Humanities Scholars' Conceptions of Research Quality

Michael Ochsner, Sven E. Hug and Hans-Dieter Daniel

Abstract The assessment of research performance in the humanities is linked to the question of what humanities scholars perceive as 'good research'. Even though scholars themselves evaluate research on a daily basis, e.g. while reading other scholars' research, not much is known about the quality concepts scholars rely on in their judgment of research. This chapter presents a project funded by the Rectors' Conference of the Swiss Universities, in which humanities scholars' conceptions of research quality were investigated and translated into an approach to research evaluation in the humanities. The approach involves the scholars of a given discipline and seeks to identify agreed-upon concepts of quality. By applying the approach to three humanities disciplines, the project reveals both the opportunities and limitations of research quality assessment in the humanities: A research assessment by means of quality criteria presents opportunities to make visible and evaluate humanities research, while a quantitative assessment by means of indicators is very limited and is not accepted by scholars. However, indicators that are linked to the humanities scholars' notions of quality can be used to support peers in the evaluation process (i.e. informed peer review).

M. Ochsner (✉) · S.E. Hug · H.-D. Daniel
Professorship for Social Psychology and Research on Higher Education,
Department of Humanities, Social and Political Sciences, ETH Zürich,
Mühlegasse 21, 8001 Zürich, Switzerland
e-mail: ochsner@gess.ethz.ch

S.E. Hug
e-mail: sven.hug@gess.ethz.ch

H.-D. Daniel
e-mail: daniel@gess.ethz.ch

M. Ochsner
FORS, University of Lausanne, Géopolis, 1015 Lausanne, Switzerland

S.E. Hug · H.-D. Daniel
Evaluation Office, University of Zürich, Mühlegasse 21, 8001 Zürich, Switzerland

M. Ochsner et al. (eds.), *Research Assessment in the Humanities*,
DOI 10.1007/978-3-319-29016-4_5

1 Introduction

In order to evaluate research performance adequately, there should be an explicit understanding of what ‘good’ research is. Thus, knowledge about research quality is necessary. However, little is known about research quality, especially in the humanities. Existing tools and procedures of evaluation or assessment of (humanities’) research do not include an explicit understanding of quality. Even more so, the literature on research evaluation actively avoids the topic, reverting to ‘impact’, which is easier to measure but not necessarily congruent with research quality.

Yet, the assessment of research performance in the humanities must be linked to the question of what humanities scholars perceive as ‘good research’. In a report, the League of European Research Universities (LERU) formulated this in the following way: ‘senior administrators and academics must take account of the views of those ‘at the coal-face’ of research when developing assessment criteria and indicators (as should governments, funders and other external agencies)’ (League of European Research Universities 2012, p. 15). If we do not know what ‘good research’ is, it is impossible to assess it, let alone to improve it. Explicating what characterizes ‘good research’ is not only important for the assessment of research, but it is also of value to the scholars themselves.

This chapter presents a project[1] in which humanities scholars’ conceptions of research quality were investigated, and an approach to research evaluation in the humanities was developed. This chapter is structured as follows: In section one, we outline a framework for developing criteria and indicators for research quality in the humanities. In the subsequent section, we present the results of two studies in which we implemented this framework: In particular, section two describes humanities scholars’ notions of quality derived from repertory grid interviews, and section three presents the results from a three-round Delphi survey that resulted in a catalogue of quality criteria and indicators as well as a list of consensual quality criteria and indicators. In section four, we discuss the advantages of basing quality criteria and indicators on scholars’ notions of quality before we conclude the chapter with a summary and an outlook.

2 Framework

The bibliometric indicators that are widely used for evaluation in the natural and life sciences should not be applied to evaluate humanities research (Archambault et al. 2006; Bourke and Butler 1996; Butler and Visser 2006; Finkenstaedt 1990; Glänzel

[1] The Swiss University Conference started a project organized by the Rectors’ Conference of the Swiss Universities (since 1 January 2015 called swissuniversities) entitled ‘B-05 mesurer la performance de la recherche’ (see also http://www.performances-recherche.ch/). The project consisted of three initiatives (i.e. (sub-)projects) and four actions (i.e. workshops and add-ons to the initiatives). This chapter presents such an initiative entitled ‘Developing and Testing Research Quality Criteria in the Humanities, with an emphasis on Literature Studies and Art History’. Even though *initiative* would be the correct term, we use the term *project* throughout this chapter for reasons of readability.

and Schoepflin 1999; Gomez-Caridad 1999; Guillory 2005; Hicks 2004; Moed et al. 2002; Nederhof 2006; Nederhof et al. 1989). Since many evaluation procedures are based on quantitative approaches, evaluation faces strong opposition by humanities scholars. Even though there have been different projects initiated to develop assessment tools that might fit to the humanities as well (e.g. Australian Research Council 2012; Engels et al. 2012; European Science Foundation 2011; Giménez-Toledo and Román-Román 2009; Gogolin et al. 2014; Royal Netherlands Academy of Arts and Sciences 2011; Schneider 2009; Sivertsen 2010; White et al. 2009; Wissenschaftsrat 2011b), they are discussed very controversially in the humanities, and some of them have even been rejected or faced boycott by the humanities scholars (e.g. the ERIH project of the European Science Foundation, see Andersen et al. (2009), or the Forschungsrating of the German Wissenschaftsrat, see e.g. Plumpe (2009)). We analysed this critique and identified four main reservations. We then developed a framework that addresses these four points of critique and that can serve as a foundation to develop criteria for research assessment. This framework has been published in Hug et al. (2014), and this section draws on this article.

2.1 The Four Main Reservations About Tools and Procedures for Research Evaluation

While humanities scholars criticize many different aspects of research evaluation and its tools and instruments, four main reservations can be identified that summarize many of these aspects: (1) the methods originating from the natural sciences, (2) strong reservations about quantification, (3) fear of negative steering effects of indicators and (4) a lack of consensus on quality criteria.

2.1.1 Methods Originating from the Natural Sciences

The first reservation relates to the fact that the methods used to assess research quality have their origin in the natural sciences (see e.g. Vec 2009, p. 6). Hence, they do not reflect the research process and the publication habits of humanities scholars, such as the importance of national language or the publication of monographs (see e.g. Lack 2008, p. 14), and this is also supported by bibliometric research (see e.g. Hicks 2004; Nederhof 2006). Furthermore, Lack (2008) warns that the existing procedures reflect a linear understanding of knowledge creation due to the natural sciences' notion of linear progress. However, humanities' and also much of the social sciences' conception of knowledge creation relies on the 'coexistence of competing ideas' and the 'expansion of knowledge' (Lack 2008, p. 14, own translation).

2.1.2 Strong Reservations About Quantification

Second, the quantification of research performance is met with scepticism. Some humanities scholars question the mere idea of quantifying research quality, as becomes evident in a joint letter by 24 philosophers to the Australian government as a reaction to their discontent with the journal ranking in the Excellence in Research for Australia (ERA) exercise: 'The problem is not that judgments of quality in research cannot currently be made, but rather that in disciplines like Philosophy, those standards cannot be given simple, mechanical, or quantitative expression' (Academics Australia 2008, p. 1). Particularly the intrinsic benefits of the arts and humanities are feared to be neglected by the use of quantitative measures. While Fisher et al. (2000) do not deny the possibility of a quantitative measurement of research performance, they stress that these indicators do not measure the important information: 'Some efforts soar and others sink, but it is not the measurable success that matters, rather the effort. Performance measures are anathema to arts because they narrow whereas the arts expand' (Fisher et al. 2000, 'The Value of a Liberal Education', para. 18).

2.1.3 Fear of Negative Steering Effects of Indicators

Third, indicators can have dysfunctional effects. Humanities scholars fear, for example, mainstreaming or conservative effects of indicators: 'Overall, performance indicators reinforce traditional academic values and practices and in trying to promote accountability, they can be regressive' (informant B in (Fisher et al. 2000), 'IV. Critiques of Current Performance Indicators', para. 8). A further negative effect frequently mentioned is the loss of diversity of research topics or even disciplines due to constraints and selection effects introduced by the use of research indicators—thus the reaction of nearly 50 editors of social sciences and humanities journals to the European Science Foundations' European Reference Index for the Humanities (ERIH). They argued as follows: 'If such measures as ERIH are adopted as metrics by funding and other agencies, [. . .] We will sustain fewer journals, much less diversity and impoverish our discipline' (Andersen et al. 2009, p. 8). On a more fine-grained scale, Hose (2009) describes the effect of a focus on citation counts as having 'the tendency to favour spectacular (and given certain circumstances, erroneous) results, and penalize fundamental research and sustainable results as well as those doing research in marginal fields' (Hose 2009, p. 95, own translation), an argument that has gained weight given the current discussion on spurious research findings in many disciplines in the life sciences (see e.g. Unreliable research. Trouble at the lab 2013; Mooneshinghe et al. 2007). Due to the poor reputation of replication and due to strong competition and the need to publish original research in high impact journals, research findings are hardly ever replicated (Unreliable research. Trouble at the lab 2013).

2.1.4 Lack of Consensus on Quality Criteria

The fourth reservation concerns the heterogeneity of paradigms and methods. If there is a lack of consensus on the subjects of research and the meaningful use of methods, a consensus on criteria to differentiate between 'good' and 'bad' research is difficult to achieve (see e.g. Herbert and Kaube 2008, p. 45). If, however, criteria do exist, they are informal, refer to one (sub)discipline and cannot easily be transformed to other subdisciplines [Kriterien werden 'informell formuliert, beziehen sich [...] auf die gleiche Fachrichtung und sind [...] nicht ohne weiteres auf andere Subdisziplinen übertragbar'] (Herbert and Kaube 2008, p. 40).

2.2 The Four Pillars of Our Framework to Develop Sustainable Quality Criteria

In order to take these criticisms into account, we developed a framework to explore and develop quality criteria for humanities research (Hug et al. 2014). It consists of four main pillars that directly address the four main criticisms. The four pillars are (1) adopting an inside-out approach, (2) relying on a sound measurement approach, (3) making the notions of quality explicit and (4) striving for consensus.

2.2.1 Adopting an Inside-Out Approach

If the goal of assessment is enhancing research or improving or assuring research quality, it is clear that we must know what quality actually is. In other words, we need to know what we want to foster. While many different stakeholders are involved in research policy (Brewer 2011; Spaapen et al. 2007, p. 79), it is also clear that only scholars can tell what really characterizes 'good research'. In 2012, the League of European Universities concluded that '[evaluators] must take account of the views of those "at the coal-face" of research when developing assessment criteria and indicators' (League of European Research Universities 2012, p. 15). It is, however, important that the different disciplines' unique quality criteria can emerge. Therefore, quality criteria for the humanities must be based on the humanities scholars' conceptions of research. This is best achieved by adopting an inside-out approach. Ideally, the development process should be rooted in the disciplines or even subdisciplines, since there are inter- and intradisciplinary differences within the humanities (e.g. Royal Netherlands Academy of Arts and Sciences 2011; Scheidegger 2007; Wissenschaftsrat 2011b). Furthermore, a genuine inside-out approach has an open outcome. This means that whatever the scholars define as a quality criterion will be accepted as such, no matter how different it might be from the already known criteria from the natural and life sciences. Finally, the inside-out approach implies a bottom-up procedure. This means that, on one hand, quality criteria should not be determined solely by political stakeholders, university administrators or a few experts in the field in a top-down manner but rather by the scholarly community in its entirety. On the other hand, this means also that not only professors should have a say in what the important quality criteria are, but also younger researchers' conceptions of quality must be taken into account, since research practices can change and new ways of doing research should be reflected in the quality criteria as well. Applying an inside-out approach and developing specific quality criteria for each discipline is the obvious answer to the reservation that the methods in research evaluation stem from the natural and life sciences and do not take into account the research and communication practices of the humanities.

2.2.2 Relying on a Sound Measurement Approach

While it might seem paradoxical to those who argue against quantification as such, we think that applying a sound measurement approach when developing quality criteria and indicators can account for the reservations about quantification. Such an approach is necessary, because in many evaluation practices, indicators are only very loosely linked to definitions of quality. If we want to measure a concept, however, we must first understand it. This belongs to the basic knowledge in empirical sciences: 'Before we can investigate the presence or absence of some attribute [...], or before we can rank objects or measure them in terms of some variable, we must form the concept of that variable' (Lazarsfeld and Barton 1951, p. 155). However, very often theoretical and empirical studies live separate lives. Goertz concludes from his study of the social sciences that 'in spite of the primordial importance of concepts, they have received relatively little attention over the years' (Goertz 2006, p. 1). This is also true for biblio- and scientometrics. Brooks, for example, concludes in her review of major quality assessments in the U.S. that '[the assessments] often still make only a weak connection between theoretical definitions of quality and its measures by asserting a single rank or rating system that obscures the methodological and theoretical assumptions built into it' (Brooks 2005, p. 1). Donovan also points to the fact that there is a weak or no link between indicators and quality criteria, since the measurement in evaluation is very often data-driven: 'This leads us to the observation that research 'quality' comes to be defined by its mode of evaluation; and it is the measures and processes employed [...] that become the arbiters of research excellence' (Donovan 2007, p. 586). Hence, research quality seems to be defined by its measures instead of the other way round. Looking at one of the most important indicators of research performance, namely citations, Moed finds that 'it is [...] extremely difficult if not impossible to express what citations measure in one single theoretical concept [...]. Citations measure many aspects of scholarly activity at the same time' (Moed 2005, p. 221).

If there is such a weak or even missing link between the concept(s) and indicators of quality while at the same time indicators are ambiguous, it is no surprise that humanities scholars have reservations about the quantification attempts. Hence, it is important to rely on a sound measurement approach, since the issue is not 'first

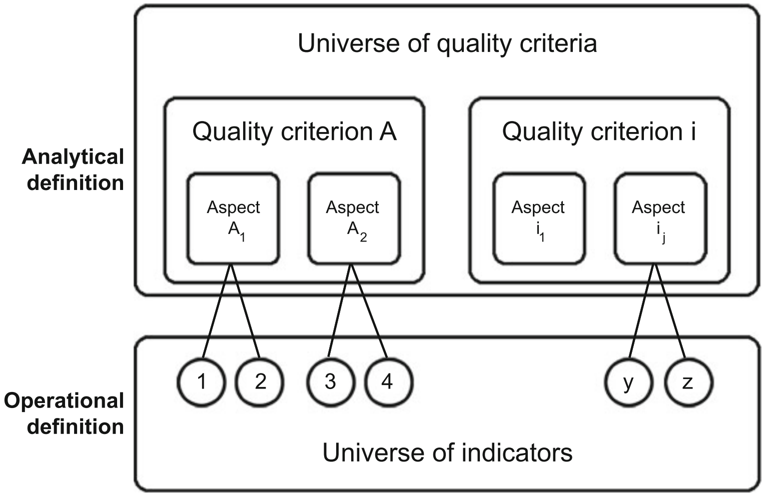

Fig. 1 Measurement model for developing quality criteria and indicators for the humanities. *Source* Hug et al. (2014)

to measure and then to find out what it is that is being measured but rather that the process must run the other way' (Borsboom et al. 2004, p. 1067). When it comes to measurement in research evaluation, it is therefore necessary to have an explicit understanding of quality (Schmidt 2005, p. 3).

We have therefore developed a measurement approach for the operationalization of research quality—the CAI-approach (Criteria, Aspect, Indicator). It is based on a measurement approach commonly used in the social sciences that includes an analytical and an operational definition of a concept (see Fig. 1) and consists of two parts. First, the concept, i.e. quality, has to be defined analytically. Every quality criterion is specified and defined explicitly by one or more aspects. These aspects can then be defined operationally: Each aspect is tied to one more indicators that specify how it can be observed, quantified or measured. Of course, it can be the case that, for a given aspect, no indicators can be found or thought of. Consequently, this aspect cannot be measured quantitatively. Therefore, this approach has the advantage that it is possible to identify quantifiable and non-quantifiable quality criteria. This might reduce scholars' reservations about quantification by disclosing what can be measured and what is exclusively accessible to the judgement of peers and by making clear that quality is not reduced to one simple quantitative indicator.

2.2.3 Making the Notions of Quality Explicit

The quotes by Brooks (2005), Donovan (2007) and Moed (2005) above show that it is not always clear what indicators are measuring. Hence, it is not evident along which criteria research is assessed and into which direction research is steered. The fact that it is not exactly known what indicators measure and, none the less important, what they do not measure might cause unintended effects of research assessment and trigger fear of negative steering effects in scholars. However, even if it is clear what the indicators of an assessment procedure do measure, scholars still might fear negative steering effects, because the criteria used might not be congruent with their notions of quality. Therefore, it is very important to make the scholars' notions of quality explicit. Yet, to explicate the scholars' notions of quality, it is important not to simply ask them what quality is. They very likely will answer something along the lines of 'I can't define what quality is, but I know it when I see it'. Lamont's study on peer review processes in the social sciences and humanities documents such statements (Lamont 2009). It shows that scholars certainly have knowledge on research quality, as they evaluate research many times during a working day. However, they cannot articulate this knowledge clearly and in detail. Polanyi (1967, p. 22) calls this phenomenon *tacit knowing* and describes it as the 'fact that we can know more than we can tell' (p. 4). *Explicit knowledge*, on the other hand, is 'capable of being clearly stated'. Since knowledge about research quality is still mainly *tacit knowing*, it is important to transform it into *explicit knowledge* in order to develop quality criteria for research assessment in the humanities. To sum up, notions of quality must be as explicit as possible, and the notions of quality of humanities scholars must be taken into account in order to reduce scholars' fears of negative steering effects—and even to reduce the probability of negative steering effects in general.

2.2.4 Striving for Consensus

If we want to develop evaluation criteria that are accepted by the majority of scholars, we must adopt an approach that allows for consensus within a discipline or sub-discipline. By including all scholars in a particular research community or discipline—that is, scholars from all sub-fields as well as methodological backgrounds, young scholars as well as senior professors—it assures the diversity of research and helps foster the acceptance of the criteria while also corresponding to the bottom-up approach described above.

2.3 The Implementation of the Framework: The Design of the Project 'Developing and Testing Quality Criteria for Research in the Humanities'

The design of the project is divided into two main phases: (I) an exploration phase and (II) a phase to find consensus. Because there was not much known about what research quality exactly is in the humanities and because the scholars' knowledge about research quality is mainly tacit, there was a need to first explore what research quality actually means to humanities scholars. Complying with the first and third pillars, i.e. to adopt an inside-out approach and to make notions of quality explicit, respectively, the exploration phase started the investigation into the notions of quality from scratch. For this aim, we conducted repertory grid interviews with 21 humanities scholars. This technique, developed by Kelly (1955), allows capturing subjective concepts that are used to interpret, structure, and evaluate entities that constitute the respondents' lives (see Fransella et al. 2004; Fromm 2004; Kelly 1955; Walker and Winter 2007). With this method, it is even possible to explicate tacit knowledge (Buessing et al. 2002; Jankowiecz 2001; Ryan and O'Connor 2009). Therefore, it is a very powerful instrument to explore researchers' notions of quality.

While it is possible to develop quality criteria from repertory grid interviews, we found it necessary to validate the criteria derived from the interviewed scholars' notions of quality, because we were able to conduct only a few repertory grid interviews due to the time-consuming nature of the technique. We also strove for consensus regarding the quality criteria according to the fourth pillar of the framework. Hence, we administered a Delphi survey to a large number of humanities scholars. The Delphi method makes use of experts' opinions in multiple rounds with anonymous feedback after each round in order to solve a problem (Häder and Häder 2000; Linstone and Turoff 1975). A Delphi survey starts with an initial round that delineates the problem. This can be done by the research team or, as in our case, by a first qualitative round surveying the experts. This was part of phase I (exploration). The result was a catalogue of quality criteria. In phase II (consensus), two more Delphi rounds, this time in the form of structured questionnaires, served to identify those quality criteria and indicators that reach consensus among the scholars. The Delphi method addresses three pillars from the above framework: By including all scholars of a discipline at the target universities, it (1) contributes to the inside-out approach; (2) it assures a sound measurement approach by structuring the communication process, that is, by linking indicators to the scholars' quality criteria; (3) it facilitates reaching a consensus.

Because both the repertory grid technique as well as the Delphi method are time-consuming methods, we could not investigate the quality notions of a broad range of disciplines. We decided to focus on three disciplines that are characterized by the fact that the commonly used approaches to research evaluation, that is, biblio- and scientometrics, are especially difficult to apply: German literature studies (GLS), English literature studies (ELS) and art history (AH).

3 Notions of Quality: The Repertory Grid Interviews

We conducted 21 repertory grid interviews with researchers from the universities of Basel and Zurich. The sample consisted of 11 women and 10 men, nine of whom were professors, five were senior researchers with a *Habilitation* qualification and seven were researchers holding a PhD.

The repertory grid interviews are built around entities and events meaningful to the respondents in the grid's thematic. These entities and events are called *elements*. We used 17 elements relevant to the scholars' research lives. They were defined by the research team and a repertory grid expert. For example, two of the elements were 'Outstanding piece of research' = Important, outstanding piece of research in the last twenty years in my discipline; 'Lowly regarded peer' = A person in my discipline whose research I do not regard highly. Using 'research' as topic for the elements, the interviewees generated words or syntagms, so-called *constructs*, they associated with pairs of elements they were presented. At the same time, they rated the constructs that they had just generated according to how much they corresponded with each of the 17 elements (for a comprehensive list of the elements as well as an in-depth description of the method and its implementation, see Ochsner et al. 2013).

Repertory grids generate qualitative, i.e. linguistic, and quantitative, i.e. numeric, data at the same time. A look at the linguistic material reveals that there is much communality between the three disciplines. The top categories in all disciplines include 'innovation' and 'approach' (see Table 1). Furthermore, 'diversity' is an important topic in all disciplines. Some differences exist between the disciplines as well. For example, 'cooperation' is mentioned quite a lot in GLS and especially in ELS but only receives a few mentions in AH. Art history is characterized further by the importance of 'scientific rigour' and 'internationality'. GLS, on the other hand, is characterized by the verbalization of 'careerist' mentality, which is not mentioned in ELS and only sparsely in AH. ELS scholars strongly emphasize 'cooperation' and do not mention 'inspiration' and 'careerist' mentality.

If we now combine the linguistic and the numeric data by using factor and cluster analysis to group the linguistic data according to the corresponding numeric data, we can reveal tacit, discipline-specific structures of the elements and constructs. In all three disciplines, the factor analysis yielded a three-dimensional representation of the elements and constructs defined by a quality dimension, a time dimension and a success dimension (in terms of success in the scientific system). In all three disciplines, the quality dimension explained the biggest portion of the variance, which means that quality is the most important factor in structuring the scholars' conception of their research lives. In GLS, the time dimension was the second factor, whereas it was the third factor in the other two disciplines (for details on the method and the statistical results, see Ochsner et al. 2013). Using these dimensions to interpret the linguistic data, we can see which constructs differentiate between, for example, 'good' and 'bad' research. This is obviously important information, since we are looking for notions of quality and quality criteria. We can show, for example, that constructs like interdisciplinarity, public orientation and cooperation have both positive and negative connotations. Interdisciplinary research and cooperation are both positively connoted if they serve diversity and complexity. However, if they are strategically used in order to obtain funding they are negatively connoted. Similarly, public-oriented research is positively connoted if it is innovative, and a connection with public issues is established. It is negatively connoted if the research is driven by public needs and, hence, is not free, or if it is economistic or career driven.

Table 1 Semantic categorization of the constructs from the repertory grid interviews

Category	Total	GLS	ELS	AH
Innovation	14.4	15.0	17.0	11.1
Approach	12.6	18.3	9.4	9.3
Cooperation	10.2	10.0	17.0	3.7
Diversity	6.6	6.7	5.7	7.4
Research autonomy	6.0	5.0	1.9	11.1
Interdisciplinarity	5.4	5.0	7.5	3.7
Skills	4.8	3.3	5.7	5.6
Public impact/applicability	4.8	3.3	5.7	5.6
Rigour	4.8	1.7	1.9	11.1
Resources	4.2	5.0	3.8	3.7
Career-oriented	3.6	8.3	0.0	1.9
Research agenda	3.6	1.7	5.7	3.7
Topicality	3.0	1.7	3.8	3.7
Inspiration	3.0	3.3	0.0	5.6
Internationality	3.0	0.0	1.9	7.4
Openness	3.0	1.7	5.7	1.9
Recognized by peers	2.4	3.3	3.8	0.0
Specialization	2.4	3.3	1.9	1.9
Varia	2.4	3.3	1.9	1.9
Column total	100.0	100.0	100.0	100.0

Note Measures in percent; Total of constructs mentioned: (n = 167); German literature studies: (n = 60); English literature studies: (n = 53); art history: (n = 54); Professors: (n = 66); Habilitated: (n = 47); PhDs: (n = 54); Male: (n = 76); Female: (n = 91); Basel: (n = 94); Zurich: (n = 73). Some columns might not sum to 100 % due to rounding

Furthermore, the combined analysis also reveals more details about how scholars structure their views regarding research. It showed that, in all disciplines, scholars differentiate between a 'modern' and a 'traditional' conception of research. 'Modern' research is characterized as being international, interdisciplinary, cooperative and public-oriented, whereas 'traditional' research is typically disciplinary, individual and autonomous. Hence, interdisciplinarity, cooperation and public orientation are not indicators of quality but of the 'modern' conception of research. It is notable that there is no clear preference for either conception of research (the 'traditional' conception received slightly more positive ratings). Hence, we can find four types of humanities research (see Fig. 2): (1) positively connoted 'traditional' research, which describes the individual scholar working within one discipline, who as a lateral thinker can trigger new ideas; (2) positively connoted 'modern' research characterized by internationality, interdisciplinarity and societal orientation; (3) negatively connoted 'traditional' research that, due to strong introversion, can be described as monotheistic, too narrow and uncritical; and finally (4) negatively connoted 'mod-

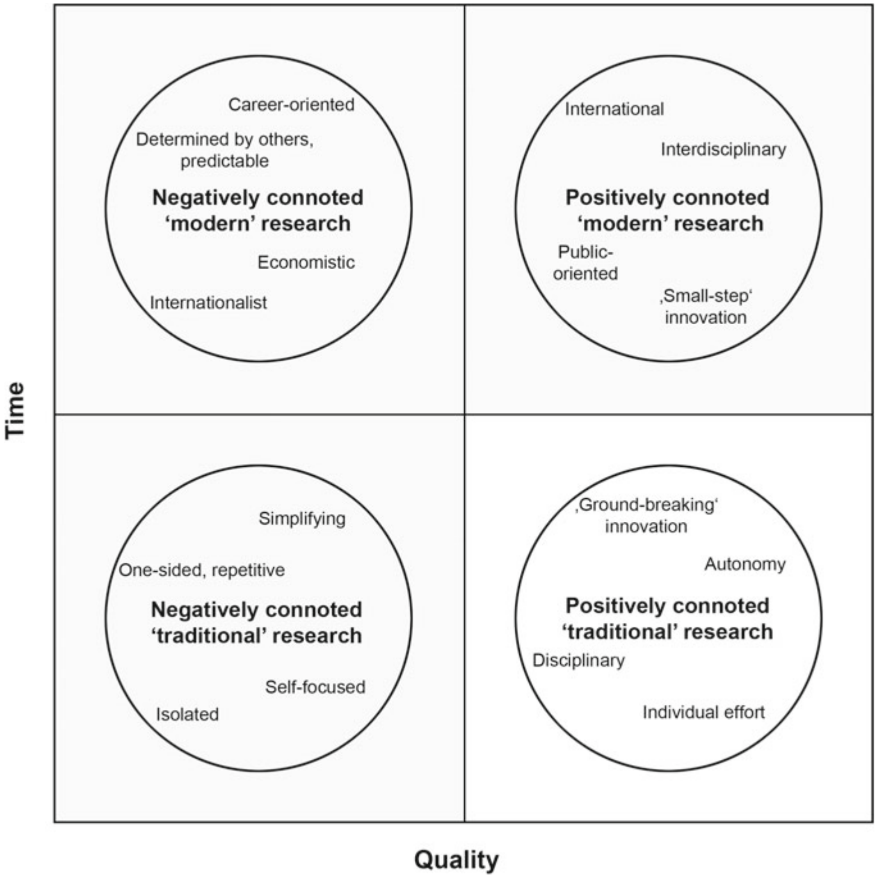
Career-oriented
Determined by others, predictable
Negatively connoted 'modern' research
Economistic
Internationalist
International
Interdisciplinary
Positively connoted 'modern' research
Public-oriented
‚Small-step' innovation
Simplifying
One-sided, repetitive
Negatively connoted 'traditional' research
Self-focused
Isolated
‚Ground-breaking' innovation
Autonomy
Positively connoted 'traditional' research
Disciplinary
Individual effort
Time
Quality

Fig. 2 Four types of research in the humanities. Commonalities across the disciplines. *Source* Ochsner et al. (2013), p. 86

ern' research that is characterized by pragmatism, career aspirations, economization and pre-structuring (see Fig. 2).

Using the time and success dimension, we can show that there are two forms of innovation. The first is connected to the 'modern' concept of research and is characterized as being an innovation of 'small steps'. It is based on new methods or current knowledge. The second is related to the 'traditional' concept of research. It is a 'ground breaking' innovation that is avant-gardist and brings about great changes (such as a paradigm shift). It is in all disciplines close to the element 'misunderstood luminary'. Hence, innovation, as a quality criterion, is double-edged along the success dimension. It can characterize successful research ('small-step' innovation) but also unsuccessful or not-yet-successful research ('ground breaking' innovation).

While the combined analysis of the quantitative and linguistic data is very useful to reveal insights into the implicit notions of quality and is therefore superior to the traditional qualitative analysis of, for example, interview data (McGeorge and Rugg 1992, pp. 151–152; Winter 1992, pp. 348–351), the interpretation of the linguistic material presented as the first results of the repertory grid reveals valuable information about the salience of some constructs, for example, that innovation, approach and diversity are used often to describe research. Additionally, we can see that internationality is salient only in art history and comes only rarely to the mind of literature scholars when describing research. They talk more often of cooperation. In German literature studies, 'careerist' behaviour is often mentioned.

Getting into the details of the notions of quality, we can see, however, that despite these differences, the notions of quality are still similar. Figures 3, 4 and 5 show a visualization of the elements and clusters of constructs for the three disciplines. In these graphs, the distances between an element and another element, or between a cluster and another cluster, can be interpreted as similarity: The closer two elements are to each other, the more similar they are. However, because the elements and the clusters are scaled differently, the interpretation of the distances between elements and clusters is accessible exclusively via their relative positioning. For example, if a cluster lies closer to an element than a second cluster does, there is greater similarity between the first cluster and the element than between the second cluster and the element (e.g. in Fig. 3, cluster 11, 'productive', is more similar to the element 'research with reception' than cluster 4, 'self-focused'). We simplified the graphical representations for this publication to increase their readability. The clusters were

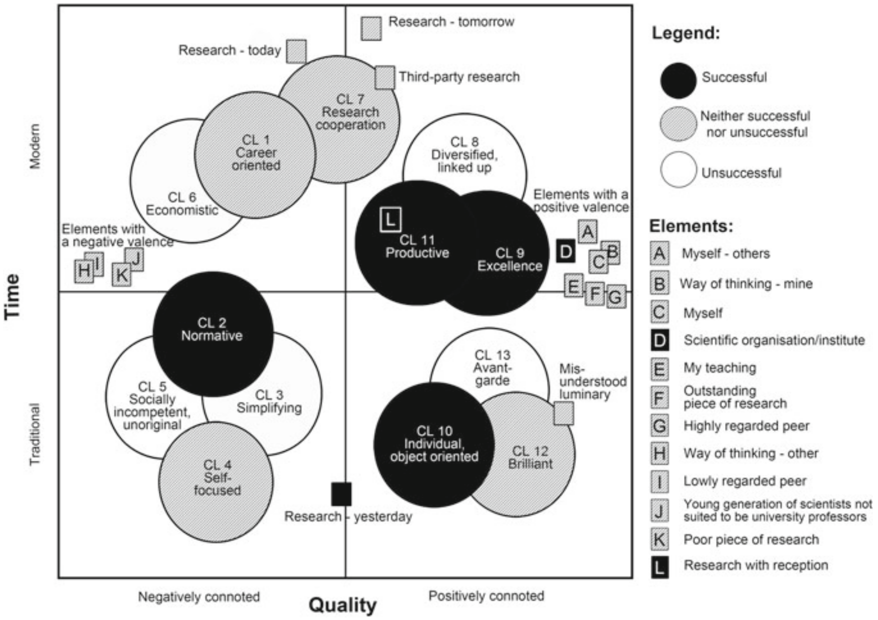

Fig. 3 Schematic representation of the clusters and elements in the discipline *German literature studies*. Slightly modified version of Ochsner et al. (2013), p. 84

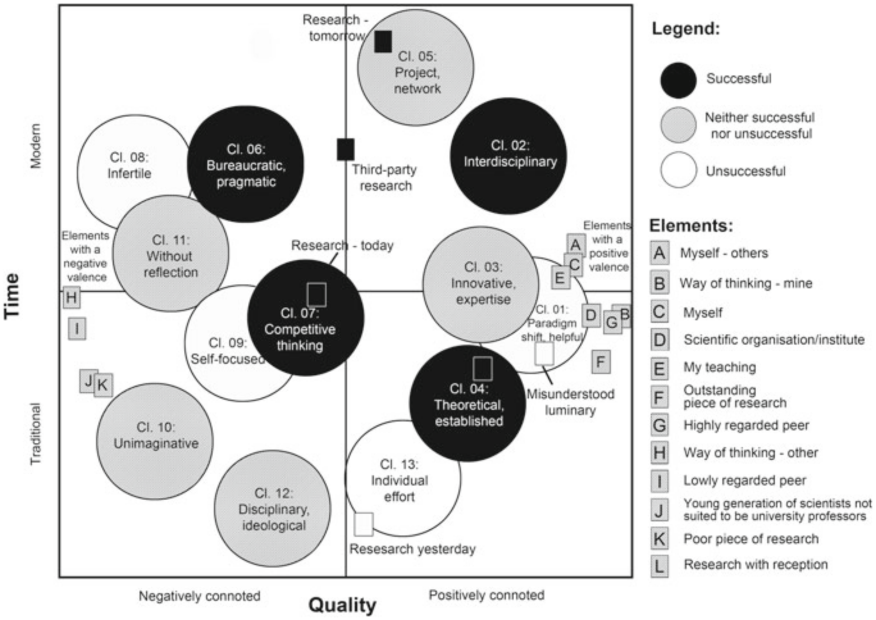
Research - tomorrow
Cl. 05: Project, network
Cl. 08: Infertile
Cl. 06: Bureaucratic, pragmatic
Third-party research
Cl. 02: Interdisciplinary
Modern
Time
Traditional
Elements with a negative valence
Cl. 11: Without reflection
Research - today
Cl. 07: Competitive thinking
Cl. 09: Self-focused
Cl. 03: Innovative, expertise
Elements with a positive valence
Cl. 01: Paradigm shift, helpful
Cl. 04: Theoretical, established
Misunderstood luminary
Cl. 10: Unimaginative
Cl. 12: Disciplinary, ideological
Cl. 13: Individual effort
Resesarch yesterday
H
I
J
K
A
C
E
D
G
B
F
Negatively connoted
Quality
Positively connoted
Legend:
Successful
Neither successful nor unsuccessful
Unsuccessful
Elements:
A Myself - others
B Way of thinking - mine
C Myself
D Scientific organisation/institute
E My teaching
F Outstanding piece of research
G Highly regarded peer
H Way of thinking - other
I Lowly regarded peer
J Young generation of scientists not suited to be university professors
K Poor piece of research
L Research with reception

Fig. 4 Schematic representation of the clusters and elements in the discipline *English literature studies*

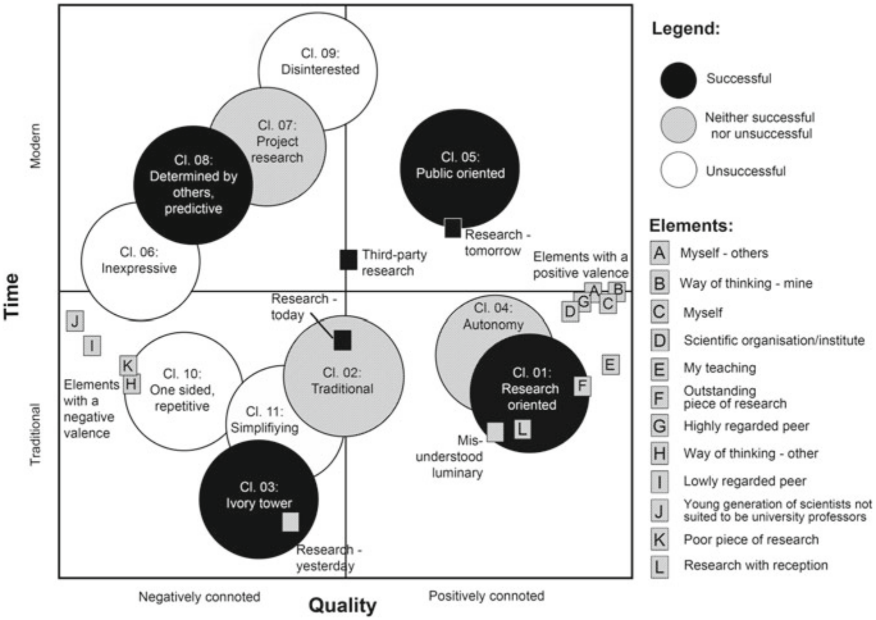

Fig. 5 Schematic representation of the clusters and elements in the discipline *art history*

placed schematically in the two-dimensional space with the axes *quality* and *time*, and the third dimension (*success*) was divided into three groups: successful, neither successful nor unsuccessful and unsuccessful.

The repertory grid for GLS is shown in Fig. 3. For example, cluster 1 represents 'career-oriented' research. Seen from the analysis of the linguistic material only, this is a concept solely salient in GLS. However, we can also find similar clusters in ELS and AH: In ELS, cluster 6, 'bureaucratic, pragmatic', describes applied research that is pragmatic and bureaucratic, associated with numbers-oriented evaluation. It is located in the negatively connoted 'modern' conception of research (see Fig. 4). In AH, cluster 8, 'determined by others', is located at a similar place in the grid and comprises research that is determined by others, elitist, overestimation of self and predictable, controllable and manageable (see Fig. 5). The three clusters encompass the same concept, career-focused strategies of research characterized by writing proposals and adapting to mainstream research. However, only the scholars of GLS clearly name it career-oriented, while in the other disciplines, it is more circumscribed and not clear-cut. However, there are also small differences. In GLS, this cluster's research is characterized by being neither successful nor unsuccessful, whereas in the other two disciplines this kind of research is characterized as successful. Furthermore, there is another cluster in ELS related to a careerist attitude: cluster 7, 'competitive thinking'. It shares the success-oriented approach to research. However, it is more focused on catching the attention of peers than on funding and social impact. This cluster is not restricted to the 'modern' conception of research but rather spreads across the time axis.

There are also clusters that are very similar in all three disciplines: Cluster 7 in GLS, cluster 5 in ELS and cluster 7 in AH are about project or network research. They are part of the 'modern' conception of research and are characterized by differentiation, cooperation, concerted activities and economization pressure. Also in the positively connoted 'traditional' conception of research, there is a cluster that is very similar in all disciplines: Cluster 13 in GLS ('avant-garde'), cluster 1 in ELS ('paradigm shift, helpful') and cluster 4 in AH ('autonomy'). They are all closely related to the element 'misunderstood luminary' and consist of research that is bringing about a paradigm shift by means of theoretical advancement and that is characterized by autonomy and unpredictability. This kind of research is not successful (yet): In GLS and ELS, it belongs to the unsuccessful clusters and in AH, to the neither successful nor unsuccessful clusters.

A peculiarity of AH is that there is only successful research in the positively connoted 'modern' conception of research. In Fig. 5, we can see that there is a positive correlation between the success and the quality dimensions in AH. There is no unsuccessful research both in the positively connoted 'modern' and in the positively connoted 'traditional' conception of research (the correlation of the two dimensions is $r = 0.43$) in AH). In the other two disciplines, the correlation is less striking (GLS: $r = 0.29$); ELS: $r = 0.26$).

4 Consensual Quality Criteria: The Delphi Survey

In order to validate our catalogue of quality criteria, we used the Delphi method. Complying with the bottom-up approach, our panel consisted of all research-active faculty at Swiss universities holding a PhD in GLS, ELS or AH. In order to ensure international standards and comparability, the panel also included all research-active faculty holding a PhD in the three disciplines at the member universities of the League of the European Research Universities (LERU). The first round of the Delphi served to complete the catalogue. The respondents could check or uncheck the existing quality criteria and aspects as well as name new criteria and/or aspects. We also asked for indicators that measure the quality aspects. Because of the heavy workload required to respond to this questionnaire, it was administered to only a part of the sample ($n = 180$) scholars). The first round achieved a response rate of 28 % and resulted in a more refined catalogue of quality criteria, comprising 19 criteria specified by a total of 70 aspects (for a description of the method and the results, see Hug et al. 2013). In the second Delphi round, which was administered to the whole sample $N = 664$), the scholars rated the aspects on a scale from 1 to 6 as to whether they agreed with a given statement. The statement consisted of a generic part that was the same for all aspects (i.e. 'My research is assessed appropriately if the assessment considers whether I . . .') and a second part consisting of the aspect (e.g. '. . . introduce new research topics') of a given criterion (e.g. *Innovation*, *Originality*); 1 was labelled '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'. The second round achieved a response rate of 30 %.

The second Delphi round showed that a broad palette of quality criteria and aspects are needed to appropriately assess research quality in the humanities. Table 2 lists the 19 criteria for research quality in the humanities (for a list of all the 70 aspects, see Hug et al. 2013). In GLS, only 10 out of the 70 aspects scored a mean of less than 4, of which only two received a median lower than 4. The same numbers apply for AH. In ELS, however, 13 aspects scored a mean of less than 4, and five aspects had a median lower than 4. The grand mean of the aspect was 4.71 (range = 3.34–5.74), 4.64 (range = 3.15–5.6) and 4.56 (range = 2.88–5.56) in GLS, AH and ELS, respectively. Of the aspects that have received a negative rating (i.e. mean lower than 4), seven were rejected in all three disciplines—namely, 'reputation in society' and 'insights are recognized by society' (*recognition*), 'continuation of research traditions' and 'long-term pursuit of research topics' (*continuity*, *continuation*), 'establishing a new school of thought' (*impact on research community*), 'responding to societal concerns' (*relation to and impact on society*) and 'research has its impact mainly in teaching' (*connection between research and teaching*, *scholarship of teaching*). Furthermore, in all three disciplines, no criterion was rejected altogether since each criterion had at least one aspect that had been rated with a 4 ('I slightly agree') by at least 50 % of the scholars ($mean > 4$). Hence, the catalogue that resulted from the repertory grid and the first Delphi round aptly reflects the notions of quality of the humanities scholars in the three disciplines.

Table 2 Quality criteria for humanities research: consensuality in the three disciplines

1.	Scholarly exchange^{GLS,ELS,AH}	8.	Continuity, continuation^{GLS}	15.	Scholarship, erudition^{GLS,ELS,AH}
2.	Innovation, originality^{GLS,ELS,AH}	9.	Impact on research community^{GLS,ELS,AH}	16.	Passion, enthusiasm^{GLS,ELS,AH}
3.	Productivity	10.	Relation to and impact on society	17.	Vision of future research^{GLS,ELS,AH}
4.	Rigour^{GLS,ELS,AH}	11.	Variety of research^{GLS,AH}	18.	Connection between research and teaching, scholarship of teaching^{GLS,ELS,AH}
5.	Fostering cultural memory^{GLS,ELS,AH}	12.	Connection to other research^{GLS,ELS,AH}	19.	Relevance^{GLS}
6.	Recognition^{ELS}	13.	Openness to ideas and persons^{GLS,ELS,AH}		
7.	Reflection, criticism^{GLS,AH}	14.	Selfmanagement, independence^{GLS,ELS}		

Note GLS = criterion reached consensus in German literature studies; ELS = criterion reached consensus in English literature studies; AH = criterion reached consensus in art history

However, regarding some aspects and criteria, the scholars were divided (i.e. while some scholars supported the aspect, a large number of others rated the same aspect very low). Therefore, and in order to comply with the fourth pillar of our framework (striving for consensus), we identified those aspects that were clearly approved by a majority and disapproved by very few scholars (i.e. consensual aspects). Consequently, we classified an aspect as consensual when at least 50 % of the discipline's respondents rated the aspect with at least a '5', and not more than 10 % of the discipline's respondents rated the aspect negatively, that is, with a '1', '2' or '3'. Accordingly, we classified a criterion as consensual when at least one of its aspects reached consensus. In GLS, 36 aspects pertaining to 16 criteria reached consensus, in AH, 31 aspects connected to 13 criteria did so and 29 aspects related to 13 criteria reached consensus in ELS. For simplicity reasons, we focus on the criteria in the further analysis. For information regarding the aspects, please refer to Hug et al. (2013).

The data revealed a set of shared criteria consisting of 11 criteria that reached consensus in all three disciplines. Note, however, that not all these criteria are specified with the same consensual aspects in the three disciplines. For example, the criterion *connection to other research* was specified differently in the three disciplines. In GLS, all three aspects of this criterion reached consensus: 'building on current state of research', 're-connecting to neglected research' and 'engaging in on-going research debates'; in ELS, the two aspects 'building on current state of research' and 're-connecting to neglected research' reached consensus; and in AH, only one aspect reached consensus: 'engaging in on-going research debates'. Moreover, six criteria were consensual in one or two disciplines and can be considered discipline-specific criteria. Finally, two criteria did not reach consensus in any discipline, namely *productivity* and *relation to and impact on society*. Table 2 indicates the consensuality of the criteria in the respective disciplines.

The fact that all criteria reached acceptable mean scores shows that in order to assess research quality in the humanities appropriately, a broad spectrum of quality criteria must be taken into account. Ten of the presented criteria are well known and are already used in evaluation procedures, and nine are less known—namely, fostering cultural memory, reflection/criticism, variety of research, openness to ideas and persons, self-management/independence, scholarship/erudition, passion/enthusiasm, vision of future research, connection between research and teaching/scholarship of teaching. Two of these criteria are also mentioned in the empirical literature on quality criteria in the humanities—reflection/criticism corresponding to reflexivity, deliberation and criticism (Oancea and Furlong 2007) and passion/enthusiasm corresponding to engagement (Bazeley 2010). However, if we look at the criteria that reached consensus, we see that all the nine less known criteria reach consensus in at least two disciplines, whereas some criteria that are very often used, i.e. productivity, recognition, relation to and impact on society and relevance, reach consensus in only one discipline or in none at all. Hence, from the point of view of the humanities scholars' notions of quality, there is doubt as to whether current evaluation criteria can capture research quality in the humanities (VolkswagenStiftung 2014, p. 1).

In order to investigate this issue further, we gathered indicators that are used or are suggested for use in evaluation procedures. These were collected in two steps. The first step consisted of an extensive literature review looking for documents that included criteria or indicators for research in the humanities and related disciplines or documents that addressed criticisms or conceptual aspects of research assessments. This resulted in a bibliography of literature on quality criteria and indicators for humanities research that is accessible on the project's website[2] (Peric et al. 2013). In the second step, the collection of indicators was expanded with indicators that were named by the humanities scholars themselves in our repertory grid interviews and the first Delphi round. Because we identified an abundance of indicators, we had to group them into clusters. The grouping procedure resulted in 62 groups of indicators by following two principles: The indicators of a group must be of similar kind and—in order to comply with our measurement model—it should be possible to assign each group to a specific quality criterion or aspect (for a detailed description of the documents used and the assigning procedure, see Ochsner et al. 2012).

By assigning the indicator groups to the quality criteria and aspects, we are able to quantify the proportion of aspects that can be measured quantitatively. We were able to identify indicators for only about half of the aspects that reached consensus,

[2] See http://www.performances-recherche.ch/projects/developing-and-testing-quality-criteria-for-research-in-the-humanities.

53 % in GLS, 52 % in ELS and 48 % in AH, respectively. In other words, indicators can capture only about half of the humanities scholars' notions of quality.

The scholars rated these groups of indicators in the third Delphi round according to a clear statement on a scale ranging, again, from 1 to 6, where (1) meant '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'. The third Delphi round was designed similarly to the second round. Again, the statements consisted of two parts: a generic part (i.e. 'The following quantitative statements provide peers with good indications of whether I ...') and an aspect (e.g. '... realize my own chosen research goals') of a criterion (e.g. self-management/independence). This statement was followed by the groups of indicators assigned to the given aspect. Because every discipline had its own set of consensual aspects, the questionnaires differed between the disciplines.

In the third Delphi round, which achieved a response rate of 20 %, most items received ratings above 4 (i.e. agreement) by at least 50 % of the respondents. However, in order to be able to use the indicators in assessment procedures, they have to be accepted by most scholars. Hence, we identified the consensual indicators (consensus was defined the same way as in round two: that is, at least 50 % of the discipline's respondents rated the item with at least a '5', and not more than 10 % of the discipline's respondents rated the item with a '1', '2' or '3'). In GLS, 10 indicator groups reached consensus (12 %); in ELS, only one indicator group reached consensus (1 %) and in AH, 16 indicator groups reached consensus (22 %). This is considerably less than in round two, where 51 % of the aspects reached consensus in GLS, 41 % in ELS and 44 % in AH.

The participants also responded to a question asking whether they think that it is conceivable that experts (peers) could evaluate the participants' own research performance appropriately based solely on the quantitative data that the participants had just rated. This question was dismissed by a vast majority of the respondents (GLS: 88 %; ELS: 66 %; AH: 89 %).

5 Discussion: Notions of Quality at the Base of Assessment

Because other projects on research evaluation in the humanities have faced strong opposition (e.g. Andersen et al. 2009; Plumpe 2009, p. 209), we expected a very low willingness of the scholars to participate in our surveys. However, the first two Delphi rounds received quite high response rates of 28–30 %, respectively. Similar studies that surveyed professors report lower or similar response rates (e.g. Braun and Ganser 2011, p. 155; Frey et al. 2007, p. 360; Giménez-Toledo et al. 2013, p. 68). However, in the third Delphi round, where the topic moved from quality criteria to indicators for research performance, only 11 % of the scholars responded to the survey within the same timeframe as in the first two rounds. Even by significantly prolonging the field period, the response rate did not exceed 20 %. This constitutes initial evidence of the fact that scholars are ready and willing to discuss research quality by defining quality criteria but are not willing to narrow down quality to purely quantitative measures, i.e. indicators. This is further confirmed by the comments we received in response to our surveys. Whereas in the first two rounds the comments were predominantly positive, in the third round a clear majority of the comments was negative (for an analysis of the comments, see Ochsner et al. 2014). Also, the data reveal a clear divide between evaluation by criteria as opposed to evaluation by indicators. In all disciplines, the ratings of the aspects were clearly higher than those of the indicators. This holds true for the grand mean, the share of aspects or indicators that received a positive rating (i.e. $mean \geq 4$) and was even more pronounced for the share of aspects or indicators that reached consensus (for a more detailed integration and comparison of the three Delphi rounds and the repertory grid interviews, see Ochsner et al. 2014).

Hence, we can conclude that humanities scholars prefer a qualitative approach to research evaluation. They are willing to talk about notions of quality and to cooperate in developing quality criteria based on those notions of quality if a bottom-up approach is applied. In order to adequately assess research performance in the humanities, a broad range of quality criteria has to be taken into account. While there is strong reluctance to accept a quantitative approach, it is not rejected altogether. However, the indicators have to be connected to the scholars' notions of quality, i.e. quality criteria.

When on one hand most indicators were accepted by most of the respondents (i.e. most indicators scored a mean of above 4) but failed to reach consensus, the question arises as to why some scholars are reluctant to accept indicators and others approve of them. There are many different reasons for this, but our studies point to two possible reasons that have not yet gained much attention. Firstly, there is a mismatch of quality criteria and indicators between evaluators and humanities scholars, and secondly some quality criteria are double-edged in nature. The mismatch can be described as follows: Some criteria that are frequently used in evaluations are not perceived as indicative of research quality by the humanities scholars (e.g. reputation, societal impact, productivity). On the other hand, there are quality criteria that humanities scholars perceive as important to assess research quality which are not known or are not commonly used in evaluation protocols (e.g. fostering cultural memory, reflection/criticism, scholarship/erudition, passion/enthusiasm). Additionally—and due to constraints of space not reported in this article—the indicators most often used in research evaluations (e.g. citations, prizes, third-party funding, transfers to economy and society) measure criteria that do not reach consensus in all disciplines (i.e. recognition, impact on research community, relevance, relation to and impact on society; see Ochsner et al. 2012, pp. 3–4). The double-edged nature of some quality criteria is revealed in the results of the repertory grid study. Interdisciplinarity, cooperation, public orientation and internationality are often used as quality criteria in evaluation schemes. However, the repertory grid interviews reveal that they are indicators of the 'modern' as opposed to the 'traditional' conception of research and are not necessarily related to quality. If these criteria are used as quality criteria, the 'traditional' conception of research would be forced to 'take a back seat'. However, it has to be kept in mind that the 'traditional' conception of research is highly regarded by the scholars and is connected to an important aspect of innovation: the 'ground-breaking' innovation that establishes new paradigms and theories. Evaluators must not confuse the dichotomy of the 'modern' and 'traditional' conceptions of research with 'new/innovative/promising' versus 'old-fashioned/conservative'. Both are valuable, innovative and important in the humanities.

If humanities research is to be assessed appropriately, it is important that indicators for the 'traditional' conception of research are also used. Using the repertory grid and the Delphi method, we were able to also identify indicators for the 'traditional' conception of research (e.g. the indicator group 'number of sources, materials and original works used in publications or presentations', which measures the aspect 'rich experience with sources' from the criterion 'scholarship/erudition'). However, it is an open question as to whether the 'traditional' conception of research can be measured prospectively at all. The repertory grid interviews point clearly towards the prerequisite of autonomy for such achievements. Quantitative assessments are even explicitly a characteristic of the 'modern' conception of research—more specifically, the negatively connoted 'modern' conception of research (see Ochsner et al. 2013, pp. 91–92). On one hand, the measurement of some characteristics of the 'traditional' conception of research could make visible important contributions of humanities research that might be overlooked otherwise. It also might help promote humanities-specific notions of quality. On the other hand, the measurement of research performance might never capture the true notion of the 'traditional' conception of research, described as an individual researcher who is bringing about a paradigm change by conducting disciplinary research locked up in his study. Hence, many humanities scholars will likely be critical if not disapproving of quantitative measurement and purely indicator-based assessments, having in mind the ideal of the erudite scholar.

6 Conclusion

The assessment of humanities research is a controversly discussed topic. Particularly, the humanities scholars' acceptance of the assessment criteria is an unresolved problem. While most initiatives investigating ways to assess research quality in the humanities focus on enlarging databases, building new rankings or ratings, expanding the quantitative measures to societal impact or studying the peculiarities of humanities' research production (see, e.g. Australian Research Council 2012; Engels et al. 2012; Guetzkow et al. 2004; Hammarfelt 2012; Hemlin 1996; Lamont 2009; Nederhof 2011; Royal Netherlands Academy of Arts and Sciences 2011; Schneider 2009; Sivertsen 2010; White et al. 2009; Wissenschaftsrat 2011a, b; Zuccala 2012), we offer a different approach by starting with the humanities scholars' notions of quality and linking indicators to the quality criteria that are generated in a bottom-up procedure from within the humanities.

We suggest a framework for developing quality criteria for the humanities that comprises a bottom-up approach, a sound measurement approach, the explication of the humanities scholars' notions of quality and the principle of consensus (Hug et al. 2014). We implemented this framework using the repertory grid technique to explicate the scholars' implicit knowledge of quality, thereby making visible the scholars' notions of quality and generating a first catalogue of quality criteria. We then applied the Delphi method to survey all scholars of the three disciplines covered in this project—German literature studies, English literature studies and art history—at the Swiss and the LERU universities, thereby following a bottom-up procedure. The Delphi method made it possible to find a consensus on quality criteria.

From the results of the four studies we conducted during this project (repertory grid and three rounds of the Delphi survey), we can formulate opportunities for and limitations of research assessments in the humanities.

The limitations of research assessments in the humanities can be formulated as follows: We could identify quantitative indicators for only about 50 % of the notions of quality of the humanities scholars. As long as this holds true, humanities scholars will be very critical of purely indicator-based approaches to research assessment. Furthermore, those indicators that are most commonly used in procedures for research evaluation measure exactly those quality criteria and aspects that are not consensual among scholars (see Ochsner et al. 2012, p. 4). While the humanities scholars emphasize the importance of the 'traditional' conception of research, most indicators used in current research assessment procedures measure the 'modern' conception of research (see Ochsner et al. 2013, pp. 85–86).

However, while the humanities scholars' opposition to purely *indicator-based* research assessments will likely persist given the issues mentioned above, an approach towards research assessment relying on *quality criteria* based on the scholars' notions of quality presents opportunities (such as e.g. the guidelines of the VolkswagenStiftung: VolkswagenStiftung 2014). If a bottom-up approach is chosen and the humanities scholars are involved in formulating the quality criteria, and if a broad range of quality criteria are applied, humanities research can be assessed adequately. Using caution when linking indicators to relevant quality criteria, quantitative data can be used to inform judgements on these quality criteria. Hence, an *informed peer review* process based on the relevant quality criteria creates an opportunity to make humanities research more visible and to assess humanities research adequately. It furthermore facilitates the communication between different stakeholders in the evaluation process, and it helps young researchers to focus on quality criteria.

Of course, the research presented has some limitations. First, it is based on three humanities disciplines only. Future research should include a broader range of disciplines in the humanities and neighbouring disciplines. Second, while the response rates were quite high given the composition of the panel and the topic of the research as well as the workload of filling in the questionnaires, the results are based only on the responses of a third of the contacted scholars. Hence, future research should involve more scholars. Third, scholars are only one of several stakeholders involved in research assessments. Our approach could be used to investigate the notions of quality of other stakeholders.

Acknowledgments 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.

References

Academics Australia. (2008). Letter to senator the honourable Kim Carr, minister for innovation, industry, science and research (tech. rep. No. 15. September 2009). Retrieved from https://web.archive.org/web/20091221195149/, http://www.academics-australia.org/AA/ERA/era.pdf.

Andersen, H., Ariew, R., Feingold, M., Bag, A. K., Barrow-Green, J., van Dalen, B., et al. (2009). Editorial: journals under threat: A joint response from history of science, technology and medicine editors. *Social Studies of Science*, *39*(1), 6–9.

Archambault, E., Vignola-Gagné, E., Côté, G., Lavrivére, V., & Gringras, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics, 68*(3), 329–342. 1007. doi:10.1007/s11192-006-0115-z.

Australian Research Council. (2012). The excellence in research for Australia (ERA) initiative. Retrieved from http://www.arc.gov.au/excellence-research-australia.

Bazeley, P. (2010). Conceptualising research performance. *Studies in Higher Education*, *35*(8), 889–903. doi:10.1080/03075070903348404.

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.

Bourke, P., & Butler, L. (1996). Publication types, citation rates and evaluation. *Scientometrics*, *37*(3), 473–494. doi:10.1007/BF02019259.

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.

Brewer, J. D. (2011). The impact of impact. *Research Evaluation*, *20*(3), 255–256. doi:10.3152/095820211X12941371876869.

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

Buessing, A., Herbig, B., & Ewert, T. (2002). Implizites Wissen und erfahrungsgeleitetes Arbeitshandeln. Entwicklung einer Methode zur Explikation in der Krankenpflege [Implicit knowledge and experience guided working: Development of a method for explication in nursing]. *Zeitschrift für Arbeits- und Organisationspsychologie, 46*(1), 2–21. doi:10.1026//0932-4089.46.1.2.

Butler, L., & Visser, M. S. (2006). Extending citation analysis to non-source items. *Scientometrics*, *66*(2), 327–343. doi:10.1007/s11192-006-0024-1.

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

Engels, T. C., Ossenblok, T. L., & Spruyt, E. H. (2012). Changing publication patterns in the social sciences and humanities, 2000–2009. *Scientometrics*, *93*(2), 373–390. doi:10.1007/s11192-012-0680-2.

European Science Foundation. (2011). European Reference Index for the Humanities (ERIH). Retrieved from http://www.esf.org/erih.

Finkenstaedt, T. (1990). Measuring research performance in the humanities. *Scientometrics*, *19*(5–6), 409–417. doi:10.1007/BF02020703.

Fisher, D., Rubenson, K., Rockwell, K., Grosjean, G., & Atkinson-Grosjean, J. (2000). *Performance indicators and the humanities and social sciences*. Vancouver, BC: Centre for Policy Studies in Higher Education and Training.

Fransella, F., Bell, R., & Bannister, D. (2004). *A manual for repertory grid technique* (2nd ed.). Chichester: Wiley.

Frey, B. S., Humbert, S., & Schneider, F. (2007). Was denken deutsche Ökonomen? Eine empirische Auswertung einer Internetbefragung unter den Mitgliedern des Vereins für Sozialpolitik im Sommer 2006. *Perspektiven der Wirtschaftspolitik*, *8*(4), 359–377. doi:10.1111/1468-2516.00256.

Fromm, M. (2004). *Introduction to the repertory grid interview*. Münster: Waxmann.

Giménez-Toledo, E., Tejada-Artigas, C., & Mañana-Rodriguez, J. (2013). Evaluation of scientific books' publishers in social sciences and humanities: Results of a survey. *Research Evaluation*, *22*(1), 64–77. doi:10.1093/reseval/rvs036.

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.

Glänzel, W., & Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and social sciences. *Information Processing & Management*, *35*(1), 31–44. doi:10.1016/S0306-4573(98)00028-4.

Goertz, G. (2006). *Social sciences concepts: A user's guide*. Princeton, NJ: Princeton University Press.

Gogolin, I., Åström, F., & Hansen, A. (Eds.). (2014). *Assessing quality in European educational research. Indicators and approaches*. Wiesbaden: Springer VS.

Gomez-Caridad, I. (1999). Bibliometric indicators for research evaluation: Interfield differences. *Science Evaluation and Its Management*, *28*, 256–265.

Guetzkow, J., Lamont, M., & Mallard, G. (2004). What is originality in the social sciences and the humanities? *American Sociological Review*, *69*(2), 190–212. doi:10.1177/000312240406900203.

Guillory, J. (2005). Valuing the humanities, evaluating scholarship. *Profession*, *11*, 28–38. doi:10.1632/074069505X79071.

Häder, M., & Häder, S. (2000). *Die Delphi-Technik in den Sozialwissenschaften. Methodische Forschungen und innovative Anwendungen*. Wiesbaden: Westdeutscher Verlag.

Hammarfelt, B. (2012). Following the footnotes: A bibliometric analysis of citation patterns in literary studies. Doctoral dissertation. Skrifter utgivna vid institutionen för ABM vid Uppsala Universitet (Vol. 5). Uppsala: Uppsala Universitet. Retrieved from http://www.diva-portal.org/smash/get/diva2:511996/FULLTEXT01.pdf.

Hemlin, S. (1996). Social studies of the humanities. A case study of research conditions and performance in ancient history and classical archaeology and english. *Research Evaluation*, *6*(1), 53–61. doi:10.1093/rev/6.1.53.

Herbert, U., & Kaube, J. (2008). Die Mühen der Ebene: Über Standards, Leistung und Hochschulreform. In E. Lack & C. Markschies (Eds.), *What the hell is quality? Qualitätsstandards in den Geisteswissenschaften* (pp. 37–51). Frankfurt a. M.: Campus.

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.

Hose, M. (2009). Glanz und Elend der Zahl. In C. Prinz, & R. Hohls (Eds.), *Qualitätsmessung, Evaluation, Forschungsrating. Risiken und Chancen für die Geschichtswissenschaft?* (pp. 91–98). Historisches Forum. Berlin: Clioonline.

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.

Jankowiecz, D. (2001). Why does subjectivity make us nervous? Making the tacit explicit. *Journal of Intellectual Capital*, *2*(1), 61–73. doi:10.1108/14691930110380509.

Kelly, G. A. (1955). *The psychology of personal constructs*. New York, NY: Norton.

Lack, E. (2008). Einleitung—Das Zauberwort 'Standards'. In E. Lack & C. Markschies (Eds.), *What the hell is quality? Qualitätsstandards in den Geisteswissenschaften* (pp. 9–34). Frankfurt a. M.: Campus.

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

Lazarsfeld, P. F., & Barton, A. H. (1951). Qualitative measurement in the social sciences. Classification, typologies, and indices. In D. Lerner, & H. D. Lasswel (Eds.), *The policy sciences*. Stanford: Stanford University Press.

League of European Research Universities. (2012). Research universities and research assessment. Retrieved from http://www.leru.org/files/publications/LERU_PP_2012_May_Research_Assesment.pdf.

Linstone, H. A., & Turoff, M. (1975). Introduction. In H. A. Linstone, & M. Turoff (Eds.), *The delphi method. Techniques and applications* (pp. 3–12). Don Mills: Addison-Wesley.

McGeorge, P., & Rugg, G. (1992). The uses of 'contrived' knowledge elicitation techniques. *Expert System*, *9*(3), 149–154. doi:10.1111/j.1468-0394.1992.tb00395.x.

Moed, H. F., Luwel, M., & Nederhof, A. J. (2002). Towards research performance in the humanities. *Library Trends*, *50*(3), 498–520.

Moed, H. F. (2005). *Citation analysis in research evaluation*. Dordrecht: Springer.

Mooneshinghe, R., Khoury, M. J., & Janssens, A. C. J. W. (2007). Most published research findings are false-but a little replication goes a long way. *PLoSMedicine, 4*(2), e28. doi:10.1371/journal.pmed.0040028.

Nederhof, A. J., Zwaan, R. A., De Bruin, R. E., & Dekker, P. (1989). Assessing the usefulness of bibliometric indicators for the humanities and the social sciences: A comparative study. *Scientometrics*, *15*(5–6), 423–435. doi:10.1007/BF02017063.

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.

Nederhof, A. J. (2011). A bibliometric study of productivity and impact of modern language and literature research. *Research Evaluation*, *20*(2), 117–129. doi:10.3152/095820211X12941371876508.

Oancea, A., & Furlong, J. (2007). Expressions of excellence and the assessment of applied and practice-based research. *Research Papers in Education*, *22*(2), 119–137. doi:10.1080/02671520701296056.

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 Erziehungswissenschaft*, *17*(6 Supplement), 111–132. doi:10.1007/s11618-014-0576-4.

Peric, B., Ochsner, M., Hug, S. E., & Daniel, H.-D. (2013). *Arts and Humanities Research Assessment Bibliography (AHRABi)*. Zürich: ETH Zurich. doi:10.3929/ethz-a-010610785.

Plumpe, W. (2009). Stellungnahme zum Rating des Wissenschaftsrates aus Sicht des Historikerverbandes. In C. Prinz, & R. Hohls (Eds.), *Qualitätsmessung, Evaluation, Forschungsrating. Risiken und Chancen für die Geschichtswissenschaften?* (pp. 121–126). Historisches Forum. Berlin: Clio-online. Retrieved from http://edoc.hu-berlin.de/e_histfor/12/.

Polanyi, M. (1967). *The tacit dimension*. London: Routledge & Kegan Paul.

Royal Netherlands Academy of Arts and Sciences. (2011). *Quality indicators for research in the humanities*. Amsterdam: Royal Netherlands Academy of Arts and Sciences. Retrieved from https://www.knaw.nl/shared/resources/actueel/publicaties/pdf/quality-indicators-for-research-in-the-humanities.

Ryan, S., & O'Connor, R. V. (2009). Development of a team measure for tacit knowledge in software development teams. *Journal of Systems and Software*, *82*(2), 229–240. doi:10.1016/j.jss.2008.05.037.

Scheidegger, F. (2007). *Darstellung, Vergleich und Bewertung von Forschungsleistungen in den Geistes- und Sozialwissenschaften. Bestandesaufnahme der Literatur und von Beispielen aus dem In- und Ausland*. Bern: Zentrum für Wissenschafts- und Technologiestudien.

Schmidt, U. (2005). Zwischen Messen und Verstehen. Anmerkungen zum Theoriedefizit in der deutschen Hochschulevaluation (evaNet-Positionen 06/2005). Retrieved from http://www.forschungsnetzwerk.at/downloadpub/messen%20und%20verstehen.pdf.

Schneider, J. W. (2009). An outline of the bibliometric indicator used for performance-based funding of research institutions in Norway. *European Political Science*, *8*(3), 364–378. doi:10.1057/eps.2009.19.

Sivertsen, G. (2010). A performance indicator based on complete data for scientific publication output at research institutions. *ISSI Newsletter*, *6*(1), 22–28.

Spaapen, J., Dijstelbloem, H., & Wamelink, F. (2007). *Evaluating research in context: A method for comprehensive assessment*. The Hague: Consultative Committee of Sector Councils for Research and Development.

Unreliable research. Trouble at the lab. (2013, October 19). *The Economist*. Retrieved from http://www.economist.com/news/briefing/21588057-scientists-think-science-self-correcting-alarming-degree-it-not-trouble.

Vec, M. (2009). Die vergessene Freiheit. Steuerung und Kontrolle der Geisteswissenschaften unter der Prämisse der Prävention. In C. Prinz, & R. Hohls (Eds.), *Qualitätsmessung, Evaluation, Forschungsrating. Risiken und Chancen für die Geschichtswissenschaft?* (pp. 79–90). Historisches Forum. Berlin: Clioonline.

VolkswagenStiftung. (2014). *What is intellectual quality in the humanities? Some guidelines*. Hannover: VolkswagenStiftung. Retrieved from http://www.volkswagenstiftung.de/uploads/media/Humanities_Quality_Guidelines.pdf.

Walker, B. M., & Winter, D. (2007). The elaboration of personal construct psychology. *Annual Review of Psychology*, *58*, 453–477. doi:10.1146/annurev.psych.58.110405.085535.

White, H. D., Boell, S. K., Yu, H., Davis, M., Wilson, C. S., & Cole, F. T. H. (2009). Libcitations: A measure for comparative assessment of book publications in the humanities and social sciences. *Journal of the American Society for Information Science and Technology*, *60*(6), 1083–1096. doi:10.1002/asi.21045.

Winter, D. (1992). *Personal construct psychology in clinical practice: Theory, research and applications*. London: Routledge.

Wissenschaftsrat. (2010). *Empfehlungen zur vergleichenden Forschungsbewertung in den Geisteswissenschaften*. Köln: Wissenschaftsrat. Retrieved from http://www.wissenschaftsrat.de/download/archiv/10039-10.pdf.

Wissenschaftsrat. (2011a). *Forschungsrating Anglistik/Amerikanistik*. Köln: Wissenschaftsrat. Retrieved from http://www.wissenschaftsrat.de/nc/arbeitsbereiche-arbeitsprogramm/forschungsrating/anglistikamerikanistik.html.

Wissenschaftsrat. (2011b). *Zum Forschungsrating allgemein*. Köln: Wissenschaftsrat. Retrieved from http://www.wissenschaftsrat.de/arbeitsbereiche-arbeitsprogramm/forschungsrating.html.

Zuccala, A. (2012). Quality and influence in literary work: evaluating the 'educated imagination'. *Research Evaluation, 21*(3), 229–241. 1093, doi:10.1093/reseval/rvs017.