that the use of the indicators should be agreed upon with the scholars. We found that irrespective of gender, tenure, and discipline, a fairly large part of the scholars oppose most indicators. However, we have chosen three 'aesthetic' disciplines considered especially difficult to evaluate in a quantitative way. It is likely that in more 'empirical' disciplines, the use of a broader set of indicators will be accepted. Nevertheless, research evaluation in the humanities, especially quantitative measurements, should be discursive as well as participatory and should focus on research quality or at least include it to an important degree.

REFERENCES

Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The concept of validity. *Psychological Review, 111*(4), 1061–1071. doi:10.1037/0033-295X.111.4.1061.

Brooks, R. L. (2005). Measuring university quality. *The Review of Higher Education, 29*(1), 1–21. doi:10.1353/rhe.2005.0061.

Donovan, C. (2007). The qualitative future of research evaluation. *Science and Public Policy, 34*(8), 585–597. doi:10.3152/030234207X256538.

Hicks, D. (2004). The four literatures of social science. In H. F. Moed, W. Glänzel, & U. Schmoch(Eds.), *Handbook of quantitative science and technology research: The use of publication and patent statistics in studies of S&T systems* (pp. 476–496). Dordrecht: Kluwer Academic Publishers.

Hug, S. E., Ochsner, M., & Daniel, H.-D. (2013). Criteria for assessing research quality in the humanities: A Delphi study among scholars of English literature, German literature and art history. *Research Evaluation, 22*(5), 369–383. doi:10.1093/reseval/rvt008.

Hug, S. E., Ochsner, M., & Daniel, H.-D. (2014). A framework to explore and develop criteria for assessing research quality in the humanities. *International Journal for Education Law and Policy, 10*(1), 55–64.

Lazarsfeld, P. F., & Barton, A. H. (1951). Qualitative Measurement in the Social Sciences. Classif1cation, Typologies, and Indices, in D. Lerner & H. D. Lasswell (Eds.), *The Policy Sciences* (pp. 155–192). Stanford, CA: Stanford University Press.

[3] From our previous studies with the same panel, we assume that there might be a selection bias regarding respondents being more open to indicators than non-respondents.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: a review. *Scientometrics, 66*(1), 81–100. doi:10.1007/s11192-006-0007-2.

Ochsner, M., Hug, S. E., & Daniel, H.-D. (2012). Indicators for research quality in the humanities: opportunities and limitations. *Bibliometrie – Praxis und Forschung, 1*(4). Retrieved from http://www.bibliometrie-pf.de/article/view/157/192.

Ochsner, M., Hug, S. E., & Daniel, H.-D. (2013). Four types of research in the humanities: Setting the stage for research quality criteria in the humanities. *Research Evaluation, 22*(4), 79–92. doi:10.1093/reseval/rvs039.

Ochsner, M., Hug, S. E., & Daniel, H.-D. (2014). Setting the stage for the assessment of research quality in the humanities: Consolidating the results of four empirical studies. *Zeitschrift für Erz ehungswissenschaft, 17*(6 Supplement), 111–132. doi:10.1007/s11618-014-0576-4.

Schneider, J. W., & van Leeuwen, T. N. (2014). Analysing robustness and uncertainty levels of bibliometric performance statistics supporting science policy. A case study evaluating Danish postdoctoral funding. *Research Evaluation, 23*(4), 285–297. doi: 10.1093/reseval/rvu016.

21ST international conference on science and technology indicators

STI Conference 2016 · València

Peripheries, frontiers and beyond

14 ·16 September 2016

Universitat Politècnica de València

BOOK OF PROCEEDINGS

www.sti2016.org