A Framework to Explore and Develop Criteria for Assessing Research Quality in the Humanities[1]

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

Abstract

In the process of developing new tools for measuring and assessing research quality in the humanities, many challenges emerge, such as technical problems (e.g., building publication databases, capturing social impact) and scholars' opposition to measuring research performance. This paper focuses on scholars' opposition and presents the four most crucial objections of humanities scholars regarding the measurement and assessment of research quality (i.e., methods employed in research evaluation originated from the natural sciences, strong reservations against quantification, fear of negative steering effects of indicators, lack of consensus on quality criteria within humanities disciplines). We suggest a framework to explore and develop quality criteria and indicators that considers scholars' objections by adopting an inside-out approach, by relying on a sound measurement approach, by making scholars' notions of quality explicit, and by striving for consensus within a discipline/sub-discipline. Finally, we outline an empirical procedure, comprising Repertory Grid interviews and a Delphi survey, to implement the framework. The framework and its empirical implementation contribute to the development of criteria and indicators for assessing research quality in the humanities.

***Keywords**: Research evaluation, arts and humanities, quality, criteria, indicators, Delphi survey, Repertory Grid, criticism.*

Bibliometric indicators are used to compare and evaluate research performance in the life sciences and natural sciences (e.g. Forslöw, Rehn, & Wadskog, 2005; Giménez-Toledo, Roman-Roman, & Alcain-Partearroyo, 2007; Lane, 2010) and are employed in the performance-based university research funding systems of several countries (Hicks, 2012). However, bibliometric indicators are not well-suited to determine the quantity and quality of humanities' research or to assess it (Archambault, Vignola Gagné, Cote, Larivière, & Gingras, 2006; Bourke & Butler, 1996; Butler & Visser, 2006; Finkenstaedt, 1990; Glänzel & Schoepflin, 1999; Gomez-Caridad, 1999; Guillory, 2005; Hicks, 2004b; Moed, Luwel, & Nederhof, 2002; Nederhof, 2006; Nederhof, Zwaan, De Bruin, & Dekker, 1989).

Therefore, different initiatives are currently developing assessment tools for the humanities or striving to make the quantity and quality of humanities' research visible. Examples include the CRISTIN database in Norway (Schneider, 2009; Sivertsen, 2010), the Flemish Academic Bibliographic Database for the Social Sciences and Humanities (Engels, Ossenblok, & Spruyt, 2012), the European Reference Index for the Humanities (European Science Foundation, 2011b), the MESUR project (National Science Foundation, 2009), the Book Citation Index (Thomson Reuters, 2011), Libcitations (White et al., 2009), the assessment of monographs through their publishers (Giménez-Toledo & Roman-

1 This paper was supported by the Rectors' Conference of the Swiss Universities (CRUS) as part of the project "Developing and Testing Research Quality Criteria in the Humanities, with an emphasis on Literature Studies and Art History".

* Swiss Federal Institute of Technology Zurich (ETH Zurich), Professorship for Social Psychology and Research on Higher Education, Zurich (Switzerland);

** University of Zurich, Evaluation Office, Zurich (Switzerland). Correspondence concerning this article should be addressed to: Sven E. Hug, ETH Zurich, D GESS, Muehlegasse 21, 8001 Zurich, Switzerland, Phone: +41 44 632 46 85, Fax: +41 44 632 12 83, E-mail: sven.hug@gess.ethz.ch

Roman, 2009), the Research Rating of the German *Wissenschaftsrat* in American Studies and English Studies (Wissenschaftsrat, 2011a), the Quality Indicators for Research in the Humanities (Royal Netherlands Academy of Arts and Sciences, 2011), Evaluating Research in Context (Netherlands Organisation for Scientific Research, 2009), the European Educational Research Quality Indicators (EERQI Consortium, 2011) and the Excellence in Research for Australia (ERA) Initiative (Australian Research Council, 2012).

In the process of developing new tools for representing, measuring and assessing research quality in the humanities, many challenges emerge, such as technical problems (e.g., building publication databases, capturing social impact) and opposition of scholars. The latter is observable in two prominent initiatives that have faced strong resistance and severe criticism from humanities scholars, namely the ERIH list of the European Science Foundation (ESF) and the pilot study of the *Forschungsrating* (research rating) of the German *Wissenschaftsrat* (WR). The ERIH project was launched in 2002 by the ESF to "represent the full range of high-quality research published in Europe in the humanities and thus also serve as a tool of access to this research" (European Science Foundation, 2011a). This was achieved in 2008 with the publication of "categorised lists of quality research journals for the humanities" (European Science Foundation, 2011a). Shortly thereafter, nearly fifty journal editors from the social sciences and humanities (SSH) asked the ESF in an open letter (e.g., Andersen et al., 2009) to remove their journal from the ERIH list as an act of protest. The ERIH Steering Committee answered by publishing a "Joint Response to Criticism" (European Science Foundation, 2009) and tried to fight further adversity by renaming the three ERIH journal categories from "A", "B", "C" into "National" and "International" (subcategorized into "INT1" and "INT2") and by revising the initial ERIH list (European Science Foundation, 2011a). After this, the ESF Standing Committee for the Humanities declared that "the next round of ERIH lists revisions has not yet been planned" (European Science Foundation, 2011a). Therefore, it seems as if the ERIH project has come to a (dead-)end. In contrast to the ERIH project, the *Forschungsrating* of the WR is still alive and underway, but with significant delays. The ultimate goal of the *Forschungsrating* is to compare and assess the research performance of all German universities (Wissenschaftsrat, 2011b). The WR successfully completed two pilot studies in 2008 in the fields of chemistry and sociology but failed to do so in the field of history because the *Verband der Historiker und Historikerinnen Deutschlands* (association of German historians, VHD) refused to take part (Plumpe, 2009; Verband der Historikerinnen und Historiker Deutschlands, 2010). Therefore, the WR devised a position paper containing general guidelines and recommendations for comparing research performance in the humanities (Wissenschaftsrat, 2010) and chose English Studies and American Studies for its pilot study. The WR intends to publish the results of the pilot study in the fourth quarter of 2012 (Wissenschaftsrat, 2011a).

I. Criticisms of Humanities Scholars

If one wants to develop tools to determine the quantity and quality of humanities' research or to assess it without facing significant delays, refusal or a dead-end like the two initiatives mentioned above, it is necessary to be aware of and account for the criticism put forward by humanities scholars. For this reason, we have conducted a thorough literature review of scholars' criticisms regarding the measurement and assessment of research quality in the humanities. The following sections discuss the four most pertinent recurring objections voiced by scholars.

A. Methods originate from the natural sciences

Vec (2009), a legal scholar, claims that "a lot of evaluation systems were modelled after the natural sciences" (p. 6, own translation) and thus, do not account for the importance of monographs, the use of national languages and the fact that research in the humanities is carried out by individual scholars. The view that characteristic research patterns and practices in the humanities are neglected is also widely shared by researchers in the fields of bibliometrics and research evaluation (for a thorough discussion see e.g. Hicks, 2004a; Nederhof, 2006; Scheidegger, 2007). Moreover, Lack (2008), a literature scholar,

asserts that existing evaluation procedures and indicators are based on a natural sciences' linear understanding of progress and, therefore, asks for tools that can cope with the humanities' conception of increasing knowledge. Lack calls these conceptions the "coexistence of competing ideas" and the "expansion of knowledge" (p. 14, own translations).

B. Strong reservations against quantification

In 2008, a group of 24 renowned philosophers wrote a joint letter to the Australian Government voicing their discontent about the ranking of academic journals in the Excellence in Research for Australia (ERA) exercise. In addition to their disapproval of journal rankings, this letter also reflects their widespread reservations regarding the quantification of research quality in the humanities: "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). Similarly, McCarthy, Ondaatje and Brooks (2004, p. 37) stated in a report for the Rand Corporation on the benefits of the arts that the "intrinsic benefits of the arts are intangible and difficult to define. They lie beyond the traditional quantitative tools of the social sciences, and often beyond the language of common experience." Other humanities scholars do not deny that research quality or performance can be expressed quantitatively, but point out that measurable output is not important in the humanities and indicators convey information that is already widely known. An example of the former is Fisher et al. (2000, p. 9) who points out in a report for the Humanities and Social Sciences Federation of Canada (HSSFC) that "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." In term of the latter, Charle (2009, p. 165), a scholar of Modern and Contemporary History, claims that citation counts contain "*dans le meilleur des cas, qu'à des truisms, puisque les ouvrages les plus cités sont ceux des universitaires ou des chercheurs des générations établies* […]".

C. Fear of the negative steering effects of indicators

Directly related to the two previous points of criticism is fear of the negative steering effects of indicators (also referred to as dysfunctional effects or perverse effects) when research performance is measured. Humanities scholars have put forth a whole host of dysfunctional effects, but only some of them will be mentioned here. For instance, Hose (2009, p. 95), a scholar of Greek philology, argues that citation counts "have 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 […]" (own translation). Moreover, Charle (2009) claims that citation counts can easily be manipulated by self-citations or by citing friends excessively (i.e., due to old-boy networks or citation cliques). Hose and Charle analyse dysfunctional effects in rather fine-grained detail, while other humanities scholars identify detrimental effects on a general level. For example, Fisher et al. (2000, p. 8) find that

> "Academics play an important social role that is overshadowed by the reductionist focus on publication as the measure of output. They contribute analysis and commentary on issues and the human dimension of all aspects of society. These are the ultimate issues. […] An over-emphasis on research activities results in less time being devoted to more direct forms of public engagement. Unfortunately, as we witness a rise in expectations of their output as researchers and their attractiveness to funders, we witness a decrease in their involvement in service".

Furthermore, humanities scholars attribute conservative effects to 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, p. 10). Humanities scholars widely believe that measuring research performance primarily threatens diversity. In their open letter to the ERIH Steering Committee, the editors of SSH journals assessed the consequences of the evaluative function of the ERIH list: "We will sustain fewer journals, much less diversity and impoverish our discipline"

(Andersen et al., 2009, p. 8). Kemp (2008), a scholar of Art History, articulates his fear regarding the imminent loss of diversity in a metaphorical manner:

> "At the moment, we cannot imagine what will happen if [evaluation frameworks in Germany are] expanded to the European scale, when in Brussels, not olives but research is funded. [...] It is going to be a Behemoth of research! [...] This glutton will feast on megatrends; it will definitively transfer quality into measurable items although occasionally, it might like to watch a swaying skirt pass its puddle. In reality, this Behemoth likes just one thing: more of the same" (p. 148, own translation).

D. Lacking consensus on quality criteria

Finally, humanities scholars are suspicious of the assessment and comparison of research performance because they believe that comparison is impossible due to the lack of shared quality criteria and standards between disciplines and sub-discipline. Herbert and Kaube (2008) point out that "In some humanities disciplines, a consensus regarding the criteria for good and bad research is not only non-existent, but there is no consensus on the subjects of research and the meaningful use of the right methods" (p. 45, own translation). If criteria or standards do exist, they "[…] exist informally, usually refer to the same discipline and are [...] not readily transferable to other sub-disciplines" (p. 40, own translation).

II. A Framework to Explore and Develop Quality Criteria

The following sections present a framework for exploring and developing criteria and indicators for assessing research quality in the humanities. The framework addresses the four points of criticism mentioned above by (a) adopting an inside-out approach, (b) relying on a sound measurement approach, (c) making the notions of quality explicit and (d) striving for consensus. These four building blocks are elaborated below.

A. Adopting an inside-out approach

To account for the criticism that methods in research evaluation originated from the natural sciences and neglect the research patterns and practices specific to the humanities, an inside-out approach should be adopted when exploring and developing quality criteria and indicators. The inside-out approach dictates that the development process be deeply rooted in the humanities themselves, ideally in each discipline or sub-discipline, so that the humanities' unique quality criteria and conceptions may emerge. The major inter- and intradisciplinary differences in the humanities have been pointed out (e.g., Royal Netherlands Academy of Arts and Sciences, 2011; Scheidegger, 2007; Wissenschaftsrat, 2010); therefore, a focus on the individual fields or subfields is appropriate. However, a genuine inside-out approach has an *open outcome* and entails a *bottom-up procedure*. An open outcome means that whatever scholars have defined as quality criteria will be accepted no matter how they may differ from the well-known criteria in the life sciences and natural sciences. A bottom-up process requires that political stakeholders, the board of a university and a few renowned scholars are removed of their power to determine the quality criteria in a top-down manner, but rather the respective research community with all its scholars is involved – or at least is represented adequately – in the development process. Scholars' need for involvement and adequate representation in this process is obvious in the open letter to the ERIH Steering Committee, wherein editors of SSH journals argued, "This committee [of four persons] cannot be considered representative. It was not selected in consultation with any of the various disciplinary organizations that currently represent our field […]. Journal editors were only belatedly informed of the process and its relevant criteria […]" (Andersen et al., 2009, p. 7). A bottom-up approach ensures that humanities scholars' particular quality criteria and metrics will be included, thus preventing the use of inappropriate indicators derived from the natural and life sciences. In addition, a bottom-up approach that is geared toward the individual and embraces all types of scholars

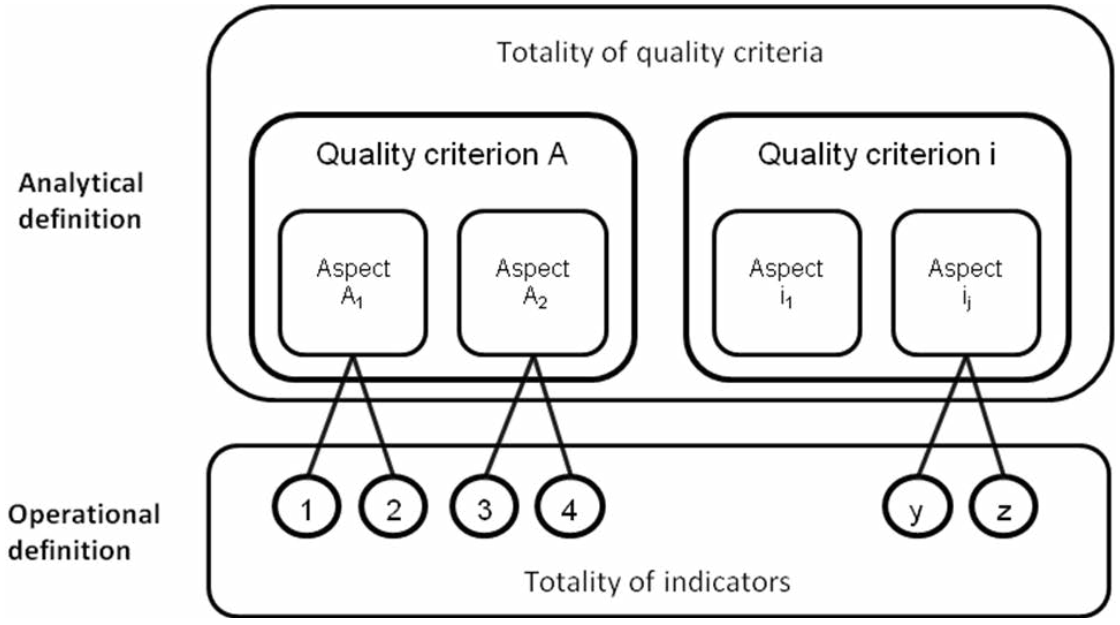
seems promising because research in the humanities is largely carried out by individual scholars (e.g., Finkenstaedt, 1990; Hellqvist, 2010; Weingart, Prinz, Kastner, Maasen, & Walter, 1991).

B. Relying on a sound measurement approach

When exploring and developing quality criteria, a sound measurement approach should be deployed to account for humanities scholars' reservations regarding quantification. To those who argue against quantification per se, this suggestion might sound paradoxical. However, the following section will briefly put forward the arguments for a sound measurement approach, describe its features and explain how it might reduce scholars' reservations against quantification, despite the apparent contradiction.

A sound measurement approach is necessary because indicators are frequently only loosely tied to definitions of quality. For example, Brooks (2005, p. 1) concludes in a review of major quality assessments in U.S. higher education 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". Donovan (2008, p. 78) goes even further by denying that definitions of quality are linked to indicators: "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". Therefore, research quality is merely defined by its measures. Furthermore, Moed (2005, p. 221) finds on the level of citation counts 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." Such weak or missing connections between quality and indicators along with the ambiguity of indicators make it difficult to understand what is being measured. In this sense, it is unsurprising that humanities scholars have strong reservations when it comes to quantification. However, it is possible to change this by relying on a measurement approach that connects the definition of quality to its measures. Such a measurement approach can be found, for instance, in the social sciences. According to social-scientific thinking, an understanding of how a construct (e.g., "research quality") is defined is a prerequisite to sensibly measuring it using indicators. Borsboom, Mellenberg and van Heerden (2004, p. 1067) also emphasise the need to define a construct before measuring it: "[The issue is not] first to measure and then to find out what it is that is being measured but rather that the process must run the other way." Schmidt (2005) formulates the social sciences measurement approach with regard to research on higher education as follows: "The attempt to create the measurability of research and teaching performances presupposes – in view of the compulsory metrics – an explicit understanding of quality" (own translation, p. 3). With this in mind, the social science measurement model can be modified for developing quality criteria and indicators as is depicted in Figure 1. Every quality criterion is specified and defined explicitly by one or more aspects (i.e., analytical definition) and each aspect is operationalised. 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 possible that there will be no suitable quantitative indicators making it impossible to measure an aspect. The measurement approach delineated above can reduce reservations related to quantification by linking quality criteria (i.e., analytical definition) to quantitative indicators (i.e., operational definition), thereby clarifying what is being measured and attributing meaning to bare figures. Moreover, this approach allows for the identification of quantifiable and non-quantifiable quality criteria by disclosing quality criteria that cannot be linked to quantitative indicators. Therefore, scholars will be able to see that not all aspects are measurable and that, consequently, some of the quality criteria will be exclusively accessible to the judgment of peers. Finally, this approach resolves scholars' reservations that research quality is reduced to one simple, quantitative expression or indicator by unfolding a wide range of metrics that can be connected to quality criteria.

Figure 1. Social science measurement model modified for developing quality criteria and indicators for the humanities.



Note. Every quality criteria is specified explicitly by one or more aspects (i.e., analytical definition) and each aspect is operationalised. 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).

C. Making the notions of quality explicit

As stated above, it is not always clear what indicators, assessment tools and procedures are measuring. Consequently, it is not always clear along which criteria research is assessed and into which direction research is being steered. This uncertainty may cause unintended effects in research assessments or trigger fears of negative steering effects in scholars. Hence, the notions of quality[2] that underlie an indicator, an assessment tool or procedure should be made as clear and explicit as possible. This reduces uncertainty regarding what is being measured and, ultimately, clarifies the direction in which research is being steered. Similarly, Moed (2005, p. 224) emphasizes in regard to citation counts and impact that

> "[…] in order to be useful and properly used in research evaluation, citation impact must be further interpreted, by assessing what it expresses regarding the aspects to be assessed in the evaluation. In other words, outcomes of citation analysis must be valued in terms of a qualitative, evaluative framework that takes into account the substantive contents of the works under evaluation."

However, even when the notions of quality are explicit, scholars might fear negative steering effects because their own notions of quality and good research do not overlap with the ones used in the research assessment. This incongruity is reflected in the criticism stated above. For example, scholars think that research outside the mainstream, contributions to society and diversity should be valued too. Consequently, to reduce fear of negative steering effects, it would be necessary to capture the scholars' notions of quality and to account for them when exploring and developing assessment criteria for the humanities. One way to capture scholars' notions of quality is to ask them about their definition of quality but the chances are very high that they will answer „I can't define what quality is, but I know it when I see it." Samples of such statements are well documented in Lamont's study (2009) on peer review processes in the SSH. It is obvious that scholars possess knowledge, which allows them to recognise

2 The term "notions of quality" is used instead of "definitions of quality" to emphasize that quality is a "slippery concept" (i.e., hard to pin down) as Donovan (2008, p. 76) says and that quality cannot be determined conclusively.

quality research; however, they cannot articulate this knowledge clearly and easily. Polanyi (1967) calls this *tacit knowing*. According to him, tacit knowing describes the "fact that we can know more than we can tell"(p. 4), whereas *explicit knowledge* is knowledge that is "capable of being clearly stated"(p. 22). Using Polanyi's terms, the task would be to transform the scholars' *tacit knowing* about quality into *explicit knowledge* in order to develop quality criteria for assessing research in the humanities. Felt (2008) advocates a similar approach and calls for "[…] techniques that allow us to translate the implicit [i.e., implicit elements of quality] into the visible, which should then form the basis for science policy decisions" (p. 279, own translation).

In sum, notions of quality in research assessments as well as scholars' notions of quality should be made as explicit as possible. Moreover, scholars' notions of quality and good research should be accounted for when exploring and developing assessment criteria for humanities research. This corresponds to the inside-out approach outlined above. By means of making these notions explicit, scholars' fears of negative steering effects, as well as actual negative steering effects, can be reduced.

D. Striving for consensus

To account for the lack of consensus on quality criteria, an approach should be adopted to allow for consensus within a discipline or sub-discipline. Such an approach should reveal to the research community those criteria that are consensual and those that are not. For this purpose, all scholars in a particular research community or discipline should be included. This corresponds to the bottom-up approach described above.

III. Implementation of the Framework

This section outlines an empirical procedure that explores and develops quality criteria for assessing research in the humanities. This procedure builds on the framework presented above and comprises interviews based on the Repertory Grid technique to make the scholars' notions of quality explicit, as well as a survey that uses the Delphi method to validate the quality criteria and reach a consensus.

A. Repertory Grid interviews

The Repertory Grid technique was developed by George A. Kelly (1955) within the framework of the Personal Construct Psychology so as to explore and map subjective concepts (so-called constructs) that individuals use to interpret, structure and evaluate the entities that constitute their lives (see Fransella, Bell, & Bannister, 2004; Fromm, 2004; Walker & Winter, 2007). Therefore, the Repertory Grid technique can be used to capture the subjective notions of quality that scholars use to interpret, structure and evaluate the entities and events in their research lives. Rosenberger und Freitag (2009) emphasise the flexibility of the technique because it allows an idiographic as well as a nomothetic approach and facilitates qualitative and quantitative analysis. This versatility is what enables the scholars to describe their notions of research quality in their own words (i.e., idiographic dimension) and permits the summarisation of the individual perceptions for each discipline or sub-discipline, which allowes for the development of discipline-specific propositions (i.e., nomothetic dimension). The versatility of this method also permits a content analysis of the interviews (qualitative dimension) and the use of statistical methods that structure the compiled concepts of research quality (quantitative dimensions). Due to its flexibility, Repertory Grids are deployed extensively in exploratory studies and applied problems in a variety of fields (for an overview see Fransella, et al., 2004). Repertory Grid interviews are intransparent to participants, thereby reducing response bias (Hicks, 1999). Since conducting and analysing Repertory Grid interviews is highly time-consuming, studies with single cases or small sample-sizes are typical. However, a major advantage of the Repertory Grid technique is that it captures tacit knowledge (Buessing, Herbig, & Ewert, 2002; Jankowiecz, 2001; Ryan & O'Connor, 2009), i.e., knowledge that is difficult or impossible to verbalise (Polanyi, 1967). In this respect, the Repertory Grid method is superior to open interviews or group discussions (McGeorge & Rugg, 1992;

Winter, 1992) which are usually used to identify quality criteria. Therefore, the Repertory Grid method addresses two building blocks of the framework presented above. First, it addresses the inside-out approach by facilitating analysis at the individual level (ideographic perspective) and at the disciplinary or a sub-disciplinary level (nomothetic perspective). Second, it makes the scholars' notions of quality explicit by translating tacit knowledge into explicit knowledge and by capturing the notions of quality in the scholars' words and intentions. For these reasons, the Repertory Grid is well suited as an initial exploratory step in the quest for quality criteria in the humanities.

B. Delphi survey

It is possible to derive quality criteria from the scholars' notions of quality obtained using the Repertory Grid technique. These criteria are derived from a small sample size due to the time-consuming nature of Repertory Grids; therefore, it is necessary to validate the quality criteria by a large group of scholars and reach a consensus. This can be achieved using the Delphi method, which is in line with the bottom-up approach outlined above. Linstone and Turoff (1975, p. 3) define the Delphi method as

> "A method for structuring a group communication process so that the process is effective in allowing a group of individuals, as a whole, to deal with a complex problem. To accomplish this 'structured communication' there is provided: some feedback of individual contributions of information and knowledge; some assessment of the group judgment or view; some opportunity for individuals to revise views; and some degree of anonymity for the individual responses".

Delbecq, Van de Ven and Gustafson (1975, p. 10) provide a definition that is more process–oriented by explaining that Delphi is "a method for the systematic solicitation and collection of judgments on a particular topic through a set of carefully designed sequential questionnaires interspersed with summarised information and feedback of opinions derived from earlier responses." In general, a Delphi survey is a method that makes use of experts' opinions in multiple rounds with anonymous feedback after each round in order to solve a problem (Häder & Häder, 2000). Following Häder (2002), a classical Delphi study starts with the identification and delineation of the problem at hand. This can be done either by a research team that organises and monitors the study or by surveying experts in an initial qualitative Delphi round through an open-ended or structured questionnaire. In the subsequent rounds, a standardised questionnaire is used to evaluate the subject under discussion, typically until a consensus is reached. Linstone and Turoff (1975, p. 4) indicate that the Delphi method tackles, inter alia, the following methodological issues:

> "The problem does not lend itself to precise analytical techniques but can benefit from subjective judgments on a collective basis. […] More individuals are needed than can effectively interact in a face-to-face exchange. Time and cost make frequent group meetings infeasible. […] Disagreements among individuals are so severe or politically unpalatable that the communication process must be refereed and/or anonymity assured. The heterogeneity of the participants must be preserved to assure validity of the results, i.e., avoidance of domination by quantity or by strength of personality ('bandwagon effect')".

Despite these advantages, it must be emphasised that a Delphi survey is a time-consuming and laborious method of collecting data due to its iterative and sequential nature (Hsu & Sandford, 2007). However, the Delphi method has been widely applied in many research fields including research on higher education (see Häder & Häder, 2000; Murry & Hammons, 1995; Palomares-Montero & Garcia-Aracil, 2011). At least two studies have employed the Delphi method in order to develop quality criteria for assessing research. Lahtinen, Koskinen-Ollonqvist, Rouvinen-Wilenius, Tuominen and Mittelmark (2005) used it to develop a set of quality criteria for health promotion research in Finland. In Canada, the method was applied in a nationwide study to identify criteria and indicators of quality and excellence for colleges and universities involving more than 20 stakeholder groups (e.g., alumni, deans, university presidents, CEO's of companies and all members of the Canadian Parliament) with a sample size of more than 15,000 people (Nadeau, 1995).

Building on the classical Delphi design and the Repertory grid interviews, a Delphi survey to develop quality criteria and indicators for the humanities could be conceived as follows: In an initial qualitative round, a large group of scholars modifies and supplements the quality criteria obtained from the Repertory Grid interviews. The scholars also name the indicators that can be used to measure the criteria. For this purpose, a structured questionnaire can be employed in which each quality criteria is specified explicitly by one or more aspects and each aspect is tied to at least one indicator that specifies how the aspect can be measured. In subsequent rounds, scholars rate the quality criteria (specifically the aspects) and the indicators obtained in the first round through a standardised questionnaire until a consensus is reached. Ideally, a whole discipline, a sub-discipline or a research community should be involved in this process or, at least represented adequately in the sample. Consequently, such a Delphi survey can address three of the framework's building blocks presented above. First, it contributes to the inside-out approach by involving a large group of scholars (e.g., a research community, a discipline or a sub-discipline) and facilitating, as well as structuring, this group's communication process. Second, it fulfils the social sciences' measurement approach by connecting the scholars' quality criteria to indicators. Finally, the Delphi method facilitates the process of reaching a consensus.

IV. Conclusion

In the process of developing new tools for measuring and assessing research quality in the humanities, many challenges must be addressed, such as technical problems (e.g., building publication databases, capturing social impact) and scholars' opposition to measuring research performance. This paper focuses on scholars' opposition. Humanities scholars criticize that the methods employed in research evaluation originated from the natural sciences and are modelled after them. They have strong reservations regarding the quantification of research quality and fear the negative steering effects of indicators. Moreover, scholars state that there is a lack of consensus on quality criteria within the humanities' disciplines or sub-disciplines and, therefore, a comparison or assessment of research quality is impossible.

To address these criticisms and objections of humanities scholars, we suggest exploring and developing quality criteria and indicators within the following framework.

Adopting an inside-out approach. Inside-out requires that the development of criteria and indicators be rooted in the humanities themselves, ideally in each discipline or sub-discipline, so that the unique quality criteria and conceptions of each discipline can emerge. A genuine inside-out approach has an open outcome and entails a bottom-up procedure. Open outcome means that whatever the scholars define as a quality criteria will be accepted, no matter how different they may be from existing criteria. Bottom-up means that a research community or a discipline should be involved entirely – or at least be represented adequately – in the development process.

Relying on a sound measurement approach. A sound measurement approach links quality criteria (i.e., analytical definition) to quantitative indicators (i.e., operational definition); thereby clarifying what is being measured. Moreover, such an approach allows for the identification of quantifiable and non-quantifiable quality criteria by disclosing those quality criteria that cannot be connected to quantitative indicators. This reveals that not all aspects can be measured and that some of the quality criteria are exclusively reliant on the judgment of peers. A sound measurement approach can also resolve scholars' beliefs that research quality is reduced to one simple, quantitative expression or indicator by unfolding a wide range of metrics that are connected to the quality criteria.

Making the notions of quality explicit. This suggestion consists of two parts. First, the notions of quality that underlie an indicator, an assessment tool or an evaluation procedure should be made as clear and explicit as possible. This reduces uncertainty about what is being measured and clarifies the direction in which research is being steered. Second, the scholars' explicit notions of quality and good research should be taken into account when developing criteria and indicators so that research is ultimately

steered in the direction of the scholars' understanding of good research, thereby reducing fear of negative steering effects and actual negative steering effects. Since the scholars' notions of quality often exist as tacit knowledge (i.e., knowledge that cannot be articulated easily and clearly; Polanyi, 1967), methods that translate tacit notions of quality into explicit knowledge (i.e., knowledge that can clearly be stated) are needed.

Striving for consensus. To account for the lack of consensus regarding quality criteria, an approach should be adopted that allows consensus to be reached in a discipline or sub-discipline. Such an approach should reveal which criteria are consensual and which are not to the research community.

The Repertory Grid technique and the Delphi method are well-suited approaches to implement this framework in an empirical manner. The Repertory Grid can be employed to explore quality criteria by facilitating the translation of scholars' tacit knowledge about research quality into explicit knowledge. A Delphi survey aimed at a discipline, sub-discipline or research community can be used to validate and reach a consensus on the quality criteria and indicators.

The delineated framework and its empirical implementation contribute to the development of criteria and indicators for assessing research quality in the humanities. In particular, the framework accounts for humanities scholars' objections to measuring and assessing research quality. Furthermore, the outlined framework may help to tackle the "lack of information on how to develop indicators" and the "problem related to the definition of indicators" (Palomares-Montero & Garcia-Aracil, 2011, p. 354).

V. References

Academics Australia. (2008). Letter to senator the honourable Kim Carr, minister for innovation, industry, science and research Retrieved 15. September 2009, from http://www.academics-australia.org/AA/ERA/era.pdf

Andersen, H., R. Ariew, M. Feingold, A. K. Bag, J. Barrow-Green, B. van Dalen 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. V., Cote, G., Larivière, V., & Gingras, Y. (2006). Benchmarking scientific output in the social sciences and humanities: The limits of existing databases. *Scientometrics, 68*(3), 329-342.

Australian Research Council. (2012). The Excellence in Research for Australia (ERA) Initiative Retrieved 30.1., 2012, from http://www.arc.gov.au/era/

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.

Brooks, R. L. (2005). Measuring university quality. *The Review of Higher Education, 29*(1), 1-21.

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

Butler, L., & Visser, M. S. (2006). Extending citation analysis to non-source items. *Scientometrics, 66*(2), 327-343.

Charle, C. (2009). How academics and scholars should be assessed: Critics and proposals. *Vingtieme Siècle-Revue D' Histoire, 102*, 159-170.

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.

Donovan, C. (2008). Das zweiköpfige Lama zähmen: Die australische Suche nach den besten Evaluierungsmethoden für die Geisteswissenschaften. In E. Lack & C. Markschies (Eds.), *What*

the hell is quality? Qualitätsstandards in den Geisteswissenschaften (pp. 74-98). Frankfurt / New York: Campus.

EERQI Consortium. (2011). EERQI - European Educational Research Quality Indicators Retrieved 30.1., 2012, from http://www.eerqi.eu/

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 Retrieved 05.1., 2012, from http://esftraining.gky06.punkt.de/research-areas/humanities/research-infrastructures-including-erih/joint-response-to-criticism.html

European Science Foundation. (2011a). ERIH foreword Retrieved 30.1., 2012, from http://www.esf.org/research-areas/humanities/erih-european-reference-index-for-the-humanities/erih-foreword.html

European Science Foundation. (2011b). European Reference Index for the Humanities (ERIH) Retrieved 30.1., 2012, from http://www.esf.org/research-areas/humanities/erih-european-reference-index-for-the-humanities.html

Felt, U. (2008). Angemessen messen? Die Qualität von Forschungsprojekten in den Geisteswissenschaften. In E. Lack & C. Markschies (Eds.), *What the hell is quality? Qualitätsstandards in den Geisteswissenschaften* (pp. 273-291). Frankfurt / New York: Campus.

Finkenstaedt, T. (1990). Measuring research performance in the humanities. *Scientometrics, 19*(5-6), 409-417.

Fisher, D., Rubenson, K., Rockwell, K., Grosjean, G., & Atkinson-Grosjean, J. (2000). Performance indicators and the humanities and social sciences. Retrieved from http://www.fedcan.ca/content/en/364/performance-indicators-and-the-humanities-and-social-sciences.html

Forslöw, B., Rehn, C., & Wadskog, D. (2005). Användning av bibliometri som delparameter för tilldelning av forskningsmedel till institutioner vid Karolinska Institutet och forskning vid SLL. Beskrivning av metodval (Vol. 09-09-05). Stockholm: Karolinska Institutet/SLL.

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

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

Giménez-Toledo, E., & Roman-Roman, 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.

Giménez-Toledo, E., Roman-Roman, A., & Alcain-Partearroyo, M. D. (2007). From experimentation to coordination in the evaluation of Spanish scientific journals in the humanities and social sciences. *Research Evaluation, 16*(2), 137-148.

Glänzel, W., & Schoepflin, U. (1999). A bibliometric study of reference literature in the sciences and social sciences. *Information Processing & Management, 35*(1), 31-44.

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

Guillory, J. (2005). Valuing the humanities, evaluating scholarship. *Profession, 11*, 28-38.

Häder, M. (2002). *Delphi-Befragungen. Ein Arbeitsbuch*. Wiesbaden: Westdeutscher Verlag.

Häder, M., & Häder, S. (2000). Die Delphi-Methode als Gegenstand methodischer Forschung. In M. Häder & S. Häder (Eds.), *Die Delphi-Technik in den Sozialwissenschaften. Methodische Forschungen und innovative Anwendungen* (pp. 11-31). Wiesbaden: Westdeutscher Verlag.

Hellqvist, B. (2010). Referencing in the humanities and its implications for citation analysis. *Journal of the American Society for Information Science and Technology, 61*(2), 310-318. doi: 10.1002/asi.21256

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 / New York: Campus.

Hicks, C. M. (1999). *Research methods for clinical therapists: Applied project design and analysis*. Edinburgh, UK: Churchill Livingstone.

Hicks, D. (2004a). 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.

Hicks, D. (2004b). 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.

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

Hose, M. (2009). Qualitätsmessung: Glanz und Elend der Zahl. In C. Prinz & R. Hohls (Eds.), *Historisches Forum. Qualitätsmessung, Evaluation, Forschungsrating. Risiken und Chancen für die Geschichtswissenschaften?* (Vol. 12, pp. 91-98). Berlin: Clio-online und Humboldt-Universität zu Berlin.

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

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.

Kemp, W. (2008). Wehe, Behemoth erwacht - harmlose und weniger harmlose Moden in den Geisteswissenschaften. In E. Lack & C. Markschies (Eds.), *What the hell is quality? Qualitätsstandards in den Geisteswissenschaften* (pp. 145-149). Frankfurt / New York: Campus.

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 / New York: Campus.

Lahtinen, E., Koskinen-Ollonqvist, P., Rouvinen-Wilenius, P., Tuominen, P., & Mittelmark, M. B. (2005). The development of quality criteria for research: A Finnish approach. *Health Promotion International, 20*(3), 306-315.

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

Lane, J. (2010). Let's make science metrics more scientific. *Nature, 464*(25 March), 488-489.

Linstone, H. A. and M. Turoff (1975) 'Introduction'. In Linstone, H.A. and Turoff, M. (eds.) *The Delphi method. Techniques and applications*, pp. 3-12. Addison-Wesley: Don Mills.

McCarthy, K., Ondaatje, E., Zakaras, L., & Brooks, A. (2004). *Gifts of the Muse: Reframing the debate about the benefits of the arts*. Pittsburgh, PA: RAND Corporation.

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

Moed, H. F. (2005). *Citation analysis in research evaluation* (Vol. 9). Dordrecht: Springer.

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

Murry, J. W., & Hammons, J. O. (1995). Delphi - a versatile methodology for conducting qualitative research. *Review of Higher Education, 18*(4), 423-436.

Nadeau, G. G. (1995). *Criteria and indicators of quality and excellence in colleges and universities in Canada: Summary of the three phases of the project / Critères et indicateurs de qualité et d'excellence dans les collèges et les universités du Canada: Bilan de trois phases du projet*. Winnipeg: Centre for Higher Education Research and Development, University of Manitoba, Canada.

National Science Foundation. (2009). MESUR project Retrieved 30.1., 2012, from http://mesur.informatics.indiana.edu/

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., Zwaan, R., De Bruin, R., & 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.

Netherlands Organisation for Scientific Research. (2009). Evaluating Research in Context (ERiC) Retrieved 29.6., 2011, from http://www.eric-project.nl

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

Plumpe, W. (2009). Stellungnahme zum Rating des Wissenschaftsrates. In C. Prinz & R. Hohls (Eds.), *Historisches Forum. Qualitätsmessung, Evaluation, Forschungsrating. Risiken und Chancen für die Geschichtswissenschaften?* (Vol. 12, pp. 121-126). Berlin: Clio-online und Humboldt-Universität zu Berlin.

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

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. [Article]. *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. *Artikel aus dem evaNet-Projekt 06/2005 vom 24. Mai 2005* Retrieved 11. September 2009, from http://www.hrk.de/de/download/dateien/05-2005_-_Theoriedefizit_in_der_deutschen_Hochschulevaluation_-_Schmidt.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 the scientific publication output at research institutions. *ISSI Newsletter, 6*(1), 22-28.

Thomson Reuters. (2011). The Book Citation Index Retrieved 30.1., 2012, from http://wokinfo.com/products_tools/multidisciplinary/bookcitationindex/

Vec, M. (2009). Qualitätsmessung: Die vergessene Freiheit. Steuerung und Kontrolle der Geisteswissenschaften unter der Prämisse der Prävention. In C. Prinz & R. Hohls (Eds.), *Historisches Forum. Qualitätsmessung, Evaluation, Forschungsrating. Risiken und Chancen für die Geschichtswissenschaften?* (Vol. 12, pp. 79-90). Berlin: Clio-online und Humboldt-Universität zu Berlin.

Verband der Historikerinnen und Historiker Deutschlands. (2010). Ausschuss des Historikerverbandes beriet über Forschungsratings in den Geisteswissenschaften und über den Historikertag 2010 Retrieved 30.1., 2012, from http://www.historikerverband.de/fileadmin/_vhd/bilder/Pressemitteilung_WR_Rating.pdf

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

Weingart, P., Prinz, W., Kastner, M., Maasen, S., & Walter, W. (1991). *Die sogenannten Geisteswissenschaften: Aussenansichten: Die Entwicklung der Geisteswissenschaften in der BRD, 1954-1987*. Frankfurt am Main: Suhrkamp.

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.

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.

Wissenschaftsrat. (2011a). Forschungsrating Anglistik/Amerikanistik Retrieved 30.1., 2012, from http://www.wissenschaftsrat.de/arbeitsbereiche-arbeitsprogramm/forschungsrating/anglistikamerikanistik/

Wissenschaftsrat. (2011b). Zum Forschungsrating allgemein Retrieved 30.1., 2012, from http://www.wissenschaftsrat.de/arbeitsbereiche-arbeitsprogramm/forschungsrating/