Assessment Criteria for Early Career Researcher's Proposals in the Humanities

Michael Ochsner[1] Sven E. Hug[2] Hans-Dieter Daniel[3]

[1] *ochsner@gess.ethz.ch*
ETH Zurich, Zurich (Switzerland)
FORS, Lausanne (Switzerland)

[2] *sven.hug@gess.ethz.ch*
ETH Zurich, Zurich (Switzerland)
University of Zurich, Zurich (Switzerland)

[3] *daniel@gess.ethz.ch*
ETH Zurich, Zurich (Switzerland)
University of Zurich, Zurich (Switzerland)

Abstract
Competitive research grants become more and more important in the careers of young scholars. If grants are making careers, the decision for the grant winners is important and needs to be fair, consistent and transparent. In this research in progress paper, we present evaluation criteria for research proposals from early career researchers in the humanities. We apply a bottom-up procedure to identify evaluation criteria that reach consensus among the humanities scholars themselves. We identified 23 aspects pertaining to 9 criteria for the assessment of research proposals. There are no differences between the selection of aspects that reach consensus among the scholars regarding whether the applicant is a doctoral student or a postdoc, nor did we find differences in the selection of aspects between disciplines. We found slight differences in the ratings between tenured and non-tenured scholars and between women and men. Tenured scholars and women each emphasized an additional aspect.

Conference Topic
Science Policy and Research Evaluation

Introduction
Competitive research grants become more and more important in the careers of young scholars (van Arensbergen, van der Weijden, & van den Besselaar, 2014a, b). The acquisition of such grants is seen as a sign of quality of scholarship by senior researchers (see e.g. van Arensbergen *et al.*, 2014a, b) as well as in evaluation procedures (see e.g. Ochsner et al., 2012). This process is closely linked to the shift to the notion of excellence in higher education policy. If a higher education institution adheres to the notion of excellence, it has to recruit excellent scholars. Therefore, governments and universities focus on talent selection processes and increase the support for early career researchers. Research policy implements 'excellence' amongst others through competitive research funding and temporary positions for early and mid-career researchers (van Arensbergen et al., 2014b; van den Akker, 2016). While competitive funding first concerned Science, Technology, Engineering and Medicine (STEM) disciplines due their need for expensive infrastructure and large teams (Krull and Tepperwien, 2016), in the humanities the acquisition of competitive grants has not been very important in the past. However, the change to a focus on talent selection and temporary employment applies also to the humanities. Therefore, the assessment and selection of research proposals become more and more important in the humanities as well.
While there are studies on peer and panel review, they focus mainly on selection biases and fairness (see e.g. Bornmann, Mutz & Daniel, 2008; Bornmann, Mutz, Marx, Schier &

Daniel, 2011; Lamont, 2009). There is a lack of knowledge about what quality of research proposals means and how it can be identified, especially so in the humanities (see e.g. Hemlin, 1993). Little is known about what criteria peers have in mind when evaluating a proposal and even less how they weight these criteria. However, judging a work without criteria is inconsistent and not adequate for judging merit, as Thorngate, Dawes and Foddy (2009) conclude their comprehensive research on decision making. They found that judging separately according to specified criteria reveals more consistent results (Thorngate *et al.*, 2009, p. 26). Such *intra*-rater reliability (in distinction of the *inter*-rater reliability) is of utter importance when judging merit (i.e. a reviewer would give the same rating for application A when rating application A before application B or after it; or give the same rating at a later point in time). It is not only crucial to have reliable judgments as a basis in the review process. When deciding for future careers, it is also important that the criteria are explicit and clear so that young scholars do know what to deliver and, in case of a negative evaluation, how to improve. Furthermore, explicit criteria serve transparency. All these points are important for the judgment of merit to be fair and consistent (Thorngate *et al.*, 2009). The growing importance of research grants for the further career of young scholars makes it particularly important that the best applicants are awarded the grant. Therefore, an adequate procedure for selecting the best proposals must be applied.

In this research in progress paper, we present quality criteria for the ex-ante assessment of research proposals from early career researchers in the humanities. Applying a bottom-up approach we base the evaluation criteria on scholars' ratings of quality criteria regarding their adequacy for the use in such an assessment situation. Particularly, we are investigating the following research questions: a) are there differences between the criteria for evaluating the proposals from PhD students and those for evaluating proposals from postdocs? b) is there a common set of quality criteria across disciplines that can be used to adequately judge research proposals? c) do tenured professors emphasize other criteria than the young scholars themselves? d) are there gender differences regarding the preferences for criteria?

This paper is organized in the following way: First, we present the approach and methods used to identify the important quality criteria for a specific evaluation situation. We then present the methods with which we investigate differences in the preferences for quality criteria between sub-groups of our sample: level of the grant (PhD or postdoc), discipline, academic status and gender. We finally conclude regarding the differences between sub-groups of the sample reflecting the generalizability of our results.

Methods and data used

For the selection of the criteria to be included in the evaluation sheet, we designed a questionnaire containing nine criteria, specified by 23 aspects, for judging the quality and potential of the research proposal. We draw from our previous research in which we developed a catalogue of criteria for research quality in the humanities in a strictly bottom-up procedure, i.e. the scholars formulated their own criteria during several steps and using different methods (Ochsner, Hug & Daniel, 2014). Based on this catalogue, we selected, adapted and expanded the criteria to the evaluation situation of ex-ante research proposal assessment of early career researchers and added some criteria usually used in such evaluation situations (i.e., information about the applicant). The questionnaire was sent out to all Swiss scholars holding a doctoral degree in the humanities (theology/religious studies were excluded due to another project fielding a similar survey at the same time in these disciplines). The scholars were to rate the criteria for their suitability for the evaluation of research proposals. To do so, they had to give their agreement on a 6-point scale with a statement that consisted of a generic part and a part specific to the aspect that is rated. The generic part read: 'A project application is assessed appropriately, if the assessment considers whether...' while the

specific part read for example ‘the project identifies gaps in existing knowledge’ (criterion *Originality*, aspect *gaps in knowledge*). The scale was labelled in the following way: 1 means ‘I strongly disagree with the statement’, 2: ‘I disagree’, 3: ‘I slightly disagree’, 4: ‘I slightly agree’, 5: ‘I agree’ and 6: ‘I strongly agree with the statement’.
Using the ratings, we identified criteria and aspects that reach consensus among the scholars. An aspect reaches consensus when it is clearly approved by the majority (at least 50% of the scholars rate the aspect with at least a ‘5’) and only a small majority disapproves the aspect (less than 10% rate the aspect negatively, that is with a ‘1’, ‘2’ or ‘3’).
To answer our research questions, we identified differences in the ratings of the aspects between different sub-groups of our sample using standardized effect sizes, i.e. Cohen’s d (Cohen, 1988). We applied a threshold of *Cohen’s d*=0.2 as suggested by Cohen (1988), who proposes the rule of thumb that $d>0.2$ equals a small effect, $d>0.5$ a medium effect, and $d>0.8$ a large effect. Because our sample is a full population survey (i.e., all humanities scholars in Switzerland were invited to participate) and not a random sample, we cannot use inferential statistics but use bootstrap resampling with 1,000 replications to estimate the stability of the results instead, using bootstrapped 95% stability intervals. This serves to account for the effect of possible outliers and as a measure of stability because not all humanities scholars in the population respond to the survey (see Schneider & van Leeuwen, 2014).

Results

The questionnaire was sent out to 2,609 humanities scholars of whom 916 filled in the questionnaire. This amounts to an overall response rate of 35%, which is a very good response rate compared to similar survey projects (see e.g. Braun & Ganser, 2011; Cardoso, Rosa & Santos, 2013; Giménez-Toledo & Román-Román, 2013). The response rate differed between disciplines in the sense that scholars from law studies participated to a quiet smaller degree: While the response rate in language and linguistics amounted to 38% and in history and cultural studies to 41%, the response rate among law scholars was at 24%. This can be explained by the fact that in law studies, our sample contained a significant number of persons that are primarily active as lawyers or judges and teach irregularly at the university. Those scholars did not participate (many of them sent us an email with the excuse that they felt it was not appropriate for them to answer the questionnaire due to a certain distance from academic life).
From the 23 aspects assigned to 9 criteria that the scholars rated, 13 aspects pertaining to 6 criteria reached consensus (see table 1). All aspects that reached the threshold of not more than 10% negative ratings also reached a median of 5. Therefore, we only list the results for the percentage of negative ratings as this was the decisive criterion. As the bootstrapped stability intervals show, the results are quite stable.
The important criteria for the evaluation of research proposals of young humanities scholars are their originality, feasibility, rigour, relevance, complexity and variety. Originality is defined by the aspects *identifying gaps* in existing knowledge, using *innovative data*, presenting *new findings*. Feasibility is defined by a realistic *timetable* and *resources*. Rigour is defined by the aspects appropriate *research process*, expression of the *state of research* and *choice of method* as well as a *stringent argumentation* and *understandable* language. Relevance is defined as *academic relevance*. Complexity is defined as *making complexity visible*. Variety is defined as contribution to the *variety of research*.
Looking at the Cohen’s d, we find little differences in general. Almost all coefficients are below the threshold of *d*=0.2. Between the assessment of proposals from doctoral students and postdocs, only two aspects reach a $d>0.2$: *independence* (*d*=0.20) and the *applicant’s publication list* (*d*=0.30) are rated less favourably for the assessment of proposals from doctoral students than for those from postdocs. However, for both groups, the two aspects do

not reach consensus. Therefore, the selection of criteria for the assessment of proposals from doctoral students is the same as for those from postdocs.

Table 1. Overall mean, percentage of negative ratings (with bootstrapped 95% stability intervals in parentheses), and Cohen's d of subgroups for the ratings of the aspects

Aspect	Mean	% of neg. ratings	Cohen's d Doc vs Postdoc	Cohen's d Lang. vs. HistCult	Cohen's d Law. vs. Lang.	Cohen's d HistCult vs. Law	Cohen's d Tenure	Cohen's d Gender
Independence	4.77 (4.69-4.85)	0.16 (0.14-0.19)	-0.20	-0.20	0.08	0.12	-0.08	0.10
Originality: Identify Gaps	5.31 (5.25-5.36)	0.03 (0.02-0.04)	-0.07	0.10	-0.18	0.07	-0.07	0.15
Originality: Innovative Data	5.18 (5.12-5.23)	0.04 (0.03-0.05)	-0.01	-0.10	-0.12	0.23	0.11	0.13
Originality: New Research Topic	4.75 (4.68-4.83)	0.12 (0.10-0.14)	0.02	0.00	0.10	-0.10	0.07	-0.01
Originality: New Approach	4.80 (4.72-4.87)	0.13 (0.10-0.15)	0.02	0.04	-0.29	0.25	-0.03	0.13
Originality: New Paradigm	4.53 (4.45-4.62)	0.19 (0.16-0.21)	0.01	0.05	0.08	-0.13	-0.05	-0.02
Originality: New Finding	4.99 (4.93-5.05)	0.08 (0.06-0.09)	0.05	0.07	0.06	-0.13	-0.19	0.16
Feasibility: Timetable	5.03 (4.96-5.09)	0.07 (0.05-0.08)	-0.03	-0.05	-0.04	0.08	0.19	0.09
Feasibility: Resources	5.10 (5.04-5.16)	0.05 (0.03-0.06)	-0.01	-0.05	0.02	0.03	0.16	0.02
Rigour: Research Process	5.31 (5.26-5.36)	0.03 (0.02-0.04)	-0.06	-0.07	0.04	0.02	0.29	0.09
Rigour: State of Research	5.29 (5.23-5.35)	0.03 (0.02-0.05)	0.01	-0.05	-0.17	0.23	0.20	0.13
Rigour: Choice of Method	4.98 (4.92-5.04)	0.06 (0.04-0.07)	0.03	-0.09	0.05	0.04	0.20	0.15
Rigour: Argumentation	5.54 (5.49-5.59)	0.02 (0.01-0.03)	-0.03	-0.03	0.02	0.01	0.23	0.16
Rigour: Understandable	5.51 (5.46-5.55)	0.02 (0.01-0.02)	-0.01	-0.15	0.18	-0.04	0.16	0.14
Relevance: Academia	5.06 (4.99-5.13)	0.07 (0.06-0.09)	0.01	-0.11	0.05	0.06	0.19	-0.08
Relevance: Societal	3.81 (3.72-3.90)	0.36 (0.33-0.39)	-0.02	-0.22	0.72	-0.50	-0.13	0.08
Cultural Heritage	4.46 (4.38-4.55)	0.19 (0.17-0.22)	-0.03	-0.06	-0.60	0.65	-0.12	0.02
Complexity	4.98 (4.92-5.05)	0.08 (0.06-0.10)	-0.15	-0.12	-0.06	0.18	0.01	0.14
Variety	4.96 (4.89-5.02)	0.08 (0.06-0.10)	-0.11	-0.05	-0.05	0.10	-0.11	0.20
Person: CV	4.79 (4.71-4.86)	0.11 (0.09-0.13)	-0.16	-0.10	0.11	-0.02	0.34	0.15
Person: Diploma	4.58 (4.51-4.65)	0.12 (0.10-0.14)	-0.12	0.02	0.14	-0.17	0.23	0.21
Person: Publications	4.52 (4.45-4.60)	0.16 (0.14-0.19)	-0.30	-0.01	0.02	-0.01	0.39	0.02
Person: Recommendations	3.90 (3.81-3.98)	0.29 (0.26-0.32)	0.02	0.07	-0.06	-0.01	0.10	0.19

Note. Doc=proposals from doctoral students; postdoc=proposals from postdocs; Lang.=language and literature; HistCult=history and cultural studies; Law=law studies

Regarding disciplines, we find six aspects reaching a *Cohen's d*>0.2 (i.e. *independence*, *innovative data*, *new approach*, *state of research*, *cultural heritage* and *societal relevance*), but only two aspects, for which ratings are different to a greater degree: the criterion *cultural heritage* is rated much higher in the disciplines language and literature as well as history and cultural studies than in law studies (*d*=0.6 and *d*=0.65), while *societal relevance* is rated much higher in law studies (*d*=0.7 and *d*=0.5). But again, the aspects do not differ regarding consensus. Thus, while we find disciplinary differences regarding the ratings of some aspects, the selection of aspects that reach consensus does not change between disciplines.
We also find differences between the ratings from tenured and non-tenured scholars. From the criterion rigour, tenured scholars rate all aspects bar one higher than non-tenured scholars: *research process* (*d*=0.29), *state of research* (*d*=0.20), *choice of method* (*d*=0.20) and *argumentation* (*d*=0.23). Furthermore, tenured scholars rate the *applicant's CV* (*d*=0.34), his or her *diploma* (*d*=0.23) and his or her *publication list* (*d*=0.39) higher than non-tenured scholars. The applicant's CV reaches consensus among tenured scholars as an important criterion for the assessment of young scholars' research proposals while it does not reach consensus among non-tenured scholars. Other than that, the selection of aspects reaching consensus does not differ between tenured and non-tenured scholars.
Regarding gender, we only find two aspects that reach the threshold of *Cohen's d*=0.2. Women rate the *variety* (*d*=0.20) and the *applicant's diploma* (*d*=0.21) higher than men. *Variety* reaches consensus among women but not among men (the bootstrapped stability interval for men amounts to 0.07-0.12), revealing slight gender differences regarding the ratings of the criteria and aspects for the assessment of young scholars' research proposals.

Conclusions

In this research in progress paper, we investigated the criteria and aspects that humanities scholars feel important and adequate to assess research proposals by young humanities scholars. We found that 23 aspects pertaining to 9 criteria reach consensus among the scholars: originality (*identifying gaps*, *innovative data*, *new findings)*, feasibility (*timetable* and *resources)*, rigour (*research process*, *state of research*, *choice of method*, *stringent argumentation*, *understandable)*, relevance (*academic relevance*), complexity (*making complexity visible)* and variety (*variety of research*).
Regarding our first research question, whether there are differences between the criteria for evaluating the proposals from PhD students and those for evaluating proposals from postdocs, we found that there are no such differences. The same selection of criteria reached consensus for both evaluation situations. Regarding our second research question, whether there are disciplinary differences between the evaluation criteria, we also found that the same selection of criteria reached consensus in the three groups of disciplines we investigated. However, there were some differences in the means between disciplines that are related to the topics of research in the disciplines: scholars in language and literature and in history and cultural studies emphasised more the criterion *cultural heritage* than scholars in law studies, while scholars in law studies rated *societal relevance* higher. This points to the fact that while the same evaluation sheet can be used in all disciplines, the weighting of the criteria (can) differ by discipline. The third research question focused on differences between tenured and non-tenured scholars. We found very little differences, however, the *applicant's CV* reached consensus among tenured scholars but not among non-tenured scholars. Finally, we investigated gender differences in the ratings. Regarding gender, the differences are rather small as well and only one criterion, *variety of research*, reached consensus among women but not among men. Thus, we can conclude that the 23 aspects pertaining to 9 criteria are rather generally applicable in the assessment of research proposals from early career researchers in the humanities.

Regarding the application of the criteria in the assessment of research proposals, it remains to be noted that we present findings from the perspective of the scholars themselves, thus referring to the academic quality of the proposal. In funding decisions, also criteria put forward by the funder might be added. In the further analysis of our data, we will delve deeper into the relation of the criteria for the assessment of proposals by early career researchers and general quality criteria identified in our previous research (Hug *et al.*, 2013) as well as the interrelations of gender, tenure and disciplines regarding the ratings of the evaluation criteria.

Acknowledgements

The authors would like to thank swissuniversities for their grant for the project "Application of Bottom-up Criteria in the Assessment of Grant Proposals of Junior Researchers" within the "Programme P-3 Performances de la recherche en sciences humaines et sociales". Matching funds were provided by the University of Zurich.

References

van Arensbergen, P., van der Weijden, I., & van den Besselaar, P. (2014a). Different views on scholarly talent: What are the talents we are looking for in science? *Research Evaluation, 23*(4), 273–284. doi:10.1093/reseval/rvu015

van Arensbergen, P., van der Weijden, I., & van den Besselaar, P. (2014b). The selection of talent as a group process. A literature review on the social dynamics of decision making in grant panels. *Research Evaluation, 23*(4), 298–311. doi:10.1093/reseval/rvu017

van den Akker, W. (2016). Yes we should; research assessment in the humanities. In M. Ochsner, S. E. Hug & H.-D. Daniel (eds). *Research Assessment in the Humanities. Towards Criteria and Procedures* (pp. 23–29). Cham: Springer International Publishing.

Bornmann, L., Mutz, R., & Daniel, H.-D. (2008). How to detect indications of potential sources of bias in peer review: A generalized latent variable modeling approach exemplified by a gender study. *Journal of Informetrics, 2*(4), 280–287. doi:10.1016/j.joi.2008.09.003

Bornmann, L., Mutz, R., Marx, W., Schier, H., & Daniel, H.-D. (2011). A multilevel modelling approach to investigating the predictive validity of editorial decisions: do the editors of a high profile journal select manuscripts that are highly cited after publication? *Journal of the Royal Statistical Society, 174*(4), 857–879. doi:10.1111/j.1467-985X.2011.00689.x

Braun, N., & Ganser, C. (2011). Fundamentale Erkenntnisse der Soziologie? Eine schriftliche Befragung von Professorinnen und Professoren der deutschen Soziologie und ihre Resultate. *Soziologie, 40*(2), 151–174.

Cardoso, S., Rosa, M. J., & Santos, C. S. (2013). Different academics' characteristics, different perceptions on quality assessment? *Quality Assurance in Education, 21*(1), 96–117. doi: 10.1108/09684881311293089

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Giménez-Toledo, E., & Román-Román, A. (2009). Assessment of humanities and social sciences monographs through their publishers: A review and a study towards a model of evaluation. *Research Evaluation, 18*(3), 201–213. doi:10.3152/095820209X471986.

Hemlin, S. (1993). Scientific quality in the eyes of the scientist. A questionnaire study. *Scientometrics, 27*(1), 3–18.

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

Krull, W., & Tepperwien, A. (2016). The four 'I's: Quality indicators for the humanities. In M. Ochsner, S. E. Hug & H.-D. Daniel (eds). *Research Assessment in the Humanities. Towards Criteria and Procedures* (pp. 165–179). Cham: Springer International Publishing.

Lamont, M. (2009). How professors think: Inside the curious world of academic judgment. Cambridge: Harvard University Press.

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 Erziehungswissenschaft, 17(6 Suppl.), 111–132. http://doi.org/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.

Thorngate, W., Dawes, R. M., & Foddy, M. (2009). *Judging merit*. Hove, UK: Psychology Press Taylor & Francis Group.