Setting the stage for the assessment of research quality in the humanities. Consolidating the results of four empirical studies

Author(s):
Ochsner, Michael; Hug, Sven E.; Daniel, Hans-Dieter

Publication date:
2014-11

Permanent link:
https://doi.org/10.3929/ethz-b-000094524

Originally published in:
Zeitschrift für Erziehungswissenschaften 17(6 Supplement), https://doi.org/10.1007/s11618-014-0576-4

Z Erziehungswiss (2014) (Suppl) 17:111–132
DOI 10.1007/s11618-014-0576-4

ZfE

Setting the stage for the assessment of research quality in the humanities. Consolidating the results of four empirical studies

Michael Ochsner · Sven E. Hug · Hans-Dieter Daniel

Abstract The assessment of research performance in the humanities is an intricate and highly discussed topic. Many problems have yet to be solved, foremost the question of the humanities scholars' acceptance of evaluation tools and procedures. This article presents the results of a project funded by the Rectors' Conference of the Swiss Universities in which an approach to research evaluation in the humanities is developed that focuses on consensuality. We describe the results of four studies and integrate them into limitations and opportunities of research quality assessment in the humanities. The results indicate that while an assessment by means of quantitative indicators exhibits limitations, a research assessment by means of quality criteria presents opportunities to evaluate humanities research and make it visible. Indicators that are linked to the humanities scholars' notions of quality can be used to support peers in the evaluation process (informed peer review).

Keywords Research evaluation · Humanities · Quality criteria · Repertory grid · Delphi method

M. Ochsner (✉) · S. E. Hug · H.-D. Daniel
D-GESS, ETH Zurich,
Muehlegasse 21,
8001 Zurich, Switzerland
e-mail: ochsner@gess.ethz.ch

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

S. E. Hug · H.-D. Daniel
ETH Zurich, University of Zurich,
Muehlegasse 21,
8001 Zurich, Switzerland

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

Voraussetzungen für die Beurteilung der Qualität geisteswissenschaftlicher Forschung: Zusammenführung der Befunde aus vier empirischen Studien

Zusammenfassung Die Beurteilung von Forschungsleistungen in den Geisteswissenschaften ist ein heikles und viel diskutiertes Thema. Es gilt, einige Probleme zu lösen, insbesondere die Frage der Akzeptanz von Evaluationsverfahren bei den Geisteswissenschaftlerinnen und Geisteswissenschaftlern selbst. Dieser Artikel präsentiert die Ergebnisse eines von der Rektorenkonferenz der Schweizer Universitäten geförderten Projektes, in welchen ein Ansatz für die Beurteilung von Forschungsleistungen in den Geisteswissenschaften entwickelt wurde, der auf dem Prinzip der Konsensualität beruht. Es werden die Befunde von vier Studien beschrieben und deren Resultate dahingehend zusammengeführt, dass Möglichkeiten und Grenzen der Beurteilung von geisteswissenschaftlicher Forschung formuliert werden. Die Befunde des Projektes weisen darauf hin, dass zwar einer Beurteilung von Forschungsleistungen basierend auf *quantitativen Indikatoren* Grenzen gesetzt sind, jedoch eine Beurteilung von Forschungsleistungen mittels *qualitativer Kriterien* Möglichkeiten und Chancen eröffnet, nicht nur bezüglich einer Evaluation, sondern auch im Hinblick auf die Sichtbarmachung geisteswissenschaftlicher Forschungsleistungen. Nichtsdestotrotz können in Evaluationsverfahren Indikatoren, die an das Qualitätsverständnis der Geisteswissenschaftlerinnen und Geisteswissenschaftler rückgebunden sind, zur Unterstützung von Gutachterinnen und Gutachtern eingesetzt werden (sog. *Informed Peer Review*).

Schlüsselwörter Forschungsevaluation · Geisteswissenschaften · Qualitätskriterien · Repertory Grid · Delphi-Methode

1 Introduction

Research assessments in the humanities are highly controversial and the evaluation of humanities research is delicate. While citation-based research performance indicators are widely used in the natural and life sciences, quantitative measures of research performance meet strong opposition in the humanities. However, the need for accountability reaches the stronghold of the humanities (Guillory 2005, p. 28). Since there are many problems related to the use of bibliometrics in the humanities (Hicks 2004; Nederhof 2006), new assessment approaches have to be considered for humanities research. This article presents an integration of the results of two studies consisting of four consecutive investigations as part of a project in which we developed an approach to assess research performance in the humanities (Hug et al. 2013, 2014; Ochsner et al. 2012, 2013). The consolidated results of the studies provide quality criteria and indicators for research in the humanities along with information on their consensuality among humanities scholars as well as a presentation of opportunities and limitations of the use of quality criteria and indicators in research assessments. It therefore sets the stage for the assessment of research quality in the humanities by focusing on what is possible (and why), yet still taking the risks and pit falls into

account, instead of focusing on what is different in the humanities and which methods do not work well, as is usually the case in the literature so far (see, e.g., Andersen et al. 2009; Hicks 2004; Nederhof 2006; Plumpe 2009). In the first part of this paper, we discuss the framework we used to develop quality criteria for assessing research performance. It includes an analysis of the humanities scholars' stance on research assessments. In the next part, we present the results of a qualitative-quantitative study that explicated the humanities scholars' implicit (i.e., non-verbalizable) conceptions of research quality. The third part describes a three-round Delphi survey. The first Delphi round consisted of a qualitative study of the scholars' perceptions of research quality and resulted in a catalogue of 19 quality criteria for humanities research. In the second Delphi round, humanities scholars rated the quality criteria. For the third Delphi round, we assigned indicators to the quality criteria and the scholars rated these indicators as to how appropriate they are to inform peers about the occurrence of the given aspect of the criterion in the research under investigation. Finally, the results of these studies are summarized and integrated into insights about the opportunities and limitations of research quality assessments in the humanities.

2 A framework to develop quality criteria for research assessment in the humanities

Humanities scholars have strong reservations against formal quality assessments that have exogenous roots (i.e., quality assessments in the light of new public management, performance based funding). This can be seen in the rejection or boycott of some recent initiatives to establish procedures or tools for the evaluation of humanities research quality (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). There are several reasons for their rejection of (quantitative) assessments of research quality. We identified four main reservations against the measurement of research quality put forward by humanities scholars (Hug et al. 2014). First, they criticize that the methods used in research evaluation originate from the natural and life sciences, and these methods were modelled according to the research process and the publication habits in the natural and life sciences (see, e.g., Lack 2008, p. 14; this is also supported by bibliometric research, see, e.g., Hicks 2004; Nederhof 2006). Second, humanities scholars have strong reservations regarding the quantification of research quality: 'Some efforts soar and others sink, but it is not the measurable success that matters, rather the effort' (Fisher et al. 2000, 'The Value of a Liberal Education', para. 18). Third, humanities scholars fear the dysfunctional effects of indicators. For example, they fear a loss of diversity. This is evident in the joint answer of nearly 50 editors of social sciences and humanities journals to the establishment of the European Reference Index for the Humanities (ERIH) by the European Science Foundation: '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). Fourth, there is a lack of consensus on the subjects of research and the meaningful use of methods. Therefore,

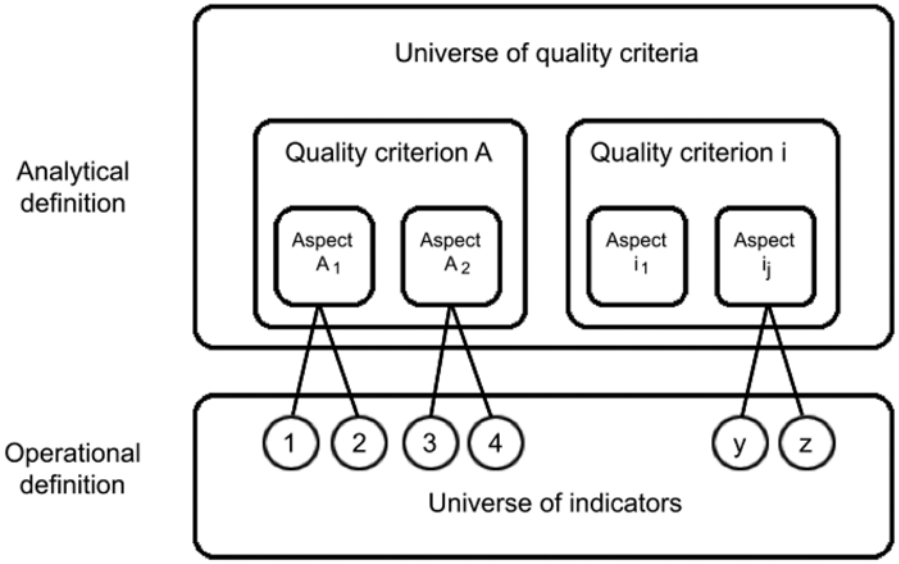

Fig. 1 Measurement model for developing quality criteria and indicators for the humanities. (Every quality criteria is specified explicitly by one or more aspects (i.e., analytical definition) and each aspect is operationalized. That is, each aspect is tied to at least one indicator that specifies how it can be observed, quantified or measured (i.e., operational definition). Naturally, it is also possible that no suitable quantitative indicators exist and, therefore, an aspect cannot be measured (e.g., Aspect i_1)). (Source: Hug et al. 2014)

a consensus on criteria to differentiate between good and bad research is far from being reached (see, e.g., Herbert and Kaube 2008, p. 45).

In order to address these objections, we suggest the following framework consisting of four pillars to develop quality criteria and indicators for humanities research (for details, see Hug et al. 2014).

The first pillar, *adopting an inside-out approach*, demands that the development of criteria and indicators be rooted in the humanities themselves, ideally in each discipline. Such a procedure, bottom-up in nature, ensures that the unique quality criteria and conceptions of each discipline can emerge. It is central to this pillar that the research community is involved or represented adequately in the development process and that the outcome is open (i.e., any quality criterion defined by the scholars is accepted, no matter how different from existing criteria).

Relying on a sound measurement approach, as the second pillar of the framework, responds to the humanities' reservations regarding quantification. To clarify what is being measured, quality criteria are specified by aspects (i.e., analytical definition of quality) that can be linked to quantitative indicators (i.e., operational definition). Figure 1 schematically illustrates such a measurement model. This approach allows for the identification of quantifiable and non-quantifiable quality criteria. If no quantitative indicators can be found to measure a quality criterion, this criterion is exclusively reliant on the judgement of peers. By unfolding a wide range of metrics that are

connected to the quality criteria, a sound measurement approach can resolve scholars' fears that research quality will be reduced to one simple, quantitative indicator.

The third pillar, *making the notions of quality explicit*, consists of two parts. Firstly, the notions of quality that underlie the measurement instrument, assessment tool or evaluation procedure should be made as explicit as possible to reduce uncertainties about what is being measured and make clear in what direction research is being steered. Secondly, when developing quality criteria and indicators for research, the scholars' notions of quality should be taken into account. This ensures that research is steered in the direction of the notions of quality put forward by humanities scholars themselves, thereby reducing their fear of negative steering effects.

The fourth pillar is *striving for consensus*. In order to be successful and not be rejected by scholars, an instrument or tool designed to determine the quantity and quality of research needs to be based on quality criteria and indicators that are accepted in the research community. Therefore, an approach should be chosen that is able to reveal which criteria are consensual to the research community and which are not.

We have implemented this framework using two specific methods: the repertory grid technique (Kelly 1955) and the Delphi survey method (Linstone and Turoff 1975).

The repertory grid technique addresses two pillars of our framework. On the one hand, it captures tacit knowledge (Buessing et al. 2002; Jankowiecz 2001; Ryan and O'Connor 2009) by exploring and mapping subjective concepts that individuals use to interpret their research lives (see Fransella et al. 2004; Fromm 2004; Walker and Winter 2007). On the other hand, it addresses the inside-out approach by generating conceptions of research quality genuine to the interviewed scholars.

While it is possible to derive quality criteria from the scholars' notions of quality using repertory grid interviews, these criteria are obtained from a small sample size due to the time-consuming nature of repertory grids. Thus, it is necessary to validate these criteria by a large group of scholars and to reach a consensus. We applied the Delphi method to this end. The Delphi method addresses three pillars of our framework. First, it contributes to the inside-out approach by involving a large group of scholars. In our case, participants include all research-active faculty in the three disciplines German literature studies (GLS), English literature studies (ELS), and art history (AH) at Swiss universities and at the member universities of the League of European Research Universities (LERU). Second, it assures the application of a sound measurement approach by structuring the group's communication process. This is achieved by systematically linking indicators to the scholars' quality criteria (see Fig. 1). Third, the Delphi method facilitates the process of reaching a consensus.

The delineated framework may help tackle the 'lack of information on how to develop indicators' and the 'problem related to the definition of indicators' (Palomares-Montero and Garcia-Aracil 2011, p. 354). The following sections present the results of the empirical implementation of this framework using repertory grid interviews and three rounds of a Delphi survey.

3 Repertory grid interviews

The repertory grid method was developed by George A. Kelly based on his Psychology of Personal Constructs (Kelly 1955). It captures the subjective conceptions (in Repertory Grids the so-called *constructs*) that individuals use to interpret, structure, and evaluate the entities (in Repertory Grids the so-called *elements*) that constitute their lives (Fransella et al. 2004; Fromm 2004; Walker and Winter 2007). According to Rosenberger and Freitag (2009), the method is very flexible because it allows an idiographic (i.e., the scholars describe their notions of research quality in their own words) as well as a nomothetic approach (i.e., the method allows for the development of discipline-specific criteria by summarizing the individual perceptions for each discipline). A great advantage of the repertory grid method is the ability to also capture tacit knowledge—that is, knowledge that can be put into words only with difficulty or not at all (Jankowiecz 2001, p. 64; Buessing et al. 2002, p. 3, 7–8; Ryan and O'Connor 2009, p. 232).

Twenty-one researchers participated in the repertory grid study (11 women, 10 men). We selected the participants according to three criteria: academic status, discipline, and university (Basel and Zurich). The sample consisted of nine professors, five senior researchers with a *Habilitation* qualification, and seven PhDs working at the University of Basel (a total of 12 academics) or Zurich (a total of 9 academics). Each of the three disciplines German literature studies, English literature studies, and art history was represented by seven interviewees.

The base of the interviews consisted of 17 entities and events in the participants' research lives (the so-called *elements*). For example, two of the elements were: *Outstanding piece of research* = Important, outstanding piece of research in the last 20 years in my discipline; *Lowly regarded peer* = A person in my discipline whose research I do not regard highly (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).

To evoke and capture the individual conceptions (the so-called *constructs*), the participants had to verbalize the similarity or difference between a pair of elements (e.g., 'way of thinking—other' and 'research with reception'). This set the initial pole of a construct (e.g., the pair of elements was rated as similar and the similarity was verbalized as 'political, calculated (money and power)'). The participants were then asked about the opposite pole, or what they saw as the opposite of the initial pole (e.g., 'object-related'). This evoking procedure was repeated several times with other element pairs, to capture all of a participant's conceptions decisive within the thematic framework of the interview. After evoking the constructs, the participants rated the 17 elements on the two-pole constructs that they had constructed in the preceding step. Repertory grid interviews thus generate linguistic (construct statements) as well as numeric data (grid ratings). The linguistic response material is interpreted based on numerical grouping by factor and cluster analysis. This makes it possible to discover implicit, discipline-specific structures of the elements and construct poles.

The analysis of the numerical and linguistic data revealed that the participants differentiate between a 'modern' and a 'traditional' conception of research. Whereas 'modern' research is international, interdisciplinary, and cooperative, 'traditional' research is disciplinary, individual, and autonomous. Both of them have positive as

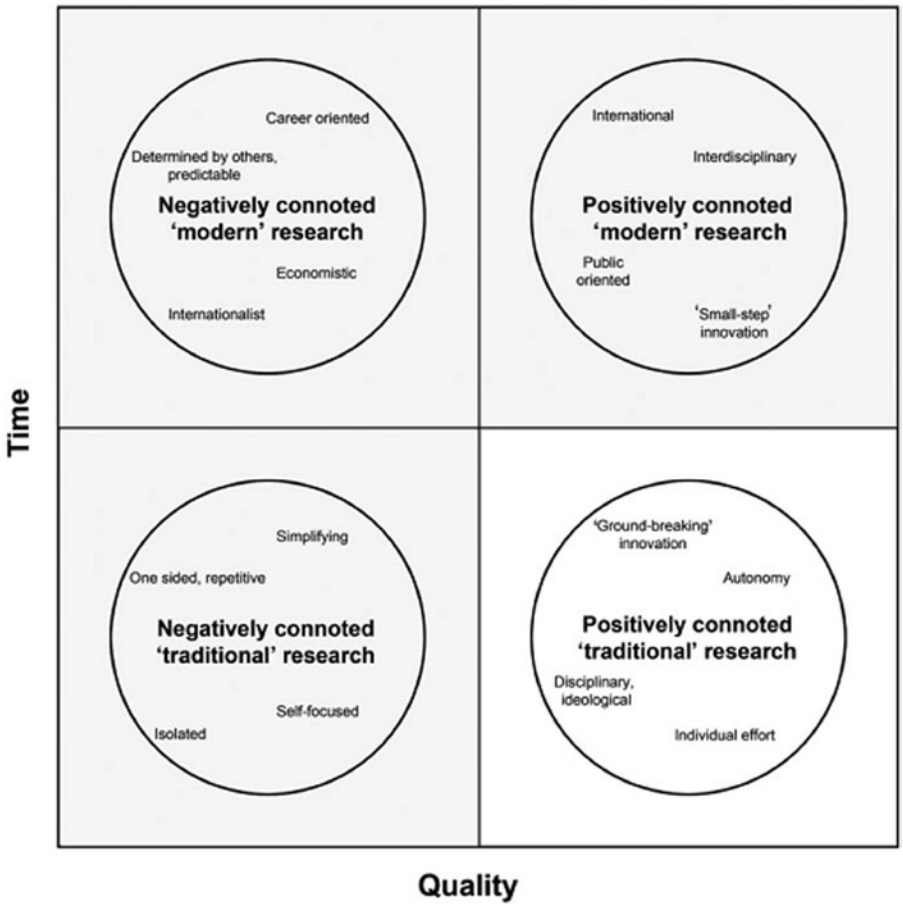

Fig. 2 The four types of humanities research. Summarizing, two-dimensional representation of commonalities across the disciplines. (Source: Ochsner et al. 2013, p. 86)

well as negative connotations. Thus, there is no clear preference for either (the 'traditional' conception received slightly more positive ratings). This analysis resulted in the identification of 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 'modern' research that is characterized by pragmatism, career aspirations, economization, and pre-structuring.

Additionally, we were able to identify two kinds of innovation connected to the two conceptions of research: 'modern' research is linked to 'small-step' innovation that finds strong reception and starts out from and ties into existing knowledge, whereas 'traditional' research is associated with 'ground-breaking' innovation that

can cause structural change but might not yet be crowned by success. Moreover, we discovered that some constructs which are commonly used as quality criteria in evaluations are double-edged in nature. Interdisciplinarity, social orientation, cooperation, and internationality are found in both the positively connoted and the negatively connoted 'modern' conception of research. At the same time, the opposites disciplinarity—interdisciplinarity, individual research—cooperation, and autonomy—social orientation reflect the differences between 'traditional' and 'modern' research but are not indicative of research quality. Since there is no difference in the perception of quality between 'modern' and 'traditional' research, it is important that evaluations take into account both of these conceptions of research. If only quality criteria for the 'modern' conception of research are used, there is the danger of sacrificing 'groundbreaking' research.

Besides these general observations about the scholars' perceptions of research quality, we were also able to derive quality criteria from the repertory grid interviews. Alongside more common quality criteria for the humanities (e.g., *innovation*, *rigour*, *connection to society*), some less-known quality criteria emerged (e.g., *continuity*, *inspiration*, *topicality*, *openness and integration*, *connection between research and teaching*, and *intrinsic motivation*). For a detailed description of the results of the repertory grid interviews, see Ochsner et al. (2013).

The quality criteria extracted from the repertory grid interviews were generated from a small sample of scholars from Switzerland only. In order to validate the results, we conducted a three-round Delphi survey with a large international group of scholars. The quality criteria derived from the repertory grid study complemented with quality criteria found in the pertinent literature on quality criteria for humanities research formed the basis for the first Delphi round.

4 Delphi survey

Delbecq et al. (1975) describe Delphi as a 'method for the systematic solicitation and collection of judgments on a particular topic through a set of carefully designed sequential questionnaires interspersed with summarized information and feedback of opinions derived from earlier responses' (p. 10). The advantages of the method are inter alia that, firstly, larger groups can be surveyed than can be handled in face-to-face group discussions; secondly, subjective judgments can be used on a collective basis when a problem does not lend itself to precise analytical tools (see Linstone and Turoff 1975, p. 4). The Delphi method is flexible and adaptable, and a valid, widely used and recognized instrument (Hsu and Sandford 2007; Landeta 2006).

The 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). For the first Delphi round in Spring 2010, the panel was comprised of 581 scholars and was subsequently updated in Winter 2010–2011 for the second and third Delphi round, finally encompassing 664 scholars (for a detailed description of the panel, see Hug et al. 2013).

We designed the Delphi survey as follows. The first round aimed at completing the quality criteria derived from the repertory grid interviews and the literature; in the second round, the scholars rated the criteria aspects; in the third round, they rated quantitative indicators attached to the criteria aspects.

The questionnaire of the first round consisted of 17 quality criteria that were further specified by 49 aspects. The participants had to (1) tick those of the 49 aspects that they personally thought fit or matched the criteria, (2) add aspects to a criterion if they thought that, according to their personal perspective, something was missing, (3) name indicators that quantify or measure a certain aspect, and (4) name additional criteria and corresponding aspects and indicators if the scholars felt something important was missing. Because of the qualitative nature of the study, the questionnaire was administered to a subset of the panel resulting in a sample of 180 persons: 30 from each of the three disciplines at Swiss universities and 30 from each of the three disciplines at LERU universities. The sample was generated using a stratified sampling procedure with each discipline at a university as a stratum.

The overall response rate of the Delphi's first round was 28 %. Because none of the 49 aspects was clearly disapproved, all aspects were retained for the second Delphi round. In addition, the participants named new aspects for the criteria provided in the questionnaire and generated new criteria. This resulted in a comprehensive list of 19 criteria for good research specified by 70 aspects. For example, the criterion *scholarly exchange* is specified by the three aspects 'disciplinary exchange', 'interdisciplinary exchange', and 'international exchange'; the criterion *recognition* is specified by five aspects, namely 'insights are recognized by the research community', 'insights are recognized by society', 'reputation within research community', 'reputation in society', and 'reputation at own university'. Table 1 lists the 19 quality criteria for research in the humanities (for a detailed description of the criteria and aspects, see Hug et al. 2013). Ten out of the 19 criteria for good research are well known and commonly used in various evaluation schemes. However, the participants also named nine criteria that are not, or at least not frequently, employed: *fostering cultural memory*, *reflection/*

Table 1 All criteria with an indication of consensuality in the three disciplines German literature studies, English literature studies and art history

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. Self-management, independence$^{GLS, ELS}$	

GLS criterion is consensual in German literature studies, *ELS* criterion is consensual in English literature studies, *AH* criterion is consensual in art history

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*.

The results of the first round of the Delphi survey served as the online questionnaire for the second round, which consisted of the 19 quality criteria specified by 70 aspects. In the second round, the participants were asked to rate the 70 aspects of the quality criteria. For that purpose we formulated a clear statement for each aspect that the respondents had to rate on a scale 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'. A statement consisted of a generic part (i.e., 'My research is assessed appropriately, if the assessment considers whether I …') and an aspect (e.g., '… introduce new research topics.') of a criterion (e.g., *innovation/originality*). The online questionnaire was administered in English and German to the whole panel (i.e., 664 scholars) from March to April 2011 (for a detailed description of the questionnaire and the translation process, see Hug et al. 2013).

With 196 returned questionnaires, the overall response rate of the second round was 30 %. In compliance with the inside-out approach of the framework, we analysed the returned questionnaires separately for each discipline. In all three disciplines, no criterion was completely rejected, since every criterion had at least one aspect with a mean score of '4' and a median of '4' or higher. Only one aspect received a median of less than '4' in all three disciplines (i.e., 'research has its impact mainly in teaching' specifying the criterion *connection between research and teaching/scholarship of teaching*). Therefore, the results indicate that in all three disciplines, every criterion and almost every aspect is seen as appropriate by at least 50 % of the respondents to assess their own research. Thus, the catalogue of criteria and aspects elaborated in the first Delphi round aptly reflects the scholars' understanding of research quality. In order to identify those criteria for assessing research quality that find acceptance in the research community, we identified consensual criteria in each discipline. We classified a criterion as consensual when at least one of its aspects was clearly approved by a majority (i.e., at least 50 % of the discipline's respondents rated the aspect at least with a '5') and disapproved only by very few scholars (i.e., not more than 10 % of the discipline's respondents rated the aspect with a '1', '2' or '3'). Eleven criteria reached consensus in all three disciplines, thus building a set of shared criteria. Note, however, that not all of these criteria are specified with the same consensual aspects in the three disciplines. For example, the criterion *scholarly exchange* was specified differently in the three disciplines: In GLS, two aspects of this criterion reached consensus: 'disciplinary exchange' and 'interdisciplinary exchange'; in ELS, the two aspects 'disciplinary exchange' and 'international exchange' reached consensus; and in AH, all three aspects reached consensus: 'disciplinary exchange', 'interdisciplinary exchange', and 'international exchange'. 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 1 indicates the consensuality of the criteria in the respective disciplines. For a detailed description and an indication of consensuality of the aspects in the respective disciplines, see Hug et al. (2013).

For the third Delphi round, we collected indicators for the aspects that were consensual in at least one discipline. Even though there is a wealth of literature on research assessment and indicators of research performance, there is no canon on indicators for research quality in the humanities because of the focus on the natural and life sciences (see, e.g., Hemlin 1996, p. 53). Therefore, we had to collect the indicators ourselves. In the first step, we conducted extensive literature research 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. In order to find as many arguments, criteria, and indicators as possible in the first step, we included a broad range of documents spanning from bibliometric and scientometric literature and government or institutional reports on how humanities are evaluated to grey literature on critiques of those procedures by humanities scholars. This resulted in a bibliography of literature on quality criteria and indicators for humanities research that is accessible on the project's webpage http://www.psh.ethz.ch/crus/bibliography (Peric et al. 2012). In the next step, we expanded our list of indicators by gathering indicators directly from humanities scholars themselves during the repertory grid interviews and in the first round of the Delphi survey. This step assured that the collection of indicators was tailored to the humanities. To our own surprise, we found an abundance of indicators, some very specific, some more vague. Because the participants had to rate the indicators in the third Delphi round, we grouped the indicators into clusters to obtain a workable amount of items. The grouping of the indicators followed two principles: Firstly, the indicators in one group should be of a similar kind (e.g., h-index, number of citations or crown indicator and the like are grouped together as 'citations'; books, articles, monographs, edited books, historical critical editions, art work, documentary films and the like are grouped together as 'publications'). Secondly, according to our measurement model, it should be possible to connect the group to a specific consensual quality criterion or aspect from the catalogue developed in the previous Delphi rounds. This resulted in 62 groups of indicators for research quality in the humanities. For a comprehensive list of the scanned documents as well as a description of the indicator groups, see Ochsner et al. (2012).

Each of these groups of indicators can be assigned to one or more aspects of the quality criteria. We assigned a group to an aspect if the occurrence of an aspect can be deduced from the indicator(s) of the group. For example, the aspect 'documentation of aspects of the past' of the criterion *fostering cultural memory* can potentially be measured by the following four groups of indicators: 'number, weighting and duration of documentation or preservation activities'; 'number and weighting of outputs reflecting documentation or preservation activities'; 'number and weighting of activities for the public (e.g., guided tours, public lectures, readings, media appearances, performances)'; 'number and weighting of outputs for the public (e.g., popular books or articles, exhibitions, documentary films)'. For some aspects, we were not able to identify indicators. For example, another aspect of the criterion *fostering cultural memory* (i.e., 'renewal of interpretations of aspects of the past') cannot be measured and it is only accessible by the judgement of peers.

Having assigned the indicators to the aspects of the quality criteria, we can quantify the amount of aspects that can be measured quantitatively. We were only able to assign indicators to 23 of the 42 aspects that were consensual in at least one discipline. This corresponds to a share of 55 % of aspects that can be measured quantitatively. The share of quantifiable aspects is slightly lower when each discipline is analysed separately: In GLS, we identified indicators for 53 % of the consensual aspects; in ELS, 52 % of the consensual aspects are potentially measurable; and in AH, we were able to assign indicators to 48 % of the consensual aspects. In other words, indicators can only capture about 50 % of the humanities scholars' notions of quality.

The questionnaire of the third Delphi round was conceptualized in a very similar way to the questionnaire of the second round. The scholars had to rate the groups of indicators 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'. 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*). The scholars rated the indicator groups that were assigned to the given aspect with respect to the statement. Because every discipline had its own set of consensual aspects, the questionnaires differed between the disciplines. In GLS, the respondents had to rate 86 items consisting of 59 unique indicator groups assigned to 19 aspects; in ELS, scholars had to rate 85 items consisting of 45 unique indicator groups assigned to 15 aspects; and in AH, the participants had to rate 74 items consisting of 44 unique indicator groups assigned to 15 aspects. Just as in the second round, the online questionnaire was administered in either English or German to the whole panel (i.e., 664 scholars). The field period lasted from October 2011 to January 2012.

A total of 133 scholars returned the questionnaire. This corresponds to an overall response rate of 20 %. However, after the same period of time as in the second round (34 days), the response rate was only 11 %. Hence, we extended the field period until we did not receive any responses for four consecutive days (extension of 45 days including Christmas and New Year). Overall, most items were approved by at least 50 % of the respondents (In GLS, 93 % of the items were rated with not less than a'4' by at least 50 % of the respondents; ELS: 91 %; AH: 97 %). However, if indicators are to be used in research assessments, the affected scholars need to accept the indicators. Therefore, we identified the consensual indicator groups in each discipline. We classified an indicator group as consensual the same way we classified an aspect as consensual, that is, if the indicator group was clearly approved by a majority (i.e., at least 50 % of the discipline's respondents rated the item at least with a '5') and disapproved by very few scholars (i.e., 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 is classified as consensual (1 %); and in AH, 16 indicator groups are consensual (22 %). Table 2 lists the consensual indicator groups along with the aspects they measure.

Table 2 Consensual indicators along with the aspects they measure for German literature studies, English literature studies, and art history

Name of Group	Definition of Group	Aspect
German Literature Studies		
Publications	Number and weighting of publications for a disciplinary audience (e.g., monographs, article in edited volume, exhibition catalogue, art work, documentary monographs, article in edited volume, exhibition catalogue, art work, documentary film)	Disciplinary exchange
Collaborations	Number, weighting and duration of collaborations within my discipline (e.g., joint research projects with other institutions, co−authorship with peer, membership in a research network)	Disciplinary exchange
Presentations	Number and weighting of presentations for a disciplinary audience	Disciplinary exchange
Collaborations	Number, weighting and duration of interdisciplinary collaborations (e.g., joint research projects with institutions from other disciplines, co−authorship with researchers from other disciplines, membership in an interdisciplinary research network)	Interdisciplinary exchange
Publications	Number and weighting of publications for an interdisciplinary audience (e.g., monographs, article in edited volume, exhibition catalogue, historical critical edition, art work, documentary film)	Interdisciplinary exchange
Organized events	Number and weighting of organized events for an interdisciplinary audience (e.g., colloquium, series of lectures, exhibition, conference)	Interdisciplinary exchange
Output of documentation activities	Number and weighting of outputs reflecting documentation and preservation activities	Documentation of aspects of the past
Teaching	What I offer in teaching (e.g., teaching hours, the time that I spend in helping and guiding junior researchers; my participation in a graduate program, graduate school or comparable program; the number and quality of further training courses I offer)	Promotion of young academics
External education	External education of junior researchers (e.g., research stays of junior researchers at other institutions; number of external further training these junior researchers have attended; the financial resources I make available to them for attending congresses or receiving additional training)	Promotion of young academics
Assessed openness	Assessment of my openness by students and junior researchers	Openness to other persons
English Literature Studies		
Publications	Number and weighting of publications for an international audience (e.g., monographs, article in edited volume, exhibition catalogue, historical critical edition, art work, documentary film)	International exchange
Art History		
Publications	Number and weighting of publications for a disciplinary audience (e.g., monographs, article in edited volume, exhibition catalogue, art work, documentary monographs, article in edited volume, exhibition catalogue, art work, documentary film)	Disciplinary exchange
Organized events	Number and weighting of organized events for a disciplinary audience (e.g., colloquium, series of lectures, exhibition, conference)	Disciplinary exchange
Publications	Number and weighting of publications for an interdisciplinary audience (e.g., monographs, article in edited volume, exhibition catalogue, historical critical edition, art work, documentary film)	Interdisciplinary exchange

Table 2 (continued)

Name of Group	Definition of Group	Aspect
Collaborations	Number, weighting and duration of international collaborations (e.g., joint research projects with institutions from other countries, co−authorship with researchers from other countries, membership in an international research network)	International exchange
Presentations	Number and weighting of presentations for an international audience	International exchange
Publications	Number and weighting of publications for an international audience (e.g., monographs, article in edited volume, exhibition catalogue, historical critical edition, art work, documentary film)	International exchange
Organized events	Number and weighting of organized events for an international audience (e.g., colloquium, series of lectures, exhibition, conference)	International exchange
Survey: renewal of interpretations	Survey of students, alumni and the public	Renewal of interpretations of aspects of the past
Research topics	Number of research topics, approaches, theories, methods, materials, disciplinary areas and languages that I use (e.g., evident in the bibliography of my publications and presentations, information on my research website)	Contributing towards variety and diversity
Discussions/ debates	Number and weighting of participation, organisation or moderation of disputes, debates or discussions about research	Engaging in ongoing research debates
Assessed openness	Assessment of my openness by students and junior researchers	Openness to other persons
Opportunities for junior researchers	Career opportunities for junior researchers (e.g., number of positions for junior researchers, number of publications by junior researchers who have been my students, number of co−authorships with junior researchers)	Openness to other persons
Sources	Number of sources, materials and original works used in publications or presentations	Rich experience with sources
Qualification of junior researchers	Qualification of students and junior researchers (e.g., number of bachelor/master/doctoral degrees; success rate (appointments to a professorship) of former students; drop-out rate of students and junior researchers; survey of alumni about the skills/competencies/ qualifications they acquired)	Arouse passion for research
Attractivity to juniorresearchers	Attractivity to junior researchers (e.g., number of Ph.D. students I have, postdoctoral researchers and researchers from abroad I have; number of participants in my courses and lectures)	Arouse passion for research
Research orientation of teaching	Student satisfaction with the research orientation of the courses	Research-based teaching

In addition to the rating of the indicator groups, the participants were asked if they think that it is conceivable that experts (peers) could evaluate the participants' own research performance appropriately based only on the quantitative data that the participants had just rated. This question was clearly rejected by the respondents of all three disciplines (GLS: 88 %; ELS: 66 %; AH: 89 %).

5 Integration and discussion of the results of the four studies

In order to integrate the findings of our studies, we first combined the results of the three rounds of the Delphi survey and then related them to the findings of the repertory grid.

In light of the fact that many projects that develop instruments or tools for research assessments in the humanities face refusal or a dead-end (as, e.g., the *Forschungsrating* of the German *Wissenschaftsrat* that was rejected by the *Verband der Historiker und Historikerinnen Deutschlands* [Association of German Historians, VHD] or the ERIH project of the ESF that reached a dead end), we expected strong opposition by the humanities scholars. Yet, the first two rounds of the Delphi survey were received astonishingly well. Both reached a response rate of about 30 %. In view of the amount of time the scholars had to put into filling in the questionnaires—especially in the first round—30 % can be considered a comparatively high response rate. 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). The comments on the first two Delphi rounds that we received from the scholars either by email—as a reaction to the invitation—or by using a free text field in the survey were quite positive as well. In the first round, six out of the eight scholars who commented on the survey (a total of 9 comments) delivered clearly positive statements (75 %) and two scholars sent clearly negative statements (25 %). In the second round, 35 % of the 34 scholars who commented on the survey (a total of 34 comments) provided clearly positive feedback and 29 % provided clearly negative feedback. The clearly positive comments of the first two rounds often pointed to the fact that it was an important topic (e.g., 'I do find this survey excellently conceived, with most pregnant issues') and that the explication of the quality criteria was useful to them (e.g., 'some time ago I completed your questionnaire and, to my surprise, found the questions and issues interesting and stimulating. I would like to show my master and doctoral students which criteria can play a potential role in the assessment of humanities research […]' [own translation]). The clearly negative comments often questioned the survey's objective (e.g., 'it is not necessary').

However, in the third Delphi round that focused on indicators, only 11 % of the scholars participated in the survey during the same timeframe as in the first two rounds and after a significant expansion of the field period, the response rate still did not exceed the 20 % mark. We also received more negative comments on the third Delphi round than during the first two rounds. Only 22 % of the 27 scholars (32 comments in total) voiced clearly positive statements and 78 % of the scholars sent clearly negative comments. It is remarkable that 47 % of these negative statements commenting on the third Delphi round emphasize negative attitudes towards quantification (e.g., 'What troubles me, and what caused me to "disagree" with so much was the use of the word "number" in almost all of the criteria mentioned. The qualitative aspects of research cannot be reduced to a quantity […]'). The positive statements voiced general support for the object of the survey (e.g., 'Anything that challenges the stranglehold of growth fundamentalism and institutionalised distrust on humanities funding deserves full support').

Table 3 Comparison of the ratings of the quality aspects (Delphi round 2) and the indicator groups (Delphi round 3) by discipline

	Second Delphi round aspects			Third Delphi round indicator groups		
	GLS	ELS	AH	GLS	ELS	AH
Number of items	70	70	70	86	85	74
Grand mean	4.71	4.56	4.64	4.14	3.92	4.4
Standard deviation of grand mean	0.59	0.64	0.56	0.5	0.38	0.43
Minimum mean	3.34	2.88	3.15	2.77	2.91	3.19
Maximum mean	5.74	5.56	5.6	5	5.02	5.21
Median of means	4.8	4.7	4.78	4.21	3.93	4.46
Items with mean ≥4	60	57	60	60	32	62
Percentage of mean ≥4	86%	81%	86%	70%	38%	84%
Items with median ≥4	68	65	68	80	77	72
Percentage of median ≥4	97%	93%	97%	93%	91%	97%
Number of consensual items	36	29	31	10	1	16
Percentage of consensual items	51%	41%	44%	12%	1%	22%

Items were rated on a scale ranging from '1' to '6' where ratings from '1' to '3' indicate a disagreement with the item and ratings from '4' to '6' indicate an agreement with the item. *GLS* German literature studies, *ELS* English literature studies, *AH* art history

A comparison of the ratings of the aspects and indicators presents a similar picture (see Table 3). In all disciplines, the grand mean (i.e., the mean of the aspect and indicator means) is clearly lower for the indicators than for the aspects. If we look at the percentage of aspects or indicators that scored a mean of '4' or higher (a rating of '4' corresponds to an approval of the aspect or indicator), the same picture emerges: In all disciplines, the share of aspects that received a positive mean was higher than the share of indicators receiving a positive mean (in AH, the difference is not very pronounced). The same holds true if we look at the share of consensual aspects and indicators. In GLS, 51% of the aspects reached consensus whereas only 12% of the indicator groups reached consensus. In ELS, 41% of the aspects are consensual in contrast to only 1% of the indicator groups. In AH, 44% of the aspects were qualified as consensual, but only 22% of the indicator groups. However, there is no difference between the share of aspects and indicators which were approved by at least 50% of the respondents: Most aspects as well as most indicators were approved by a majority.

Despite these observations, it is not necessarily valid to conclude that humanities scholars are strictly against quantification or—to make an even bolder statement—that indicators cannot provide meaningful information on research performance. Firstly, it is surprising that 20% of the humanities scholars actually participated in the third Delphi round which focused on quantitative indicators after having already been asked to participate in two previous surveys on quality criteria. At the same time, most indicators were even approved by a majority. Secondly, we also received two slightly positive comments on quantification in the second and third waves, which correspond to 13% of all the comments on quantification.

But we can conclude that humanities scholars prefer a qualitative approach to research evaluation and that indicators must be linked to their notions of quality: Humanities scholars are willing to think about research quality and develop quality

criteria if a bottom-up approach is applied and a clear link between the scholars' own research and the quality criteria is established (Delphi rounds one and two). However, when it comes to rating quantitative indicators for research quality, humanities scholars are more critical (Delphi round three). A comment directed at a specific indicator illustrates the critical but not necessarily disapproving stance towards the use of indicators: '[…] I think that the quantitative criterion works very well for people at the top of their field who have been professors for some ten years or more […]. For younger researchers, these criteria are no good, except perhaps publication in refereed journals—so for younger people potential cannot be measured quantitatively […]. There is another exception, namely those visionary people who do not publish a lot and withdraw from associations and get on with writing the great stuff, which may not be much in terms of quantity. Those are getting rarer because of the continual evaluation being practised already as part of the economization of the universities. Yet it is these people who really are the most important […]'.

Why are (some) humanities scholars critical of indicators, even though many scholars would agree with the use of some quantitative indicators that measure certain aspects of research quality? While there are plenty of reasons for a critical stance towards research indicators, our studies point to two possible reasons that have not received much attention to date: firstly, a mismatch of the quality criteria and indicators between evaluators and scholars (Hug et al. 2013, p. 9; Ochsner et al. 2012, pp. 3–4) and, secondly, the double-edged nature of some frequently used indicators (Ochsner et al. 2013, p. 86). The mismatch, as the first possible reason, can be described as follows: On the one hand, not all criteria that are frequently used in evaluation schemes are shared by humanities scholars (e.g., *social impact*, *reputation*, *productivity*); and on the other hand, humanities-specific criteria are not known or not used in evaluation schemes (e.g., *fostering cultural memory*, *reflection/criticism*, *scholarship/erudition*). Additionally, the most frequently used indicators (e.g., citations, prizes, third-party funding, transfers to economy and society) measure quality criteria and aspects that are not consensual in all three disciplines (i.e., *recognition*, *impact on research community*, *relevance*, *relation to and impact on society*; see Ochsner et al. 2012, pp. 3–4).

The repertory grid interviews reveal the second possible reason (i.e., the double-edged nature of some frequently used indicators): For example, interdisciplinarity, cooperation, public orientation, and internationality are indicators of the 'modern' as opposed to the 'traditional' conception of research and not necessarily related to quality. If, for example, interdisciplinarity serves someone's career or is used as an end in itself or as lip service to get funding, it is clearly negatively connoted; when it serves diversity, it is positively connoted. By applying these indicators, the positively connoted 'traditional' conception of research would be forced to the back seat. However, the 'traditional' conception of research is highly regarded within the research community. Evaluators must not confuse the dichotomy of the 'traditional' and 'modern' conception of research with 'old-fashioned/conservative' and 'new/innovative/promising' research, respectively. Both conceptions of research can stand for innovative and promising research. Whereas research according to the 'modern' conception serves the direct needs of the public and adheres to contemporary paradigms, the 'traditional' conception of research comprises critical thinking and taking new directions.

Using the repertory grid and a Delphi survey to develop quality criteria genuine to humanities research, we were able to identify indicators that also reflect the 'traditional' conception of research (e.g., the indicator group 'Number of sources, materials and original works used in publications or presentations' that measures the aspect 'Rich experience with sources' from the criterion '*scholarship/erudition*'). If humanities research is to be assessed appropriately, it is important to include indicators that also represent the 'traditional' conception of research. However, there is one caveat concerning the measurement of the 'traditional' conception of research: It is an open question whether the purely 'traditional' conception of research that ideally brings about 'ground-breaking' innovation and paradigm change can be measured prospectively at all. The repertory grid interviews point clearly towards the prerequisite of autonomy for such achievements. Quantitative assessments, publication pressure or economization as such are explicit characteristics of the 'modern' conception of research, more specifically, the negatively connoted 'modern' conception of research (see Ochsner et al. 2013, pp. 91–92). A measurement of some characteristics of the 'traditional' conception of research could, on the one hand, help to make visible important contributions of humanities scholars that otherwise might be overlooked and, on the other hand, help to promote humanities-specific notions of quality. However, a measurement of research performance might never capture the essence of the 'traditional' conception of research of the individual researcher who brings about important changes to the scientific community by undertaking disciplinary research locked up in his study. Having in mind this ideal of the erudite scholar, many humanities scholars will most likely be critical of the use of indicators and disapprove of purely indicator-based assessments.

We can conclude from the results of our studies that an assessment of research performance by means of *indicators* will be met with some resistance in the humanities: We have found that (1) only about 50 % of those quality criteria and aspects which are rated as most important can be measured with quantitative indicators. As long as 50 % of the most relevant criteria and aspects cannot be measured with indicators, humanities scholars will be very critical of purely quantitative approaches to research assessments; (2) while most indicators are accepted for use in peer review-based assessments, a minority, who are not to be underestimated, does not approve of the use of most indicators; (3) some indicators that are often used in evaluation schemes are not measuring the research quality but differentiate between more 'traditional' and more 'modern' research, both of which can be of high quality and importance; and (4) purely indicator-based research assessments are disapproved of by a vast majority of the humanities scholars.

However, we can also conclude that concerning an assessment of research performance by means of *quality criteria* humanities scholars are willing to think about quality and take part in the development of quality criteria if a bottom-up approach is chosen, and a performance assessment on the basis of relevant criteria is possible if the humanities scholars are involved. We have found that (1) a broad range of quality criteria has to be applied to adequately assess research quality in the humanities; (2) there are shared criteria that are consensual in all disciplines that have been studied; (3) the disciplines should not be lumped together as we found discipline-specific criteria; and (4) with a certain amount of care, research indicators linked to the relevant criteria can be used to support the experts in research assessments (informed peer review).

6 Conclusion

The assessment of research performance in the humanities is an intricate and highly discussed topic. Many problems have yet to be solved, foremost the question of the humanities scholars' acceptance of evaluation tools and procedures. Currently, different initiatives are investigating ways to make the quantity and quality of humanities research visible. Some focus on the building or expansion of databases or the building of rankings and lists of journals or publishers, for example the RESH and DICE databases (Evaluation of Scientific Publications Research Group 2010, pp. 11–13; 2012), the VABB-SHW database, (Engels et al. 2012), the CRISTIN database (Schneider 2009; Sivertsen 2010), the ERIH project (European Science Foundation 2009), the MESUR project (National Science Foundation 2009), the Book Citation Index (Thomson Reuters 2011), Libcitations (White et al. 2009), and a label for peer-reviewed books (Verleysen and Engels 2013). Others focus on the development of evaluation procedures, for example the Research Rating of the German Council of Science and Humanities (Wissenschaftsrat 2011a, 2011b), the Quality Indicators for Research in the Humanities (Royal Netherlands Academy of Arts and Sciences 2011), and the ERA initiative (Australian Research Council 2012). A third way is to research the peculiarities of the humanities in citation and publication behaviour (e.g., Hammarfelt 2012; Nederhof 2011; Zuccala 2012) or productivity (e.g., Hemlin 1996; Hemlin and Gustafsson 1996). And finally, some investigate peer review processes and peer reviewers' quality criteria in the humanities (e.g., Guetzkow et al. 2004; Lamont 2009).

We add a different approach by explicating the scholars' notions of quality and linking indicators to the quality criteria that are generated in a bottom-up procedure from within the humanities based on the notions of quality of the scholars. Following our framework for developing quality criteria for research in the humanities and using the repertory grid and Delphi method, we explicated the humanities scholars' notions of quality, linked indicators to the humanities scholars' quality criteria according to a measurement model, and identified quality criteria and indicators that are consensual in each of the three disciplines covered in this project (i.e. German literature studies, English literature studies, and art history). In all steps, we followed an inside-out approach to ensure quality criteria and indicators were tailored to humanities research.

Drawing from the four studies we conducted in this project, we can formulate limitations and opportunities of assessments of research performance in the humanities. While an assessment by means of *indicators* exhibits some limitations, an assessment by means of *quality criteria* derived from humanities scholars' notions of quality presents opportunities to make humanities research quality visible.

As long as only about 50% of the criteria and aspects derived from the humanities scholars' notions of quality are measurable by quantitative indicators, humanities scholars will be critical of purely *quantitative assessments* of research performance. Moreover, the most frequently used quantitative indicators in assessment schemes reflect quality criteria that are not consensual in the humanities (see Ochsner et al. 2012, pp. 3–4). Besides, these commonly used indicators measure the 'modern' conception of research, (see Ochsner et al. 2013, p. 86) while the humanities scholars emphasize the importance of the 'traditional' conception of research as well, which is characterized by attributes like autonomy, individual effort or having no specific addressee

(see Ochsner et al. 2013, p. 85). However, while purely indicator-based assessments are clearly disapproved, an assessment of research performance by means of *quality criteria* is possible if a bottom-up approach is chosen and the humanities scholars are involved in the definition of the quality criteria. A broad range of quality criteria needs to be applied if humanities research is to be adequately assessed. Besides general quality criteria, discipline-specific criteria need to be applied. Using the consensual quality criteria specific to each discipline opens the opportunity to link research indicators to criteria and aspects that are relevant to the humanities scholars. These indicators can be used to support the experts in research assessments (informed peer review).

Four limitations of our studies need to be kept in mind and can be reformulated as desiderata of further research. In all our studies, we focused on the three disciplines German literature studies, English literature studies, and art history. While these disciplines offer a good picture of disciplines that elude bibliometric approaches to assessment, more disciplines must be investigated to develop assessment tools that are adequate for a broader range of humanities disciplines. Furthermore, the response rates were adequate and relatively high given the composition of the panel and the intricate nature as well as the topic of the questionnaire. However, our results are based on slightly less than one third of the panel. Therefore, studies that include a broader range of scholars are needed to minimize or exclude selection bias. Moreover, our studies present humanities scholars' notions of quality and their acceptance of indicators reflecting these notions of quality. Yet, research assessments involve many more stakeholders. Our approach could be used to find quality criteria and indicators for other stakeholders. Finally, the feasibility of the implementation of a research assessment according to our framework needs to be investigated.

Funding This work was supported by the Rectors' Conference of the Swiss Universities (CRUS) in the framework of the cooperative project 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 project were provided by the University of Zurich.

Acknowledgements The authors thank Esther Germann for proofreading.

References

Andersen, H., Ariew, R., Feingold, M., Bag, A. K., Barrow-Green, J., van Dalen, B., … Zuidervaart, H. (2009). Editorial: Journals under threat: A joint response from history of science, technology and medicine editors. *Social Studies of Science, 39*(1), 6–9. doi:10.1177/0306312709039001070 2.

Australian Research Council. (2012). *The Excellence in Research for Australia (ERA) Initiative*. http://www.arc.gov.au/era/. Accessed 29 Sept 2013.

Braun, N., & Ganser, C. (2011). Fundamentale Erkenntnisse der Soziologie? Eine schriftliche Befragung von Professorinnen und Professoren der deutschen Soziologie und ihre Resultate [Fundamental knowledge of sociology? A written survey of male and female professors of German sociology and its results]. *Soziologie, 40*(2), 151–174.

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

Delbecq, A. L., Van de Ven, A., & Gustafson, D. H. (1975). *Group techniques for programm planning. A guide to nominal group and Delphi processes*. Glenview: Scott, Foresman.

Engels, T. C. E., Ossenblok, T. L. B., & Spruyt, E. H. J. (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. (2009). *The European Reference Index for the Humanities: A reply to the criticism*. http://www.universitaetsverlagwebler.de/inhalte/qiw-3%2B4-2009.pdf. Accessed 29 Sept 2013.

Evaluation of Scientific Publications Research Group. (2010). *Difusión y Calidad Editorial delas Revistas Españolas de Humanidades y Ciencias Sociales y Jurídicas (DICE)*. http://dice.cindoc.csic.es/. Accessed 29 Sept 2013.

Evaluation of Scientific Publications Research Group. (2012). *Revistas Españolas de Ciencias Sociales y Humanidades (RESH)*. http://epuc.cchs.csic.es/resh/. Accessed 29 Sept 2013.

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

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

Frey, B. S., Humbert, S., & Schneider, F. (2007). Was denken deutsche Ökonomen? Eine empirische Auswertung einer Internetbefragung unter den Mitgliedern des Vereins für Socialpolitik im Sommer 2006. *Perspektiven der Wirtschaftspolitik, 8*(4), 359–377. doi:10.1111/j.1468-2516.2007.00256.x.

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

Giménez-Toledo, E., Tejada-Artigas, C., & Mañana-Rodríguez, 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.

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

Guillory, J. (2005). Valuing the humanities, evaluating scholarship. *Profession (MLA: Modern Language Association), 1,* 28–38. doi:10.1632/074069505X79071.

Hammarfelt, B. (2012). Harvesting footnotes in a rural field: Citation patterns in Swedish literary studies. *Journal of Documentation, 68*(4), 536–558.

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.

Hemlin, S., & Gustafsson, M. (1996). Research production in the arts and humanities: A questionnaire study of factors influencing research performance. *Scientometrics, 37*(3), 417–432.

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: 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. 473–496). Dordrecht: Kluwer Academic.

Hsu, C. C., & Sandford, B. A. (2007). The Delphi technique: Making sense of consensus. *Practical Assessement, Research & Evaluation, 12*(10), 1–8.

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), 55–68. http://www.psh.ethz.ch/research/publications/ijelp_inpress.pdf. Accessed 29 Sept 2013.

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: 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: Campus.

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

Landeta, J. (2006). Technological forecasting and social change. Current validity of the Delphi method in social sciences. *Technological Forecasting and Social Change, 73*(5), 467–482. doi:10.1016/j.techfore.2005.09.002.

Linstone, H. A., & Turoff, M. (1975). *The Delphi method: Techniques and applications*. Reading: Addison-Wesley.

National Science Foundation. (2009). *MESUR project*. http://mesur.informatics.indiana.edu/. Accessed 29 Sept 2013.

Nederhof, A. J. (2006). Bibliometric monitoring of research performance in the social sciences and the humanities: A review. *Scientometrics*, *66*(1), 81–100.

Nederhof, A. J. (2011). A bibliometric study of productivity and impact of modern language and literature research. *Research Evaluation, 20*(2), 117–29.

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. URN: urn:nbn:de:bvb:355-157-7.

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*(2), 79–92. doi:10.1093/reseval/rvs039.

Palomares-Montero, D., & Garcia-Aracil, A. (2011). What are the key indicators for evaluating the activities of universities? *Research Evaluation, 20*(5), 353–363. doi:10.3152/095820211x13176484436096.

Peric, B., Ochsner, M., Hug, S. E., & Daniel, H.-D. (2012). *AHRABi. Arts and Humanities Research Assessment Bibliography*. ETH Zurich. http://www.psh.ethz.ch/crus/bibliography/. Accessed 29 Sept 2013.

Plumpe, W. (2009) ‚Stellungnahme zum Rating des Wissenschaftsrates aus Sicht des Historikerverbandes'. In Prinz, C. and Hohls, R. (eds.) *Qualitätsmessung, Evaluation, Forschungsrating. Risiken und Chancen für die Geschichtswissenschaften? Historisches Forum, Bd. 12,* (pp. 121–126). Clio-online und Humboldt-Universität zu Berlin: Berlin.

Rosenberger, M., & Freitag, M. (2009). Repertory grid. In S. Kühl & P. Strodtholz (Eds.), *Handbuch Methoden der Organisationsforschung: Quantitative und qualitative Methoden* pp. 477–496. Wiesbaden: VS Verlag für Sozialwissenschaften.

Royal Netherlands Academy of Arts and Sciences. (2011). *Quality indicators for research in the humanities*. Amsterdam: Royal Netherlands Academy of Arts and Sciences.

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.

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 the scientific publication output at research institutions. *ISSI Newsletter, 6*(1), 22–28.

Thomson Reuters. (2011). *The Book Citation Index*. http://wokinfo.com/products_tools/multidisciplinary/bookcitationindex/. Accessed 29 Sept 2013.

Verleysen, F.T., & Engels, T.C.E. (2013). A label for peer reviewed books. *Journal of the American Society for Information Science and Technology, 64*(2), 428–430. doi:10.1002/asi.22836.

Walker, B. M., & Winter, D. A. (2007). The elaboration of personal construct psychology. *The Annual Review of Psychology, 58,* 453–477.

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.

Wissenschaftsrat. (2011a). Forschungsrating Anglistik/Amerikanistik. http://www.wissenschaftsrat.de/arbeitsbereiche-arbeitsprogramm/forschungsrating/anglistikamerikanistik/. Accessed 29 Sept 2013.

Wissenschaftsrat. (2011b). Zum Forschungsrating allgemein. http://www.wissenschaftsrat.de/arbeitsbereiche-arbeitsprogramm/forschungsrating/. Accessed 29 Sept 2013.

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