Ingrid Gogolin · Fredrik Åström
Antje Hansen *Editors*

RESEARCH

Assessing Quality in European Educational Research

Indicators and Approaches

Springer VS

Assessing Quality in European Educational Research

Ingrid Gogolin • Fredrik Åström
Antje Hansen (Eds.)

Assessing Quality in European Educational Research

Indicators and Approaches

Springer VS

Editors
Ingrid Gogolin
University of Hamburg
Hamburg, Germany

Fredrik Åström
Lund University Libraries
Lund, Sweden

Antje Hansen
Koordinierungsstelle für Mehrsprachigkeit und sprachliche Bildung (KoMBi)
University of Hamburg,
Hamburg, Germany

ISBN 978-3-658-05968-2
DOI 10.1007/978-3-658-05969-9

ISBN 978-3-658-05969-9 (eBook)

The Deutsche Nationalbibliothek lists this publication in the Deutsche Nationalbibliografie; detailed bibliographic data are available in the Internet at http://dnb.d-nb.de.

Library of Congress Control Number: 2014939070

Springer VS

Printed on acid-free paper

Springer VS is a brand of Springer DE.
Springer DE is part of Springer Science+Business Media.
www.springer-vs.de

Table of Content

V. Intrinsic Indicators & Peer Review

VI. Relevance for other Research Fields

Contributing Authors

Fredrik Åström

Fredrik Åström received his PhD in Library and Information Science (LIS) at Umeå University, Sweden, in 2006; and holds a position as Postdoctoral Research Fellow at Lund University Libraries, also in Sweden. His research on scholarly communication has been focused on mapping and visualizing research areas utilizing bibliometric methods; and in particular through citation analysis. A particular interest in Dr. Åströms research has been directed towards the organization of multidisciplinary research areas with connections to the humanities, the social sciences and the computer sciences, as well as with strong connections to fields of professional practices. Dr. Åströms research has been published in leading LIS journals, as well as in journals and handbooks in e.g. research policy and entrepreneurship research; and presented internationally at e.g. the biannual conferences of the International Society for Scientometrics and Informetrics (ISSI).

Alexander Botte

Alexander Botte, is a former teacher and was trained as a scientific documentalist. He is now deputy chief of the Department "Information Centre Education" at the German Institute for International Educational Research, Frankfurt Main. Since 1979, he has been working with bibliographic databases on educational topics and since 1992 he has been project manager of the German Education Index, a cooperative initiative of more than 30 documentation units of research and higher education institutions in the German-speaking countries. Since 1999, Alex has also been project manager of the German Education Server, an online portal for educational research and practice. He has been personally involved in bibliometrics projects since 2003.

Ingrid Gogolin

Dr. phil. Dr. phil. h.c. Ingrid Gogolin is Professor for international comparative and intercultural education research at the University of Hamburg. Her research is focused on problems of migration and linguistic diversity in education. She was coordinator of the EERQI-project. Recent research projects deal with the

following topics: Linguistic diversity management in urban areas (Research Cluster of Excellence at the University of Hamburg; www.lima.uni-hamburg.de); Support of migrant children in schools (www.foermig.uni-hamburg.de); Multilingualism and Education (www.kombi.uni-hamburg.de). Examples of publications: Book series 'Hamburg Studies on Linguistic Diversity' (2013f, ed. with Peter Siemund, Amsterdam: John Benjamins. – The Bilingualism Controversy (ed. with Ursula Neumann). Wiesbaden: Springer VS 2009. – Migration, gesellschaftliche Differenzierung und Bildung (ed. with Bernhard Nauck). Opladen: Leske + Budrich 2000. – Der monolinguale Habitus der multilingualen Schule. Münster, New York: Waxmann 1994 (2nd edition 2009).

Stefan Gradmann

Dr. Stefan Gradmann is a Professor teaching knowledge management and semantic knowledge architectures at the Katholieke Universiteit Leuven, University Library (since 2013). Before he taught at the School of Library and Information Science of Humboldt-Universität zu Berlin (since 2008). Other focal areas in teaching and research are digital libraries, library automation as well as the use of information technology in the realm of signification and interpretation in the 'Digital Humanities' with a specific focus on the 'document'-notion and its deconstruction. He has been substantially involved in the creation of Europeana from its beginnings, where he is responsible for semantic interoperability and one of the architects of the Europeana Data Model (EDM).

Antje Hansen

Antje Hansen studied Economics at the University of Hamburg and European Studies at the Bosphorous University in Istanbul. For 1,5 year she worked at the German research centre DESY (Deutsches Elektronen Synchrotron) in the EU project office. After that she worked at the University of Hamburg as EU-counsellor in the presidential department, assisting scientists with their application for the EU's 7th Framework Programme. She then changed to the faculty of education, psychology and human movement where she focused on research funding in the area of social sciences and humanities. Antje also took over the management of the EERQI project and is working on a follow-up project. Antje currently works the coordination office for multilingualism and language education, supporting 14 German research projects in the aera of multilingualism and

language education, funded by the German ministry for education and research and located at the University of Hamburg.

Frank Havemann

Dr. rer.nat. Frank Havemann, a trained physicist, is working as a researcher and lecturer (bibliometrics) at the School of Library and Information Science of Humboldt-Universität zu Berlin. Since 1990 he has done research on bibliometric topics like science indicators, collaboration in science and technology, and growth dynamics of science. Together with others he is now engaged in the ACUMEN EU-project and in a project (financed by the German government) for measuring diversity of research. Since 2007 he is board member of the Collnet Journal of Scientometrics and Information Management. He has written a text book that introduces students to bibliometrics and that is used by him in a master course.

Eberhard R. Hilf

is CEO of the ISN, Institute for Science Networking Oldenburg GmbH at the Carl von Ossietzky University, founded in 2001. after his retirement from professorshipss as a Theoretical Physicist at Oldenburg, and before at the Technical University Darmstadt and the University Düsseldorf. As an early user of upcoming new technical means for the information management in Science he headed a group implementing an early web-server in 1993 and advised the installation of web-servers in most German Physics Departments. Starting in 1993 he pleaded for Open Access of scientific publications and was involved in the many attempts to persue this goal, such as advising scientific Society- as well as commercial publishers, and designing joint projects with them (e.g. the EU-application DDD Distributed Documents Database in Physics; and the German-wide project Global-Info of the main Learned Societies and publishers).
ISN is also serving the global Physics Department Network which was served by the European Physical Society.

Axel Horstmann

is professor at the University of Hamburg. He received a PhD in classical philology and completed his habilitation in philosophy. Until 2010 he was member of

the executive board of the Volkswagen Foundation, Hanover, where he headed the Humanities and Social Sciences Division. Since his retirement, he has been a freelance consultant on issues related to science promotion and science management. He is chairman of the academic advisory board on the mapping project on "Endangered Subjects" (Kleine Fächer) in cooperation with the German Rectors' Conference (HRK), and chairman of the board of trustees of the Hanns-Lilje-Stiftung, Hanover. In addition, he is engaged in advisory committees of other German foundations and institutions. He has published books and articles on classical philology, history and theory of the humanities, especially on hermeneutics, history of ideas, the reception of classical antiquity in modern times as well as on higher education and science policy.

Sarah McMonagle

Sarah McMonagle is currently a DAAD visiting Postdoctoral Fellow at the University of Hamburg in Germany. She received her PhD in Language Policy and Planning from the School of Languages, Literatures and Cultures at the University of Ulster in Northern Ireland in 2010. Her research on societal multilingualism, linguistic identities and language attitudes is interdisciplinary in nature and seeks to inform public policy through academic investigation. She has worked as a research assistant for the Department of Education in Northern Ireland and has spent some time with the Council of Europe's Secretariat of the European Charter for Regional or Minority Languages in Strasbourg. Sarah also holds a MA in Contemporary European Studies, completed at the University of Bath, Charles University Prague and Humboldt University in Berlin. She has a BA in European Studies from Trinity College Dublin where she majored in German, having spent two semesters at University of Tübingen.

Ton Mooij

Ton Mooij is functioning as a manager and researcher at ITS (Institute for Applied Social Sciences) of Radboud University, Nijmegen, the Netherlands. He also holds a position as professor by special appointment for educational technology at CELSTEC (Centre for Learning Sciences and Technologies) of Open University of the Netherlands, Heerlen. Professor Mooij received his PhD in social sciences at Radboud University in 1987. His research combines ICT-supported improvement of school processes including cognitive and social learning effects with pupils and teachers in primary and secondary education. Main

topics of interest concern development and application of ICT in relation to high ability of pupils and the support of safety for both pupils and teachers in and around schools. He was awarded several prizes for his publications focusing on theory and experimental research on cognitive excellence in school practice. He also conducted six national surveys into pupils' and teachers' bullying and school violence in primary and secondary education. Moreover, he investigated school-based intervention possibilities to effectively prevent cognitive and social problem behaviour of pupils. Ton Mooij has published in many international and national journals and books on both issues of high ability and school safety. In 1998 he participated in the foundation of the international research network 'ICT in education and training' of the 'European Educational Research Association' (EERA). Since then he is the chairman of this network.

Jenny Oltersdorf

Jenny Oltersdorf studied Library and Information Science and Theology at Humboldt-Universität zu Berlin. She is working as research assistant at Institute for Research Information and Quality Assurance (iFQ). Her research interests revolve around bibliometrics and scholarly communication. In her PhD project she focuses on the measurement of research output in the Humanities.

Sybille Peters

Sybille Peters studied applied computer science at the Hochschule Hannover. She is a senior software developer with experience in search engines. She joined the SearchEngineLab of the RRZN, Leibniz Universität Hannover in 2008 and worked full-time for the project EERQI.

Wolfgang Sander-Beuermann

Dr. Wolfgang Sander-Beuermann is the head of the Search Engine Laboratory at the Regional Computer Center of Lower Saxony (RRZN) of the Leibniz University of Hannover. The most well know product of the Search Engine Laboratory is the popular German meta search engine www.metager.de.
Furthermore he is director of SUMA-EV - Association for Free Access to Knowledge" (the acronym SUMA derives from the abbreviation of the German word for "search engine", SUchMAschine, see www.suma-ev.de). He received

his doctorate in 1980 from the Institut of Thermodynamics, Leibniz University of Hannover. Thereafter he was awarded a DAAD-Nato post-doc fellowship at the University of California, Santa Barbara. Back in Germany he worked as the deputy head of a research center of the Fraunhofer-Gesellschaft before he returned to Hannover University in 1984. He has completed and is involved in numerous projects, presentations, articles and speeches, including radio and TV, about search engines and free access to digital knowledge.

Ágnes Sándor

Dr. Ágnes Sándor is a project leader at Xerox Research Centre Europe (XRCE). She has been working at XRCE since 1996 as a computational linguistics researcher. She has a PhD in linguistics from the University Lyon 2. Her activities have included the construction of morphological analyzers and part-of-speech taggers for English, French and Hungarian. For several years, she was involved in projects concerning information extraction in the biomedical domain. In the past years, she has been focusing on discourse analysis and content extraction based on meta-discourse. She has published papers on all these fields.

Thomas Severiens

Thomas Severiens is a Physicist, who works in the field of Information-Engineering since 1995. He is a researcher at the University of Osnabrück and at the Institute for Science Networking in Oldenburg. He works on the fields of Vocabulary Management and tools for Open Access publishing. He is a member of the Advisory Boards of DINI e.V. and of the Dublin-Core Metadata Initiative. His interests are the research of the publication chain, trying to enlighten alternatives which are faster, cheaper and more naturally interwoven into the research process then the traditional workflows.

Verena Stumm

Verena Stumm is a lecturer and research assistant at the university of Hamburg for psychological methods. She also works for the project 'sumdata.org' with focus on the development of alternative ways of electronic surveys and quality criteria in evaluating processes.

In the EERQI project she was involved in the development of the questionnaire for the peer review exercise and its testing and revision for the EERA.
The results were made public in project meetings and international conferences. Currently her research interests also cover the variation of quality criteria in longitudinal processes.

Angela Vorndran

Angela Vorndran is a research assistant at DIPF (German Institute for International Educational Research). She holds a master´s degree in Scandinavian studies as well as in library and information science. She was involved in different research strands over the whole duration of the EERQI project such as bibliometric analysis, linguistic text analysis, multilingual search and research monitoring of educational science. The results were made public in project meetings and international conferences. Currently her research interests also cover information behaviour of professionals.

Approaches on Assessing Quality in European Educational Research
Introduction to the volume

Ingrid Gogolin, Fredrik Åström, Antje Hansen

Across the world, structures and control mechanisms of publicly funded research have changed dramatically in the last decades. Input governance of research funding has increasingly been replaced by output and control related mechanisms inspired by economic models – a transformation of all publically funded activities since the 1980s and onwards, referred to as 'new public management' (NPM) – rather than traditions in academia. These trends have been, and continue to be, accompanied by a decrease in public funding of research, especially in the social sciences and the humanities (Brinkley, 2009; Halevi & Bar-Ilan, 2013). These developments arise in parts from issues related to the scientific work process per se. However, they are also driven by external factors, such as economical or technical challenges and their impact on academic life. The introduction of competition based models for research policy and management at all steps in the research work process – from attracting research funding to publishing the results – belongs to this overarching reform of the science system. The change from "classical government" to "governance" of the scientific sphere (Hornbostel, 2011), 9) introduced an entrepreneurial perspective on the management of scientific and scholarly activities for the purpose of increasing both the quality and effectiveness of academic research. One important aspect of introducing this business oriented form of managing academic research is the large scale introduction of quality control methods utilizing various forms of performance indicators for measuring research activities.

It is within this context that the EERQI (European Educational Research Quality Indicators1) project was initiated, the outcome of which is presented in the contributions to this volume. The project is based on the observation that the concept of quality is explicitly used or resonates implicitly in the discourses that legitimize new governance mechanisms and modes of research funding. The emphasis on the quality of research, and the measurement thereof, is perceived as the driving force for the tendency to re-evaluate and redevelop the structures in

[1] The project was funded under the SSH theme of the 7th framework program of the European Commission. Its funding period lasted from 01.04.2008 – 31.03.2011.

I. Gogolin et al. (eds.), *Assessing Quality in European Educational Research,*
DOI 10.1007/978-3-658-05969-9_1, © Springer Fachmedien Wiesbaden 2014

research areas, for redesigning the system for funding research institutions and projects, and for implementing systems for control with the purpose of facilitating the work of research funding decision and policy makers. However, very little attention has been paid to an explicit discussion on the quality of the mechanisms established to assess the quality of research. Thus, the purpose of the EERQI project is to address the question of how quality can be identified or measured.

In particular, the project has focused on the following questions. What are the characteristics of the current quality control systems applied in the contexts of research governance and funding? What are the possible effects of these systems on research conducted in the European Research Area; and in the Social Sciences and the Humanities (SSH)? To address these questions, the educational sciences were selected as an exemplary discipline of investigation, considering how educational research shares characteristics with a great deal of other fields within the Social Sciences and the Humanities.

One aspect of this is the wide spectrum of theoretical and methodological approaches found within the educational sciences: from philosophical-historical methodologies to psychologically or sociologically based empirical observations; from hermeneutical interpretation, over single case studies, to statistical analyses of large scale survey data sets. This range of theoretical and methodological perspectives reflects most modes of knowledge production found in the Social Sciences and the Humanities.

Another aspect which the educational sciences are sharing with other areas of SSH is the relevance of language. In medical research and the natural sciences, knowledge production and dissemination may function irrespective on the language which is used, which has led to English becoming the lingua franca of the sciences.2 SSH-research, however, is to a large extent deeply rooted in the cultural and intellectual traditions of the regional or national languages in which it is carried out. The usefulness and necessity of discourse and knowledge dissemination in a global working language – which is English today, that was German less than a century ago, and that may be Chinese within a century – cannot be denied. Irrespective of this intermediate usage of a lingua franca, the discretionary utilization of the language in which the knowledge producers live, carry out their work and feed into discourse is necessary for the advancement of insight in the majority of SSH-related research problems.

[2] This position is also increasingly contested, even from within the natural sciences (e.g. Mocikat, 2010).

Since the majority of SSH fields are deeply rooted in local cultural and language traditions, much European SSH-research is at a disadvantage in current systems of 'quality detection', where a basic assumption is that research is published internationally in journal articles. Distinctive and fruitful traditions of work are locked into national intellectual resources and enabling them to move across borders is a slow process. This problem was approached within the EERQI project. An important aim was to contribute to the development of, and agreement on, common standards paving the way to a virtual working space for European researchers – regardless of which European language they produce knowledge in.

An ex ante review of the appropriateness of instruments and strategies for quality assessment that are actually applied to SSH research, and the educational sciences in particular, resulted in a generic conclusion: existing instruments for quality assessment do not lead to a valid identification of 'quality' since they do not measure what they claim to measure. One example being quality assessment based on citation indices and journal rankings, one of the more common approaches in contemporary research quality assessment.

A central quality criterion used in many instruments for measuring research quality is the 'international visibility' of research findings, as expressed in research published in journals with good reputation and of high impact, as determined by the number of citations to the journal. Typically, this approach builds on data from the Web of Science (WoS) databases Science Citation Index and Social Science Citation Index, and the citation analyses of journal citation data in Journal Citation Reports (JCR), a set of interlinked commercials products provided and owned by the US-American publishing group Thomson Reuter. The JCR journal rankings often play an important role in systems for reporting research quality and effectiveness. Analyzing the journals representing the educational sciences in JCR3 (Social Science Edition 2009), the following information can be found.

In total, there are 201 educational research journals indexed in the JCR. Approximately 52% of the journals are published by US-American publishers, whereas 24% comes from British publishing houses. Aside from the Anglo-American publishers, the Netherlands contributes with 4% and Germany with 3% of the JCR educational research journals). Altogether, publishing houses from 15 nations across the world are represented with educational research jour-

[3] Journal Citation Reports® (JCR)is a commercial product offered by the US-American publishers' group Thomson Reuters, see http://thomsonreuters.com/products_services/science/science_products/a-z/journal_citation_reports/ [May 2011]. JCR builds on citation data from the Science Citation Index and Social Science Citation Index databases, available through Web of Knowledge/Web of Science (formerly the ISI databases).

nals in the Journal Citation Reports. Another aspect of the 'internationality' of journals is the language in which they publish: among the educational research JCR journals, 89% are publishing in English, whereas journals publishing in German, Spanish and Turkish makes up about 2% each. In total, eleven languages are represented among the JCR educational research journals, French not being one of them.

These findings reveal a heavy bias towards publication in English from Anglo-American publishing houses in the WoS databases, with the consequence that the use of WoS based journal rankings is not a viable option for assessing research quality4, not the least since the intended international relevancy of the included publications cannot be proven. In the case of using WoS for assessing research from the educational sciences, international visibility as a quality criterion becomes translated to the visibility of a certain form of publications from a selection of national research spaces to the rest of the world. It substantiates the dominance of a 'minority' of regional and linguistic research areas, labeled as a 'majority' in terms of power relations and prosperity.

Consequently – and considering the lack of adequate coverage of European scientific publications from the SSH research areas – if European science and scientific institutions are evaluated using citation based metrics and WoS citation data, not only will individual researchers and institutions be widely ignored, but also, complete subject domains and language areas will be eliminated as contributors to the production of scientific knowledge.

A primary consideration when developing the EERQI-project, is the observation that many metrics based strategies for research assessment that may be appropriate for the 'hard sciences', are heavily criticized for their methodological weakness and lack of validity –not only from a social sciences and humanities point of view. Simultaneously, there is also a serious desire to develop approaches that serve better for detecting research quality. This desire unites the research community as well as stake holders from other related spheres, such as publishing houses, research funding institutions and policy makers.

The general intention of the EERQI-project was to develop useful tools supporting the process of quality detection. An intelligent combination of quality assessment tools – that was our assumption – would be able to assist the assessor in their task of determining the quality of the research being evaluated. The tools should meet two requirements:

a) They should increase the transparency and quality of the process of quality detection itself; and

[4] Aside from bias issues, there are also other problems using e.g. the WoS journal impact factor for assessing research, as discussed by e.g. Seglen (1997).

b) They should make the task better manageable and less time consuming.

It was not EERQI's objective to develop one single method, such as an indicator or an index for quality assessment. The aim was the development and testing of a set of prototype tools supporting the process of detecting research quality in texts – from the identification of a text relevant for a given question to the final assessment of the quality of the text per se.

The tools were developed within a broader prototype framework, where each tool addresses a specific part of the assessment process. Together, the instruments and tools form the EERQI Prototype Framework, allowing for an intelligent combination of different approaches complementing each other. These products and methods can serve as alternatives to citation based metrics in processes of quality assessment in SSH research. An important part of the prototype framework is the EERQI multilingual search engine and automatic semantic analysis tool, addressing issues of multilingual assistance in assessment procedures and tailor-made for strengthening the European research space. The EERQI Prototype Framework consists of:

- A content base with educational research texts in the four European languages included in the EERQI project: English, German, French and Swedish.
- A multilingual search engine including query expansion: an effective tool, capable of finding educational research texts on the Web in the four 'EERQI languages'.
- An automatic semantic analysis tool for the detection of key sentences in texts; applicable to educational research publications in (at least) the four 'EERQI languages'.
- A combination of bibliometric/ webometric approaches for the measuring 'extrinsic' indicators – i.e. indicators functioning as 'proxies' of quality.
- First tests of a citation analysis method with the potential for further development for the application to educational research (and other SSH) texts.
- A set of 'intrinsic indicators' – i.e. indicators immanent in the text per se – for the detection of quality in educational research publications, presented to, and positively evaluated by, the research community.
- An accompanying peer review questionnaire tested for reliability and practicality.
- A set of use-case scenarios advicing on how and when to use different combinations of the above-mentioned tools.

- Analyses to detect relations between 'extrinsic' and 'intrinsic' quality indicators.

The EERQI Prototype Framework attends to the full range of the process of detecting quality in research text. The process begins with the detection of potential quality through the identification of relevant texts from different sources, aspects that within the prototype framework is covered by the EERQI content base (educational research texts provided by the EERQI publisher partners) and the multilingual search and query engine (see chapters 2 and 3 in this volume). To address the assessment of texts through 'extrinsic' indicators, the 'aMeasure' application was developed: a stack of tools and programs to measure the impact of research publications, through e.g. citations and Web mentions (see chaper 6in this volume). To assist in the assessment of the internal qualities of a text, automated semantic analyses were developed and applied to identify key sentences, indicating which parts of documents the peer reviewers should pay particular attention to (see chapter 4). The process of assessing research texts through reading is also supported by a Peer Review Questionnaire containing a tested operationalization of the intrinsic indicators of quality that were developed within the EERQI project, supporting the readers' final judgment on the quality of a text.

The results of the EERQI project were presented on several occasions to the scientific community. These presentations addressed international educational research associations as well as individual experts in the field, representatives of research funding agencies as well as promotion and evaluation bodies at national and European levels. They took place in EERQI Workshops, as expert consultations and at international conferences. Since one aim was to apply the prototype framework on other SSH disciplines, the transferability was tested using the political sciences as an example (see chapter 11).

The EERQI project was developed within a truly interdisciplinary context in a European research consortium, bringing together a unique composition of experts in educational science, biblio- and webometrics, information and communication technology and computational linguistics, as well as European publishing houses. The results presented in this volume are small contributions towards the conscientious detection and assessment of research quality – independent of the scientific, cultural or linguistic area it comes from.

1 References

Brinkley, A. (2009). The Landscape of Humanities Research and Funding. Retrieved from http://www.humanitiesindicators.org/essays/brinkley.pdf

Halevi, G., & Bar-Ilan, J. (2013). Trends in Arts & Humanities Funding 2004-2012. *Research Trends, 3*. Retrieved from http://www.researchtrends.com/issue-32-march-2013/trends-in-arts-humanities-funding-2004-2012/

Hornbostel, S. (2011) Resonanzkatastrophen, Eigenschwingungen, harmonische und chaotische Bewegungen. *Vol. 9. Evaluation: New Balance of Power?* (pp. 7-14). Berlin: IFQ.

Mocikat, R. (2010). Qualitätsbewertung in den Naturwissenschaften mithilfe quantitativer Parameter: Ein Paradox? *Denkströme. Journal der Sächsischen Akademie der Wissenschaften,* (5), 90 - 102. Retrieved from http://www.denkstroeme.de/heft-5/s_90-102_mocikat

Seglen, P.O. (1997). Why the impact factor of journals should not be used for evaluating research. BMJ, 314 (497).

Finding Quality: A Multilingual Search Engine for Educational Research

Aaron Kaplan, Ágnes Sándor, Thomas Severiens, Angela Vorndran

Short Summary
To develop a field specific and multilingual search-engine, numerous algorithms are needed in addition to a general-purpose search engine. Here we describe the focal areas of development done in EERQI: Automatic classification for educational research, multilingual retrieval, query extension and relevance ranking. The classification algorithms, developed in EERQI enable a crawler to identify relevant objects with respect to a scientific field; the multilingual algorithms allow the retrieval of documents in several languages; query extension proposes related query terms to the user; relevance ranking is enhanced by semantic analysis.

1 An Automated Decider: Which Objects are Relevant for Educational Research?

Having a general web search engine, it would be impossible to decide which of the harvested objects are relevant for Educational Research, and which ones are not. One could only select the starting addresses for the crawling process wisely, but it would be impossible to detect new clusters of relevant material online in an automated way. To avoid this constraint, we developed and tested an algorithm deciding which of all crawled objects may be of relevance for Educational Research.

To train this machine-based learning algorithm, it was necessary to extract a number of full texts from the EERQI database of published articles and books. As the developed algorithm is highly sensitive to the language of the object to be tested, we had to train four different algorithms for the four EERQI languages: English, French, German, and Swedish. At least for the German and English algorithms, we had a sufficient number of training objects.

The technique used for the algorithms is quite old and well tested, but before the age of Cloud Computing, it was hard to find use-cases small enough to be implemented in real scenarios. Thus, one of the challenges was to boost the technical implementation and to make it usable.

I. Gogolin et al. (eds.), *Assessing Quality in European Educational Research,*
DOI 10.1007/978-3-658-05969-9_2, © Springer Fachmedien Wiesbaden 2014

The technology used for duplicate detection is described by e.g. Monika Henzinger (2006). Her work is based on algorithms developed by Broder in 1995-1997, who in turn refined algorithms described theoretically by Rabin (1981). The technology described by Henzinger is current state of the art for comparing big textual collections. Sorokina et.al. (2006) describe some basic rules, to reduce the number of shingles (an *k* words long phrase is called a *k-shingle*) to be handled, such as the rule to remove all shingles crossing sentence boarders, to remove capitalization, to replace stop words by an asterisk, etc. Empirical tests showed that 4-shingles are the optimum size for our deciding algorithm.

We made use of all these rules and trained deciding algorithms for all the four EERQI languages, using published articles and books as in-put. We were careful to train the algorithms, taking into account information on authors, publishers, from any genres and subfields in Educational Research. As the number of available publications in Swedish was too low, we decided to focus our activity on English, French and German. For the French algorithm, we had to add several articles from other sources to reach the necessary amount of documents for training, which is about 500 full texts. At the end, the tests showed that only the German and the English algorithms were usable, while the other two were unable to appropriately take into account information on subfields.

Part of the training procedure is to have a 'negative group' of full texts from other, but ideally adjacent fields, where phrases (shingles) available in both text collections are removed from the list of field specific phrases. At the end, one has a list of uni-lingual phrases (shingles) which are typical for Educational Research and represent the whole field. Most programmers call this kind of list a 'finger print'.

We made use of these finger prints to compare them with the list of shingles extracted from objects to be tested. If the percentage of shingles extracted from the object, and also being available in the finger print, exceeded a critical value (individually determined for every finger print case), an object was marked as being of potential relevance for Educational Research.

This service was coupled to the search engine using a REST-based[5] web-service. This allows other software to connect to our service in a defined and open way.

To test our algorithms, we used 50 relevant and 50 non relevant documents from the EERQI database and from other Open Access institutional repositories. Out of the relevant documents, the algorithm for English documents identified 91% as relevant, while 3% of the relevant documents were not identified. The

[5] REST: "Representational State Transfer", a dialect for a web service

corresponding results when testing the German algorithm was a recognition rate of 89% of the relevant documents while missing 5%, i.e. results that were slightly worse; and the French algorithm only recognized 73% while failing to identify 12% of the relevant documents. We could not develop and test a Swedish algorithm because there are too few publications available for training and testing.

The developed software, as well as all fingerprints are published under the BSD[6]-license on the EERQI web-server[7], to be reused by other projects. It already has been re-used in the field of biotechnology[8].

2 Multilinguality and query expansion

To enhance the field-specific search engine we built a software module that performs query translation and identifies relevant term suggestions, and we created a user interface that makes this functionality available to users via the web.

To support query translation and term suggestion, we use a number of different lexical resources: term networks that were compiled by DIPF and IRDP expressly for the purposes of this project, existing multilingual controlled vocabularies (TESE9, EET, and TheSoz10), and the general-purpose (i.e. not education-specific) query translation service from the CACAO project11 (the CACAO query translation service was graciously provided to the EERQI project by CELI). To translate a query, the software tries first the term networks, then the controlled vocabularies, and finally the CACAO service. For term suggestion, only the term networks and the controlled vocabularies are used. We also integrated into the query translation and suggestion software the same linguistic processing modules that were used in the indexer, so that the base forms of query words can be matched with the base forms of words in the indexed documents.

We built a web interface that allows the user to enter a query in English, French, German, or Swedish, and retrieve documents in one or more of these languages. Results for all desired languages are returned in a single list, ranked by estimated relevance to the query. When term suggestions are available, they are displayed (in the query language) next to the results. Clicking on a suggestion causes that term to be added to the query. In an earlier version of the interface we allowed the user to modify how the query was translated, but testing

[6] BSD-license: An open source license, formerly known as Berkeley Software Distribution
[7] Decider Software and Fingerprints published at: http://www.eerqi.eu/sites/default/files/EERQI-Classifier-and-Fingerprints.tar
[8] http://www.bibliometrie.info/forschung/teilprojekte.html
[9] http://ec.europa.eu/education/news/news1907_en.htm
[10] http://www.gesis.org/en/services/tools-standards/social-science-thesaurus/
[11] http://www.cacaoproject.eu/

indicated that some users were confused by this functionality, so in the current version the translation is displayed but cannot be modified. Users who are not satisfied with the automatic translation can simply use a monolingual search in the target language.

To determine how well the multilingual search functionality works and to identify opportunities for improvements, we performed several rounds of user testing of increasing size and formality. The initial rounds involved a few participants among the EERQI partners. After taking into account the feedback from the earlier rounds, we ran a larger set of tests in which education researchers worldwide were invited to participate.

While there are some testing methodologies for comparing cross-language information retrieval systems that have emerged as standards in the research community, these techniques are only applicable when the systems being compared are used to index the same set of documents, and when the query process consists merely of submitting a textual query and retrieving a list of results. Since the EERQI content base was compiled expressly for this project, it has not yet been indexed by any competing search engine; and since our search engine allows interactive query refinement via term suggestions, an evaluation methodology designed for one-shot query mechanisms is not applicable. In light of this, our goal in designing a testing methodology was not to compare our system directly to others, but to identify opportunities for improvement and to establish tools for tracking improvements from one version of our system to the next.

A number of independent factors affect the quality of search results, including coverage and quality of the collection being searched, of the lexical resources used, of the linguistic software for finding base forms, the appropriateness of the ranking formula, and the design of the user interface. To have a detailed understanding of the performance of the system, it would be interesting to design tests that isolate each of these factors. In some cases this would also facilitate comparison with other search engines. However, given the resources allocated, such detailed evaluation was out of the scope of the EERQI project. In some cases subsystems have already been evaluated elsewhere, e.g. the CACAO query translation system has participated in the CLEF evaluation campaign (Bosca and Dini 2009).

The evaluation methodology has two parts: quantitative analyses of user log data, and qualitative feedback in the form of a questionnaire and interviews.

Quantitative measurement:

Each time a query is submitted, the server logs an entry that includes a timestamp, the text of the query, the method that was used to submit the query (typing in the query box or clicking on a term suggestions), the query language and the requested result languages, and any term suggestions made by the sys-

tem. When a user clicks on a link in the result list to read a document, or advances to a subsequent page of results, these clicks are also logged and associated with the query from which the result list was generated.

In the first two weeks of the final round of testing, 1152 queries were logged in 289 sessions, where a session corresponds (roughly) to a series of queries made from the same computer within a period of ten hours. 46% of the queries submitted were cross-language searches. The total number of documents viewed was 516, or 0.45 documents per query on average. More specifically, in 81% of the cases, none of the results were viewed; in 10% of the cases one document was viewed; in 4% of the cases two documents were viewed; and in the remaining 5% of the cases three or more documents were viewed.

One measure of the quality of a query translation system is the ratio of cross-language search performance to monolingual search performance. With an ideal query translation system, one would find as many relevant results by using automatic translation as one does when searching in each language separately, resulting in a ratio of cross-language performance to monolingual performance of 1. In our tests, the average number of viewed documents per query was .30 for cross-language queries and .57 for monolingual queries, for a ratio of .53.

The system suggested additional terms for 81% of the queries. In cases where suggestions were made, the user clicked a suggestion 12% of the time.

Qualitative feedback:

All test participants were requested to fill out a questionnaire after using the system, but we made no attempt to enforce compliance with this request. We received 15 questionnaire responses, which is only 5% of the number of sessions observed on the search engine. Reactions were generally quite positive, but since the respondents were self-selected and the response rate was so low, statistics compiled from the responses would be difficult to interpret. The value of the responses is primarily that they describe problems that users encountered, indicating ways in which we can improve the search engine in the future.

In addition to the questionnaire, which was widely distributed via email lists, we contacted a small number of users personally to arrange telephone interviews to discuss their experiences in depth. We have performed five such interviews.

The most frequent comments in the questionnaire responses and the interviews were the following:

Many users requested an ”advanced search“ mode that gives more control over the search, particularly Boolean operators and constraints on metadata fields, e.g. constraining the search to documents published in certain years.

This remark was often linked to the complaint that a search returned ”too many results“, leaving the users with a need for options to cull the list. Since

results are ranked according to a scoring function giving higher scores to documents with more query terms, a document containing all query terms would be at the top of the list. Our expectation was that users would be reading the list from the top down, and then stops reading when they perceived that the remaining results were no longer relevant. However, feedback shows that many users read the whole list without considering any difference in relevance of the retrieved documents.

This mismatch in expectations is related to the difference between curated electronic library catalogs and web search. Curated collections typically have rich and reliable metadata, and support Boolean search with field constraints, whereas web search engines rely on ranking-based techniques with less user intervention in order to deal with noisier, non-curated data. Since the EERQI document base is a mixture of curated data from publishers and non-curated documents from the web, we chose to use a web-style approach, but testing revealed that many users were expecting a tool similar to a digital library. If we have an opportunity to develop the system further, we will approach this problem in two ways: by making a more fine-grained control of the search terms when possible; and by better managing user expectations, e.g. by explaining the ranking criteria.

Several users complained that the "title" metadata field was often missing or containing inadequate or irrelevant information. This is, again, a result of using documents crawled from the web, with metadata extracted by an error-prone automatic method rather than curated. It will never be possible to achieve 100% accuracy in automatically-extracted metadata, but there may be ways to improve on the methods we are currently using.

Translation of German compound words was often seen to be problematic. When a German compound word is not present in the term networks, its individual components are translated independently, and documents containing the translations are retrieved. This proved to be too broad in many cases. To narrow the search, it might be preferable to set as requirement that the individual components of translated compound words occur near each other.

3 Enhancing Relevance Ranking

In Chapter 4 we outlined a method for defining and detecting salient sentences in social science research articles. Similarly to the way content-oriented metadata – title and abstract - are used in digital libraries, we have used these sentences as additional metadata in the EERQI search engine, and we tested the performance.

The basic algorithm applied by the search engine includes term frequencies (TF) and inverse document frequencies (IDF) for ranking the retrieved documents. These measures are based on the frequency of occurrence of search terms in the documents. The so-called TF-IDF formula weighs the number of times a search term occurs in a document against the number of times a term occurs in the whole document collection. If a search term thus appears in one document frequently but only rarely or not at all in most of the other documents in the document collection, the document is ranked highly (cf. Manning et al., 2009).

The method developed for the EERQI search and query engine is meant to support the ranking of retrieved documents by assigning a higher weight to the query terms retrieved in sentences detected as salient sentences by XIP (see Chapter 5). We suggest that as a consequence the precision concerning the relevance of the retrieved documents will increase, since the likelihood that the query term represents the content of the whole document rises. While a retrieved term with the TF-IDF method can be located in any part of the document and thus may be irrelevant to the gist and main content of the article, a term retrieved in a salient sentence bears high resemblance to the general topic of the article.

In the following paragraphs we provide indications for comparing the results provided by basic EERQI search engine with the query "sport AND school". We evaluated[12] an article as relevant if its main topic was related to both school and sport.

We evaluated the relevance of the first 15 articles returned by the basic relevance ranking algorithm. Our evaluation found 3 relevant articles with respect to the query. None of these articles were selected as relevant by XIP.

XIP selects an article as relevant with respect to the query if it contains at least one salient sentence that contains both query words. We evaluated our tool on the 330 articles (out of the 1200 retrieved by the basic search engine) that contain at least one sentence with both query words.

Out of the 330 articles 85 were selected by our program, i.e. in 85 articles at least one salient sentence contained both query words.
The following list shows the human evaluation of these 85 articles:

- The number of relevant articles according to human evaluation: 23 (most of these are ranked low by Lucene)
- In 4 articles out of these the salient sentence is detected on an erroneously selected sentence
- The number of non-relevant articles according to human evaluation: 62

[12] The evaluation was carried out independently by the two authors. The inter-annotator agreement was almost 100%.

Analysis of the errors:

- Error due to format transformation[13]: 29
- The automatic sentence-type detection is correct but the sentence is not relevant with respect to the query: 15
- The automatic sentence-type detection is correct and the sentence is relevant with respect to the query, but the whole article is not relevant: 7
- Erroneous sentence-type detection: 11

Out of the remaining 245 articles, 35 have been evaluated as being relevant to the query. In these articles, we checked sentences containing both query words to search for salient messages that were missed by the tool, and we found one such example.

In all we found 58 relevant articles while evaluating our tool. They were all ranked low (beyond 100) by the basic ranking algorithm.

This test allowed us to conclude that salient sentences detected by XIP are indicators of relevance for queries, and they provide complementary results with respect to the TF-IDF method. Salient sentences have been given additional weight in the final EERQI search engine, and they are also used as snippets that present the retrieved documents.

4 References

Bosca A., L. Dini, Cacao Project at the TEL@CLEF Track. Working Notes for the CLEF 2009 Workshop, Corfu, Greece. ISSN: 1818-8044

Henzinger, Monika (2006): Finding near-duplicate web pages: a large-scale evaluation of algorithms. SIGIR '06 Proceedings. New York: ACM

Manning, C.D., Raghavan, P. & Schütze, H., 2009. *Introduction to Information Retrieval.* Online edition. Cambridge: Cambridge University Press.

Rabin, M. (1981): "Fingerprinting by random polynomials". Report TR-15 81, Center for Research in Computing Technology, Harvard University.

Sándor, Á., Vorndran, A. (2010): Extracting relevant messages from social science research papers for improving relevance of retrieval. Workshop on Natural Language Processing Tools Applied to Discourse Analysis in Psychology, Buenos Aires, Argentina, 10-14 May 2010.

[13] The retrieved articles (pdf, html, doc) are transformed to plain text for the NLP analysis

Sorokina, Daria; Gehrke, Johannes; Warner, Simeon; Ginsparg, Paul (2006): “Plagiarism Detection in arXiv”. http://www.computer.org/plugins/dl/pdf/proceedings/icdm/2006/2701/00/270101070.pdf

The EERQI Search Engine

Sybille Peters, Wolfgang Sander-Beuermann

Summary
Search engines typically consist of a crawler which traverses the web while retrieving any kind of documents, storing them in a database, and a search front-end which provides the user interface to the acquired information within that database. The EERQI search engine however is able to distinguish and retrieve just documents referring to the subject of this project. The search front-end gives sophisticated options to the user and is augmented by a multilingual interface. It accepts input in any of the four project languages (English, French, German, Swedish), showing results in each of these languages.

1 Introduction

The two basic tasks of the EERQI search and query engine are:

1. Finding new documents in the field of educational research within the WWW and making these accessible to users and project partners. This kind of search engine is called a "vertical" search engine (because it goes into the depth of a specific subject).

2. Besides the public WWW a second source of information is given by non-public educational research documents, provided by publishing companies being partners in the EERQI project. Indexing the content of these non-public educational research documents (subsequently referred to as *local document corpus)* by the use of intelligent search technology is the second basic task of the EERQI search engine.

The EERQI crawler is based on Nutch (Nutch, 2009), which is an open source web crawler, that is highly configurable and extensible via plug-ins. It is scalable across CPU clusters by incorporating the Apache Hadoop (Hadoop, 2009) framework. The following sections discuss the implementation of the search engine for the significant goals mentioned here.

I. Gogolin et al. (eds.), *Assessing Quality in European Educational Research,*
DOI 10.1007/978-3-658-05969-9_3, © Springer Fachmedien Wiesbaden 2014

2 The Crawler of the EERQI search engine

This type of crawler, designed to only harvest documents from the WWW with a specific content, is called a *focused* crawler. The Nutch crawler used within this investigation was substantially optimized to fulfill this task, because the Nutch software itself is not implemented for focused crawling but is extendable in this respect. The crawl is initialized with a seed list: a set of start URLs. Most of these start URLs have been selected from lists of electronic educational research journals . These URLs are injected into the Nutch crawl database ("crawldb"), which includes some information about each URL, such as the current status (e.g. fetched or unfetched) and time of last fetch. Each crawl cycle generates a list of top scoring unfetched URLs, or, URLs which need to be re-fetched. These URLs are then retrieved from the WWW and the resulting files are parsed. The URLs and corresponding anchor texts are also extracted and inserted into the link database ("linkdb"). This contains a list of inlink URLs and anchor texts for each URL. The parsed text is indexed if the document meets the Educational Research Document Detection (ERDD) criteria. A partial index is created for each crawl cycle. Duplicate documents are deleted from the indexes ("dedup"). At last, the indexes from each crawl cycle are merged into the final index. The modified status information for each URL is rewritten to the 'crawldb'. The score for each URL is adapted for EERQI focused crawling ("rescore"). Nutch uses the OPIC (On-line Page Importance Computation) (Abiteboul et al., 2003) algorithm to assign scores to each URL.

2.1 Focused Crawling Based on Link Analysis

A basic premise in OPIC and PageRank is (Abiteboul et al., 2003): a page is important, if important pages are pointing to it and important pages should be fetched first and more often. Within the EERQI crawler, we know which pages are important, aka relevant, as soon as we have fetched and analyzed them. These are the pages that have been indexed after being detected as Educational Research Documents (ERD). We must learn to predict, which pages will be important before they are fetched, and follow the most promising paths.

Some samples from the WWW have shown that the ERDs, most often do not link to other important ERD, if they link to anything at all. However, the pages linking to ERDs can be regarded as important pages, because they often consist of tables of content pages for an entire journal volume or year. They will not be indexed but are important in finding links to other relevant pages. It makes sense to use back-propagation for boosting the relevance score of pages

which link to ERDs. These pages are comparable to the hubs in Kleinberg's HITS (Hyperlink-Induced Topic Search) algorithm (Kleinberg, 1999). The HITS algorithm assumes that a good hub is a document, that links to many good authorities (authorities are important pages, comparable to ERD). Simply using the above mentioned link importance algorithms (such as OPIC, HITS or PageRank) is not feasible because we will not crawl a significant portion of the WWW and these algorithms do not take into account whether a document is an ERD.

The web may be displayed as a directed graph. Intuitively, an ideal crawl path would retrieve a very high number of ERD and a small number of non-ERD pages. The ratio of ERD pages to the total number of fetched pages should be as high as possible. When considering specific URLs, pages are important, if they link to a high number of pages classified as ERD. Indirect outlinks (outlinks of outlinks) will be considered up to a certain distance. Effectively, the high score of an ERD will be back-propagated to pages linking to it. The resulting score must then be passed on to the outlinks of these pages, until they reach a significant amount of unfetched pages.

2.2 Anchor Text Analysis

Analyses of anchor texts was taken into account as well. It may be assumed that words such as "pdf", "full", "article" and "paper" are good indicators of research documents but they do not contain any information about whether the referenced document is about educational research. The word "abstract" is a good hint, that the referenced document contains only an abstract, which is currently not considered as ERD by the search engine. SVMlight (SVMLight, 2009) was used to train the anchor texts. SVMlight is a Support Vector Machine based classifier. Single-word anchor texts that are a good indicator of a direct link to research texts ("pdf") obtained almost the same result as single words that would most likely not point to research documents ("sitemap" and "abstract"). It is assumed that this is due to the large number of non-ERD documents (for example research from other fields) that were also linked with potentially promising anchor text words. However, the classifier works well on anchor texts containing typical educational research terms, for example "Teacher" received a score of 4.28, "Learning" a score of 4.84. When training the classifier, not only the anchor texts with direct links to ERD were used, but also anchor texts of indirect links up to a level of three. A SVMlight score above 0 may be interpreted as a positive hit. The higher the score, the higher the probability of being in the trained class. The maximum score obtained in a list of 30000 samples was 4.89 while the minimum

was −4.99. While using this score may optimize the focused crawler, it may also bias the search engine towards documents with “typical” mainstream titles.

3 Educational Research Document Detection

Before analyzing how an ERD may be detected, we must first define the term ERD more precisely: An ERD is a digital scientific research document which may be classified within the topic “educational research”. It may be for example a journal article, a conference paper, a thesis or a book. An ERD may consist of one or more ERDs as in conference proceedings or entire journals. Abstracts are a part of an ERD but are not considered as a fully qualified ERD. Educational Research Document Detection may be regarded as a combination of identifying scientific research documents and a topical classification (educational research).

A large number of publications have analyzed the use of Vector Space Model based algorithms for document classification. Sebastiani (Sebastiani, 2002) provided an overview. These methods may be used for matching new documents with existing categories, such as specific topics (e.g. physics, biology), spam / no-spam etc. The document is represented as a vector. Each dimension of the vector represents a term, the value is a representation of the frequency that the term exists in the document (e.g. "term frequency/inverse document frequency" may be used). When classifying a document, the term vector of the document is matched with the term vectors of the classes. ERDD may be regarded as a binary classification problem, because there is only one class (ERD), or a ranking problem where the documents are sorted by their ERD ranking score. For supervised learning text classification, a collection of documents is required, which may be used as a training base. This collection should cover all areas of “educational research”. A negative collection should be provided as well, which covers documents that should not be considered as ERD, such as research documents from other fields and non-research documents. The detection mechanism is implemented using the following:

1. A rule based content analysis is used in order to ensure a high probability that the document is a research document. The document must have a minimum text length, it must contain a set of keywords (such as references, abstract) and it must contain references which may be existing in various formats.
2. A number of significant “educational research” keywords must exist in the document. Further work needs to be done to replace or augment this with a vector space model based classifier.

4 Technical Background about Database and Query Engine

Lucene is a software library which provides indexing and searching functionality. It takes its input data from the above described Nutch-based focused crawler. An index is used to provide rapid access to specific parts of the available information. *Lucene* makes use of an inverted index. An inverted index contains a list of terms with a reference to where they can be found in the text, similar to the index of a book. It is called an inverted index because it maps the terms to their location in the documents, contrary to a mapping from the documents to the terms. This results in high-performance lookups when querying for a term. *Lucene* uses a process called analyzing, which is explained in the next section, to create an index from a set of documents. However, *Lucene* is just a software library and not an indexing program in itself, so a program needs to be written which uses the *Lucene* library to create an index. *Lucene* is highly configurable and extensible and a number of design decisions must be made as to the construction of the index. When indexing text, the text may be indexed as one term (for example when indexing a URL) or it may be tokenized (see next section). The content may optionally be stored as well. Stored content is necessary if the content is to be displayed in the search results. An alternative is to store the content in an external data structure. An index created using the *Lucene* library consists of a number of *Lucene* documents. Each *Lucene* document contains one or more *Lucene* document fields. Fields are name/value pairs. For example the name could be "publisher" and the value "Symposium". These fields may be addressed when querying the index. It is possible to search for a specific author, title, publisher etc. The index is then used by the querying module of the search engine. A search query can consist of simple search terms or a more complex nested query with boolean operators.

Once the index has been created, *Lucene* can be used to obtain search results from the index. The input by the user is transformed into a query string which is used to query the index. The results of the query contain fragments of the original text with highlighted search terms (usually referred to as snippets) and additional information such as the title of the document. *Lucene* supplies several search possibilities. One can do a simple term search (example: blue searches for the term blue case-insensitively) or search for a range (example: 1997 TO 2000). It is possible to do wildcard queries (example: blue* searches for all words beginning with blue like bluetooth, bluefish, blueprint, blues etc.) or search within a specific field (example: creator:a* searches for all authors beginning with a). Searching for a phrase (e.g. "very blue") is also possible. Simple queries can be combined into a complex request using AND, OR and NOT (boolean query).

5 Multilingual Search Issues

Supporting multiple languages in a search engine raises additional issues that a monolingual search engine will not need to deal with. The *EERQI* search and query engine implemented some features that are not common in search engines today, for example the query translation. General issues that need to be addressed by multilingual search engines include:

- A multilingual search interface needs to be designed. This means that the user has the possibility of selecting his or her preferred language or the system selects it automatically. All text appearing in the user interface will need to be translated to the selected language.
- The text processing needs to be language aware, specifically the use of stemming, lemmatizing, usage of stop words, synonyms and thesauri. For the *EERQI* search and query language, lemmatization and decompounding was done by the *Xerox* tools for each project language. Multilingual thesauri were integrated.
- It should be possible to specify the language of the search terms to avoid cross-language ambiguity. For example if a user searches for the search term gut, this has a very different meaning in English and German and different results will be returned. In the *EERQI* multilingual search engine it is possible to choose the language of the search terms and the target languages.
- For cross-language searching, the query terms need to be translated automatically to the other languages.
- The documents that are returned as results may be automatically translated. Other issues that need to be addressed are for example the amount of user interaction.

Should the system for example select the term translations automatically, or should it be possible for the user to influence the selection resulting in a semi-automatic selection? If the system chooses the translated terms, should they be displayed or will this confuse the user? Studies conducted by Petrelli et al. (Petrelli, Daniela; Levin, Steve; Beaulieu, Micheline; Sanderson, Mark, 2006) showed that more users preferred the automatic query translations without user interaction even though the interface with user interaction achieved better results for recall and precision. Automatic translation may cause problems with polysemic terms, as there will be several possible translations with several very different results.

Multilingual Index Design

A new index design (subsequently referred to as version 2.0) is used for the multilingual search engine and incorporates the following features: it is optimized for cross-language queries and it uses XML files as input which are provided by the *EERQI* partner *Xerox*. These files contain already lemmatized and decompounded tokens. This functionality is not included in *Lucene* and has been provided as an enhancement to the indexing process. This is a language dependant feature which was available for the four project languages English, French, German and Swedish. Additionally, the key sentences that were provided by the *Xerox* parser have been added to the index. The multilingual search engine uses translations, thesauri and term networks to expand the query.

Solutions were found for a number of issues regarding the new index design:

- The previous index design used separate indices for the content base, the WWW etc. Using *Lucene* it was possible to address multiple indices at once for a query. As the multilingual user interface does not include the possibility of index selection, this feature is no longer necessary and all documents are inserted into one index.
- A feature of the new multilingual search engine is that query terms will optionally be translated to other languages. Without structural changes in the index this has the effect, that the *inverse document frequency* may be higher for terms of languages with fewer documents in the index (e.g. French) than for dominant languages (English). This will result in a distortion of the ranking with the undesirable effect of higher ranking of documents in the language with fewer documents. This was addressed by adding a language component to the field name (e.g. "title_en", "title_fr") and thus splitting up the languages. The query will then effectively be separated and each term only applied to the fields in the language of that term. This has the additional effect that cross-language ambiguities (one word existing in different languages and having different meanings) will not be an issue when querying because each term will be applied to the field for the correct language only. The splitting up via languages must be done for each field individually, as content, title etc. may exist in different languages.
- When documents of various languages exist, it is sometimes suggested to create an index for each language. This is not a helpful option for our document collection, because the collection may contain a document in

one language and a title and or abstract in another language or even several other languages.

- Another option is to index all text that will be used to query on into one language dependant field. While this may be a good idea for optimization purposes, it results in the problem that the title, abstract, content and key sentences will effectively be concatenated together and the length normalization (as described previously) will not boost title, abstract and key sentences as desired.
- In the multilingual search interface it is not possible to use field queries. For this reason, it was only necessary to index the fields that are to be used for a query by default: title, abstract, key sentences and content. These fields receive an additional suffix describing the language (e.g. "title_en" etc.). Fields that do not have a language component, such as the author are indexed without the language suffix.

6 Result Ranking

The ranking of a result within a *Lucene* query result set is referred to as the *similarity score* of a document. This is a measure which is calculated for each *Lucene* document in a result set and will determine the order of the results (ranking). The following factors are used in *Lucene* to calculate the *similarity score* (Gospodnetic, Otis; Hatcher, Erik, 2010):

- *Term frequency* factor: The importance of a document regarding a certain term increases proportionally to the frequency of a word in the document field.
- Normalization factor based on the number of terms in a document field. This causes a higher boost of terms found in short fields than in long fields. For example if a term is queried within the title of a document and the content of a document, a hit in the title will result in a higher normalization factor than in the content, because the title will obviously be shorter.
- *Inverse document frequency*: This is obtained by using the number of all documents (in the index) divided by the number of documents containing the term in the specified field(s). This has the effect, that rare terms which exist in only a few number of documents will have a higher impact on the ranking than terms that are very common.

- *Boost:* This is an optional boost factor that can be set during indexing. A boost factor may be specified for the entire document and for individual fields.
- Factor based on the number of query terms that are found in the corresponding document field. If more query terms are found in the document field, this value will be higher
- Factor based on query term boost factors. This is relevant for queries consisting of multiple parts.

7 Search Engine Front-end

The *EERQI* search and query engine offers search for specific documents using a combination of search criteria within the document full-text and metadata (Figure 1). The first version of the search engine provided the possibility of full use of the *Lucene* query syntax (OpenSearch).
Some of the features provided by the *Lucene* query syntax:

- Field search: it is possible to specify which fields the query terms are to be searching, for example the title in the field journal.
- Phrase search: it is possible to search for specific phrases, such as ”educational research”.
- Wildcard search: it is possible to search using wildcards, for example edu* will search for all terms beginning the prefix edu.

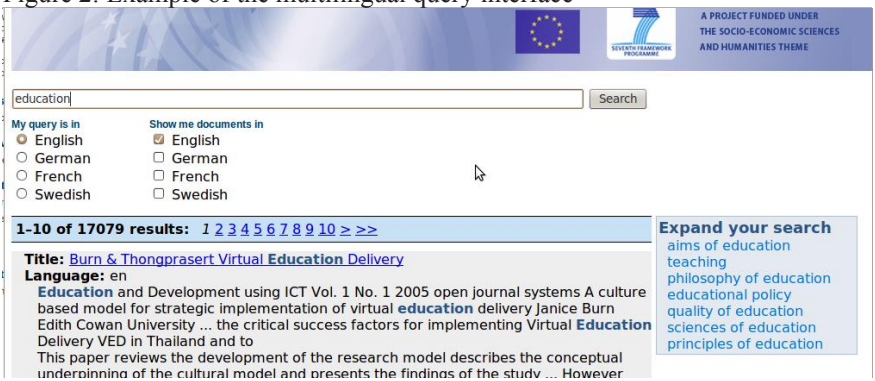
Several terms may be combined with the boolean operators OR, AND or NOT.

This version of the search engine is directed towards users who are familiar with constructing queries in this way. It is required for internal use within the project because it is possible to construct queries using the *Lucene* query syntax without limitations.

Figure 1: Example of the search interface with results

The multilingual search interface was designed by *Xerox* in cooperation with *RRZN* and *DIPF* (Figure 2). The user interface was tested during the final phase of the project. It provides the possibility to select one or more target languages from the project languages English, French, German and Swedish. Once one or more query terms are entered, additional suggestions are made with related phrases from the term networks and thesauri.

Figure 2: Example of the multilingual query interface



8 Prospects and further work

An educational research content base and search engine were successfully accomplished. It would be desirable to use the acquired knowledge and tools and apply them to other fields as well. The EERQI search and query engine should be used to make European educational research (especially work not published in English) more visible. An extension with other European languages within the multilingual search would be highly useful.

9 References

Abiteboul, S., Preda, M., and Cobena, G. (2003). Adaptive On-Line Page Importance Computation. In Proceedings of the 12th international conference on World Wide Web, pages 280–290. ACM.

Gospodnetic, Otis; Hatcher, Erik (2010): Lucene in Action, Manning Publications, Second Edition.

Hadoop (2009). Apache Hadoop. URL: http://hadoop.apache.org/.

Inverted Index, Wikipedia, http://en.wikipedia.org/wiki/Inverted_index (last accessed: March 28, 2011).

Kleinberg, J. (1999). Authoritative Sources in a Hyperlinked Environment. Journal of the ACM, pages 604–632.

Nutch (2009). Apache Nutch. URL: http://lucene.apache.org/nutch/.

OpenSearch, http://www.OpenSearch.org/ (last accessed: March 30, 2011).

Petrelli, Daniela; Levin, Steve; Beaulieu, Micheline; Sanderson, Mark (2006): Which user interaction for cross-language information retrieval? Design issues and reflections, Journal of the American Society for Information Science and Technology, John Wiley & Sons.
Sebastiani, F. (2002). Machine Learning in Automated Text Categorization. ACM Computing Surveys, 34:1–47.
SVMLight (2009). SVMlight. URL: http://svmlight.joachims.org/.

Highlighting Salient Sentences for Reading Assistance[14]

Ágnes Sándor, Angela Vorndran

Short Summary
The traditional process of the intrinsic evaluation of the quality of scholarly publications is peer reviewing. This is a time-consuming activity. In order to facilitate and speed up peer reviewing processes we have developed an exploratory natural language processing system implemented within the Xerox Incremental Parser for English, French, German and Swedish. The system highlights sentences that convey the most important threads of the article's content in order to focus the reviewer's attention on the design of the argumentation in the article. We have tested the results in several experimental settings.

1 Introduction

Peer reviewing is a very time-consuming assignment, and Natural Language Processing (NLP) technologies might provide tools that could shorten the time that peer reviewers take to process the articles.

We have set up this goal, and have developed a tool for providing assistance to peer reviewers in educational sciences, and in social sciences and humanities in general. We do not know of any other work with this perspective.

Detecting information in research articles is a long-standing and important task of Natural Language Processing. Information extraction tools usually provide structured pieces of factual information conveyed by digital texts, primarily in text genres where the messages of are mostly factual. Among academic disciplines this holds primarily for exact sciences. However, salient messages of social science texts are typically not facts, but arguments, interpretations, analyses, etc. Thus traditional information extraction technologies are not suitable to grasp them. Our approach consists in highlighting salient sentences in the articles that can be regarded as the logical backbone of the article.

[14] This chapter contains the revised version and extension of Sándor and Vorndran (2009).

I. Gogolin et al. (eds.), *Assessing Quality in European Educational Research,*
DOI 10.1007/978-3-658-05969-9_4,

Our tool does not evaluate, but aims at focusing the evaluator's attention on the parts of the texts that are relevant as a basis for his/her judgment. Nor does this tool check if the texts conform to some formal norms of scientific writing.

We regard highlighting salient sentences as a complement to the processing guidance that the structural layout of the articles provides. The structural layout of scientific articles – title, abstract, keywords, section headings – guide the reader in processing the logical, argumentative and content-wise development of the article at different levels: The title is the brief indication of the topic, the keywords yield the conceptual context of the topic, the abstract provides a concise summary of the problems and results, and the section headings guide the reader step by step in the development of the article. Besides these waymarkers, the highlighted salient sentences are meant to be an intermediary representation of content development between the title, the keywords, the abstract and the section headings on the one hand and the whole article on the other hand.

Since we define salient sentences as those sentences that sum up the main messages of the articles, and since peer reviewing consists in judging the scientific value of the main messages, we assume that highlighting salient sentences both helps understanding and provides evidence for the peer reviewer's evaluation. By highlighting we intend to add a relevant and coherent dimension of the representation of the flow of the article, which is otherwise hidden, and which the reader has to discover in order to understand the article.

Highlighting is carried out using the Xerox Incremental Parser (XIP) (Ait-Mokhtar et al., 2002).

2 Related Work

Our work is in line with the growing amount of research in documentation sciences and natural language processing that takes into account the argumentative structure of research articles in tasks such as information retrieval, information extraction, navigation within documents and summarization.

In the domain of information retrieval as far back as the beginning of the 1990's Liddy (1991) claimed that additional functions for search instruments could benefit from including the discourse-level context of the retrieved search terms in the interpretation of the results. Liddy stressed the "semantic roles" of concepts in a document as opposed to the simple occurrence of search terms. Oddy et al. (1992) proceed in this line of research and state that discourse-level structures in research texts could be useful to support retrieval for the user because they represent structural qualities recognized by the reader independent of the topic of the research. Both concentrate on the analysis of abstracts of research

articles and propose a system to combine topical with structural information in the retrieval process.

Kando (1997) also emphasizes the importance of the discourse-level context of search terms in the retrieved documents. The allocation of retrieved passages to functional units and thus the possibility to gain information about article structures provides a valuable opportunity to improve the user's assessment of the retrieved documents. A similar method of annotating text passages according to their function in the text is conducted by Mizuta et al. (2006) with the objective of categorizing articles in different document genres.

Teufel and Moens (2002) base automatic summarization on extracting sentences annotated with respect to their discourse function in the text.

Lisacek et al (2005) detect sentences in biomedical articles that describe substantially new research based on analyzing discourse functions.

Another line of research to exploit the argumentative structure for navigation and information extraction is inspired by the semantic web. Instead of automatically discovering argument structures in texts, the approach aims at creating conceptually motivated processing editors in which the users insert content according to its argumentative function. (See for example Uren et al., 2007, Couto and Minel, 2007.)

3 The Structure of Educational Research Articles

Research articles in the educational sciences tend to display a very heterogeneous structure, like articles in many other fields in social sciences and humanities. While the thematic contents of the articles are structured according to the requirements of the topic, frequent occurrences of a unifying structure are introductory and concluding chapters. However, where these chapters appear they do not display uniform headings (cf. Fiedler, 1991:98). Likewise Ruiying and Allison (2004) show that the structure of research articles in linguistics does not conform to a common model, and section headings in many cases do not refer to the function of the chapter but to the thematic contents. Brett (1994) and Holmes (1997) observe basic structural features in the articles in political sciences and sociology. They state, however, that the section headings are usually not standardized.

The structural heterogeneity in the social sciences and the humanities, - multidisciplinary fields with close connections to fields of professional practices - , derives from the coverage of a wide range of research problems and the consequential variation of the methods applied. This field includes theoretically embedded discussions as well as empirical studies or material for school praxis.

These differences in the referenced subjects are reflected in the way the research articles are organized and presented. Montesi and Owen (2008:151) notice a high grade of liberty granted by the educational sciences journals for the presentation of submitted papers. They also describe a clear distinction between qualitative and quantitative approaches in research articles, the latter displaying a closer connection in structural aspects to the exact sciences than the former.

In contrast to the heterogeneity of the structure and section headings of research articles in social sciences and humanities those in the hard sciences show a relatively uniform structure, and often follow the well-known pattern of Introduction – Methods – Results – Discussion, which renders their reading easier.

In the framework of this study we compared the structural properties of fifteen articles from three journals: the British Journal of Educational Studies (BJES), the Educational Psychology Review (EPR) and the International Journal of Educational Research (IJER). These are educational research journals covering a wide variety of topics from educational psychology to school instruction. We have made the following observations:

a) Some section headings follow the functional structuring of natural science articles, some do not. About half of the articles contain an 'Introduction' and/or a 'Conclusion', one third has a 'Methods' section and 26% of the articles has a section entitled 'Results', 'Findings' or 'Conclusion'. Thus a basis for a functionally oriented article structure can be perceived in the first and last chapters of most of the articles. Nearly 60% of the section headings, however, are oriented towards aspects of the content of the articles and show no predefined form.
b) All of the articles are preceded by an abstract and eleven of them have keywords assigned to them.

   The keywords play an important role in our highlighting approach, since they are supposed to convey the basis for topical relevance. The number of keywords assigned per article is between two and nine. While some keywords are applied only a few times in the article, others are used 60 or even over 100 times. In some cases the keywords are very common words ('teachers', 'education') and they are used frequently throughout the text. In these cases the highlighted sentences are supposed to indicate relevant, terminological uses of those common, non-specialised words. In other cases the keywords are rare, but they are terms used in specialized contexts, for example, terminological expressions related to the field of research. Those are very useful for a quick overview over the research topic. Keywords appearing very rarely or not at all in the text itself often belong to a more general level of terminology.

From an information extraction point of view the importance of the terms in the thread of the article is known to be related to their places of occurrence: in the title, the abstract, and the section headings or even in the titles of the bibliography terms have more significance than in the rest of the article. This property of terms is used in search options in digital libraries. An appearance of the query term in the introduction or conclusion could also be a hint for the term being relevant for the scientific context or the results of the study whereas terms referring to the methodology or rather non-specific terms do not convey much information about the central contents of the text.

c) The abstract is supposed to sum up the most important aspects of a research article. The articles analyzed show that in general the sentences in the abstract correspond to assertions made throughout the articles in most of the different sections. In a few cases most sentences of the abstract were also taken up in the introductory or concluding part of the article with a summarizing function.

4 The Detection of Salient Sentences

In defining the characteristic features of salient sentences that serve as a basis for their detection we rely on the kinds of judgments peer review evaluations are supposed to make (Bridges 2008).[15] We sum up these judgments as follows: the relevance of the topic, the clarity of the problem statement, the coherence of the argumentation and the well-foundedness of the conclusions. These criteria of judgment are often presented as questions in the evaluation forms that peer reviewers are asked to fill in. Based on these evaluation criteria we define salient sentences as sentences that describe research problems, purposes and conclusions related to the topic of the articles as indicated by the keywords.

The salient sentences receive two types of labels in our system: SUMMARY – the sentences that convey either the goal or the conclusion - or PROBLEM – the sentences that mention research problems. Some sentences get both labels. Labeling is carried out by rules, which rely on the conceptual definition of SUMMARY and PROBLEM sentences.

The following examples illustrate summary sentences:
The purpose of this article is to develop the idea that …

[15] In a preliminary experiment we tried to identify salient sentences in an example-based way. Six scholars marked the salient sentences in four articles from four domains according to the same evaluation criteria. There were hardly any overlaps. This led us to define salient sentences.

The perspective I shall use in this essay relies heavily on the view …
This paper explores …
Taken together, the study indicates …

Summary sentences express argumentative functions and announce themes all along the article through metadiscourse expressions, as the sentences above illustrate. They explicitly convey the discursive development of the article, and thus they are supposed to reiterate the development announced in the abstract: they state aims, claims, conclusions, present the subject matter, problems, methods, etc. All of the argumentative roles of the summary sentences imply the presence of salient messages in them, but we also propose other kinds of sentences as bearers of salient messages: sentences that convey in some way the problems handled in the article. The reason for this is twofold: on the one hand, the authors do not systematically use summary sentences as they develop their article, and on the other hand, the automatic detection is never exhaustive.

Besides summary sentences our tool is designed to detect another kind of sentence as bearer of salient messages. The definition of this kind of sentence is motivated by the consideration that the *raison d'être* of every research article is to contribute to the development or solution of a research issue. However, the explicit expression of the research issue by metadiscourse similar to that of synthesizing in the summary sentences is relatively rare (cf. Ruiying & Allison, 2004). We grasp the salient expressions concerning the author's contribution to the development or solution of research issues in sentences that speak about contesting, questioning or pointing out as significant or new research-related ideas, facts, or theories, indicate a gap in knowledge, or point out any flaw or contrast related to the research topic (cf. Sándor & Vorndran 2010). We will refer to these sentences as "problem" sentences. The following sentences convey research problems in our sense:

My interest of inquiry emerged in 1997 from a new idea in school pedagogy and sport pedagogy.

This sentence points out the author's new idea in school pedagogy and sport pedagogy that will be detailed in the subsequent sentences in the article.

With an absence of detailed work on masculinities and sport in South African primary schools (for an exception, see Bhana 2002) this paper goes some way towards addressing the issues around young boys' developing relationship with sport.

This sentence describes a flaw concerning previous research and proposes to carry out some of the missing work.

> *However, the rest effect, the first order and the second order effect, some are negative effects and some are positive effects which contrast with prior research results due to two main reasons that ...*

In this sentence the author gives reasons why some concepts contradict prior research.

While the category and role of summary sentences corresponds to traditionally recognized rhetorical or discourse functions, the concept of problem sentences is less straightforward to define.

Contrary to summary sentences, as the examples above illustrate, problem sentences do not fulfill particular rhetorical, argumentative or discursive functions considered in textual analysis.

The recognition of this category of sentences is motivated by the theory of scientific progress developed by Kuhn (1962), which is based on the conception of science as essentially a problem solving activity. With the category of problem sentence we aim at capturing the expressions of some kinds of summaries of the problem solving activity at sentence level, as expressed in scientific discourse. These sentences can also be considered as synthetic sentences, but not at the level of the argumentative development of the article but at an argumentation-independent level, which aims at capturing the theoretical issues discussed in the article.

5 Tests

In order to assess the validity of our approach we have carried out three tests that evaluate the performance of the system from three points of view. The first test assesses the effectiveness of the assistance in the peer-reviewing process with respect to a process where no assistance is yielded. The second test examines if the system provides the same results when used with different sub-genres within the educational science literature. Finally the last test compares the output of the system with summaries provided by educational scholars.

5.1 The effectiveness of the assistance in the peer-reviewing process

Six scholars evaluated five articles with and without highlighting from three educational science journals: the British Journal of Educational Studies (BJES), the Educational Psychology Review (EPR) and the International Journal of Educational Research (IJER). In a table they marked in both cases the time needed for the evaluation as well as the notes they gave from 1 to 5 for the five intrinsic quality criteria determined by the EERQI project: significance, originality, style, integrity and rigour (Table 1).

Table 1. Comparative study of peer-reviewing highlighted and not highlighted articles

Scores		1.BJES1						2.BJES2						3.EPR1						4.EPR2						5.IJER					
		1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
Time	HL	6	5	20	8			5	5		6			6	5		8			2	5	20	3			1	5	10	4		
	FR	17	15	50	45			15	15		28			21	40		33			33	50	45	35			12	30	50	?		
Signifi-cance	HL	4	3	-	2			4	3		3			4	5		2			4	5	-	-			4	5	-	2		
	FR	3	4	4	1			4	3		4			4	5		4			3	5	4	3			4	5	5	2		
Origi-nality	HL	4	5	-	3			4	3		3			5	5		2			-	5	-	1			-	5	-	2		
	FR	4	5	4	2			4	3		3			5	5		4			3	5	4	2			4	5	5	3		
Style	HL	5	4	- (5)	1			3	2		3			-	5		3			-	5	-	-			-	-	-	3		
	FR	4	4	5	1			4	5		4			4	5		3			4	5	4	3			4	5	4	3		
Integ-rity	HL	4	-	-	2			4	3		4			-	5		2			-	5	-	-			-	-	-	-		
	FR	4	5	5	2			4	5		4			4	5		4			3	5	4	4			4	5	4	2		
Rigour	HL	-	-	-	2			-	2		4			-	-		1			-	-	-	-			-	-	-	-		
	FR	3	5	5	1			5	5		4			5	5		4			3	5	4	4			3	5	4	1		

Our conclusions are the following:

- Highlighting allows to evaluate according to the criteria of significance, originality and style, but not according to integrity and rigour
- Highlighting makes it possible to rapidly filter out bad quality: processing the highlighted texts took 4 times shorter time.

5.2 The application of highlighting in different sub-genres

Within the same journals mentioned in the previous test we analyzed 36 papers in different genres in 3 domains: sociology, psychology and history. The sentences automatically selected as salient sentences were evaluated by the two authors as for their correctness: i.e. if they convey a summary or a problem within the article. In sociology and psychology there were 2 sub-genres: theoretical and empirical articles (Table 2).

Table 2. Comparison of results according to article genres

DOMAIN	THEORETICAL		EMPIRICAL	
	SUMMARY (error-rate)	RES. ISSUE (error-rate)	SUMMARY (error-rate)	RES. ISSUE (error-rate)
sociology	4% (4%)	17% (2%)	10% (16%)	15% (19%)
psychology	3% (21%)	21% (61%)	7% (7%)	11% (50%)
history	3% (24%)	17% (12%)	-	-

The first number is the percentage of the sentence type automatically detected out of all the sentences and the second number is the error rate of the automatically detected sentences according to the manual evaluation.

In theoretical articles the proportion of problem sentences detected is substantially higher than that of summary sentences, whereas in empirical articles this difference is much smaller. The ratio of the two sentences can thus be an indicator of the sub-genre. This difference is expected: theoretical articles focus on solving research problems, whereas empirical articles have more clearly formulated goals.

The other observation is that the error-rate is high in problem sentences in psychology articles. This is due to the fact that psychology articles treat questions related to problem-solving, which is exactly the content of research-issue sentences. However, in research-issue sentences that we aim at detecting problem-solving belongs to the theoretical issues of the paper, whereas in psychology articles it belongs to the subject-matter of the paper. This problem is very difficult to overcome.

The results show that the method is effective in all the domains and both sub-genres, and it also clearly shows the differences between the sub-genres: while theoretical articles contain more research-issue sentences, empirical articles contain more summary sentences. These differences in publication cultures in sub-genres of one discipline should be taken into consideration when comparing and interpreting automatically determined values.

5.3 Comparison of the highlighted sentences with peer-reviewers' summaries

We asked scholars in educational sciences to briefly summarize the goals, problems and conclusions described in the articles evaluated in the EERQI peer-review exercise (see Chapter 8). We determined for every sentence in the expert summaries if it is comparable to one or several sentences in the article. (The criteria of comparability are described in De Liddo et.al. in press). If we found comparable sentences, we determined if they fulfilled the criteria of salient sentences, i.e. if they described a summary or a problem. Finally we tested if the sentences were automatically highlighted or not. This test indicates to what extent the automatically highlighted sentences are considered relevant for the reader in the comprehension of the article.

We evaluated 189 summaries of 44 articles in English and 123 summaries of 25 articles in French. The same article was summarized by several persons. The majority of the articles were theoretical articles of the philosophy or the history of educational science (Table 3).

Table 3

	French		English	
Percentage of the sentences of the expert summaries that are comparable to sentences in the articles	48%		81%	
Percentage of the comparable expert summary sentences that satisfy the conditions of salient sentences	40%		57%	
Percentage of the sentences in the row above that are detected by XIP	53%		70%	
Estimated precision of XIP	97,7%		96%	
Average number of sentences per article	detected by XIP	written in the expert summaries	detected by XIP	written in the expert summaries
	23	5	48	6
Correlation of salience and frequency of sentences used in the summaries	Number of times the same article sentence appears in different summaries	Percentage of salient sentences	Number of times the same article sentence appears in different summaries	Percentage of salient sentences
	1 2 3 4	42% 66% 100% 72%	1 2 3 4	42% 60% 76% 71%

	5 6	100% 100%	5 6 7 8 9 10	100% 100% 100% 100% 100% 100%

Taken together these results show that the automatically detected sentences cover a considerable proportion of human summary sentences. This suggests that automatic highlighting does have the potential of providing key sentences for peer-reviewers.

These three testing scenarios represented different aspects of how automatic detection of salient sentences can be included in scientific work contexts. Problems that might appear in the application have been addressed and the quality of results has been evaluated. It has thus been shown that in regard to effectiveness of work completion, applicability to different intra-disciplinary contexts and comparison to intellectual execution of the task the tool shows positive results.

6 Conclusion

We have presented an exploratory system for highlighting salient sentences in order to support the peer reviewing process. The selected sentences are supposed to help peer reviewers of articles in educational sciences to focus their attention on some relevant textual evidence for formulating their judgments. We have argued that even if the structural characteristics— the abstract, the keywords and the section headings—guide the reader in following the development of the article, content-oriented highlighting of salient sentences might enhance a rapid understanding of the core contents.

Although the subjects of educational science research articles display very heterogeneous structures and contents, the system could identify a number of sentences containing the main statements of the articles. Text-inherent developments not accompanied by structural signs like the outcomes of empirical studies or the contents of a theoretical discussion about abstract terms could be identified using automatic text analysis, and this can possibly save intellectual effort of scientists. The time-consuming task of reviewing a growing number of research publications, hardly manageable when studying each submitted manuscript thoroughly, could thus be facilitated and supported.

The results of our tests suggest that the salient sentences detected are relevant for peer reviewing, since they describe the problems, aims and results in the articles. We have found that sentences conveying definitions, especially in theoretical articles, should also be highlighted as key sentences.

7 References

Aït-Mokhtar, Salah, Chanod, Jean-Pierre and Roux, Claude (2002): Robustness beyond shallowness: incremental dependency parsing. Natural Language Engineering, 8(2/3):121-144.

Brett, Paul (1994): A genre analysis of the results section of sociology articles. English for Specific Purposes, 13(1):47-59.

Bridges, David (2008): Criteria of Quality in Educational Research. Working Group Report of the 1st EERQI Workshop, 20-21 June 2008. Leuven. Project Internal Document.

Couto, Javier and Minel, Jean-Luc (2007): NaviTexte : a Text Navigation Tool. Artificial Intelligence and Human-Oriented Computing, Lecture Notes in Artificial Intelligence, 4733, Springer, Berlin, Heidelberg.

De Liddo, Anna, Sándor, Ágnes and Buckingham Shum, Simon (in press): Contested Collective Intelligence: Rationale, Technologies, and a Human-Machine Annotation Study. Special issue of the CSCW Journal on "Collective Intelligence in Organizations".

Fiedler, Susanne (1991): Fachtextlinguistische Untersuchungen zum Kommunikationsbereich der Pädagogik dargestellt an relevanten Fachtextsorten im Englischen. Lang, Frankfurt a.M.

Holmes, Richard (1997): Genre analysis, and the social sciences: An investigation of the structure of research article discussion sections in three disciplines. English for Specific Purposes, 16(4):321-337.

Kando, Noriko (1997): Text-level structure of research papers: Implications for text-based information processing systems. Proceedings of the 19th British Computer Society Annual Colloquium of Information Retrieval Research, Sheffield University, Sheffield, UK, 68-81.

Kuhn, T. S. (1962): The Structure of Scientific Revolutions, Chicago: Univ. of Chicago Pr.

Liddy, Elizabeth D. (1991): The discourse-level structure of empirical abstracts: an exploratory study. Information Processing and Management, 27(1):55-81.

Lisacek, Frédérique, Chichester, Christine, Kaplan, Aaron and Sándor, Ágnes (2005): Discovering paradigm shift patterns in biomedical abstracts: application to neurodegenerative diseases. First International Symposium on Semantic Mining in Biomedicine, Cambridge, UK, April 11-13, 2005.

Mizuta, Yoko, Korhonen, Anna, Mullen, Tony and Collier, Nigel (2006): Zone analysis in biology articles as a basis for information extraction. International Journal of Medical Informatics, 75(6):468-87.

Montesi, Michaela and Owen, John Mackenzie (2008): Research journal articles as document genres: exploring their role in knowledge organization. Journal of Documentation, 64(1):143-167.

Oddy, Robert N., Liddy, Elizabeth D., Balakrishnan, Bhaskaran, Bishop, Ann , Elewononi, Joseph and Martin, Eileen (1992): Towards the use of situational information in information retrieval. Journal of Documentation, 48(2):123-171.

Ruiying, Yang and Allison, Desmond (2004): Research articles in applied linguistics: structures from a functional perspective. English for Specific Purposes, 23(3):264-279.

Sándor, Á., Vorndran, A. (2009): Detecting key sentences for automatic assistance in peer reviewing research articles in educational sciences. In Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries, ACL-IJCNLP 2009, Suntec, Singapore, 7 August 2009 Singapore (2009), pp. 36--44.

Teufel, Simone and Moens, Marc (2002): Summarizing scientific articles: experiments with relevance and rhetorical status. Computational Linguistics, 28(4):409-445.

Uren, Victoria, Buckingham Shum, Simon, Mancini, Clara and Li, Gangmin (2007): Modelling Naturalistic Argumentation in Research Literatures: Representation and Interaction Design Issues. International Journal of Intelligent Systems, (Special Issue on Computational Models of Natural Argument, Eds: C. Reed and F. Grasso), 22(1):17-47.

Enhancing Relevance Ranking of the EERQI Search Engine [16]

Ágnes Sándor, Angela Vorndran

Short Summary
In this chapter we describe the application of the detection of salient sentences for enhancing relevance ranking in the EERQI search engine as well as for the presentation of document snippets in the results lists. In a proof-of-concept experiment we show that the presence of the query word(s) in the salient sentences detected is an important indicator of the relevance of the article. We have compared the relevance of the articles retrieved with our method with those retrieved by the Lucene search engine as configured for the EERQI content base with the default relevance ranking, which is based on word frequency measures. The results are complementary, which points to the utility of the integration of our tool into Lucene.

1 Introduction

The EERQI project developed a publicly available multilingual search engine (makalau/xrce.xerox.com/eerqi) dedicated to the retrieval of educational research papers in the documents in the project data base both from heterogeneous data source collections and from the web. The search engine retrieves educational science research literature as an example, but it is proposed to be usable for retrieving research literature in social sciences in general. The EERQI search engine uses the freely available Lucene library. As an inbuilt functionality Lucene ranks the results according to an algorithm that uses TF-IDF measures, i.e. it takes into account the frequency of the query words in the document and the inverse document frequency which relates to the occurrences of search terms in all documents of the collection.

In the framework of the EERQI project we propose to improve both relevance ranking and the presentation of the information snippets of the retrieved documents by integrating into Lucene the detection of salient sentences. In order

[16] This adapts and and extends Sándor and Vorndran (2010).

I. Gogolin et al. (eds.), *Assessing Quality in European Educational Research,*
DOI 10.1007/978-3-658-05969-9_5, © Springer Fachmedien Wiesbaden 2014

to justify our choice we have carried out a proof-of-concept experiment the results of which are promising.

2 Proof-of-concept Experiment

In Chapter 4 we outlined our method for defining and detecting salient sentences in social science research articles. In order to show in an empirical experiment that the sentences detected do in fact carry salient messages, we have used these sentences as additional metadata in the Lucene search engine, similarly to the way content-oriented metadata are used in digital libraries, and we tested whether salient sentences can be used successfully as supporting material for document retrieval.

The search and ranking algorithm applied by the search engine Lucene, which was used in the selection of the documents from the EERQI project content base, includes term frequencies (TF) and inverse document frequencies (IDF) for ranking the retrieved documents. These measures are based on the frequency of occurrence of search terms in the documents. The so-called TF-IDF formula weights the number of times a search term occurs in a document against the number of times a term occurs in the whole document collection. If a search term thus appears in one document frequently but only rarely or not at all in most of the other documents in the document collection, the document is ranked highly (cf. Manning et al., 2009).

The method developed for the EERQI search and query engine is meant to support the ranking of retrieved documents by assigning a higher weight to the query terms retrieved in sentences detected as salient sentences by XIP (see Chapter 4). We suppose that as a consequence the precision concerning the relevance of the retrieved documents will increase, since the likelihood that the query term represents the content of the whole document rises. While a retrieved term with the TF-IDF method can be located in any part of the document and thus may be irrelevant to the gist and main content of the article, a term retrieved in a summary sentence or problem sentence bears high resemblance to the general topic of the article.

In the following paragraphs we provide indications for comparing the results provided by Lucene and those of using XIP.

We retrieved 1200 research documents with the EERQI search engine with the query "sport AND school". We evaluated[17] an article as relevant if its main topic was related to both school and sport.

We evaluated the relevance of the first 15 articles returned by Lucene with the basic relevance ranking algorithm. Our human evaluation found 3 relevant articles with respect to the query. None of these articles were selected as relevant by XIP.

XIP selects an article as relevant with respect to the query if it contains at least one salient sentence (i.e. a SUMMARY or a PROBLEM sentence) that contains both query words. We evaluated our tool on the 330 articles (out of the 1200 retrieved by Lucene) that contain at least one sentence with both query words.

Out of the 330 articles 85 were selected by our program, i.e. in 85 articles at least one SUMMARY or PROBLEM sentence contained both query words.

The following list shows the human evaluation of these 85 articles:

- The number of relevant articles according to human evaluation: 23 (most of these are ranked low by Lucene)
- In 4 articles out of these the salient sentence is detected on an erroneously selected sentence
- The number of not relevant articles according to human evaluation: 62
- Analysis of the errors:
- Error due to format transformation[18]: 29
- The automatic sentence-type detection is correct but the sentence is not relevant with respect to the query: 15
- The automatic sentence-type detection is correct and the sentence is relevant with respect to the query, but the whole article is not relevant: 7
- Erroneous sentence-type detection: 11

Out of the remaining 245 articles 35 have been evaluated as being relevant to the query. We checked if in these articles the sentences that contain both query words express salient messages that were missed by the tool, and we found one such example.

In all we found 58 relevant articles while evaluating our tool. They were all ranked low (beyond 100) by the basic Lucene ranking algorithm.

[17] The evaluation was carried out independently by the two authors. The inter-annotator agreement was almost 100%.

[18] The retrieved articles (pdf, html, doc) are transformed to plain text for the NLP analysis

3 Conclusion

In this experiment we compared the results of the basic frequency-based relevance ranking algorithm of the Lucene search tool used in the EERQI search and query engine with the content-based selection method of the detection of salient sentences. The results show that the relevant articles returned by Lucene among the top ranked articles and those selected by our tool are disjoint, i.e. the two approaches are complementary. Since our tool, despite its very strict selection rule (the presence of both query words in a sentence labeled as expressing a salient sentence), returns a considerable number of relevant articles that would appear late in Lucene's ranked list[19], we consider that our approach is promising and that the integration of the two tools is beneficial for the user. This experiment helped us to identify a number of systematic error types which we hope to be able to fix, and thus improve the precision.

4 References

Manning, C.D., Raghavan, P. & Schütze, H., 2009. *Introduction to Information Retrieval.* Online edition. Cambridge: Cambridge University Press.

Sándor, Á., Vorndran, A. (2010): Extracting relevant messages from social science research papers for improving relevance of retrieval. Workshop on Natural Language Processing Tools Applied to Discourse Analysis in Psychology, Buenos Aires, Brasil, 10-14 May 2010.

[19] The most obvious cases where statistical ranking of unstructured document collections is bound to fail are relevant articles in a collection of articles on various topics or articles written in a foreign language containing an English abstract.

Studies in Correlative Assessing of Intrinsic and Extrinsic Indicators of Quality

Stefan Gradmann, Frank Havemann, Jenny Oltersdorf

Short Summary
Taking up work done by other partners in EERQI and more specifically the quality indicators identified we tried to establish a significant correlation between these intrinsic quality indicators and available extrinsic indicators (bibliometric, webometric and usage indicators, including new resources from the 'social web'). Although the data used were partly incomplete and fragmentary in some respect testing of uni-variate and linear correlations was not successful and any correlation one could imagine would probably be non-linear and complex. As a consequence, the most plausible conclusion for the time being seems to assume complementarity rather than correlation of intrinsic and extrinsic indicators of research publication quality.

1 Introduction

The assessment of research quality is one of the most important, yet one of the most difficult aspects of the scientific process. Evaluation procedures are in the center of many debates in academic, professional, and public policy circles. In the multidisciplinary field of educational research, an important issue in the debate is the lack of consensus on specific standards for assessing research quality and of a commonly agreed definition of the concept of quality.

The traditional method of evaluation is the judgment by peers. Advantages and disadvantages have been extensively discussed in the literature. (Bornmann, 2008; Cicchetti, 1991; Williamson, 2003) One often mentioned disadvantage of peer review is that papers are assessed based on the reputation of the author rather than their quality. Also, the process is time consuming and expensive; and very often the review is performed either by narrowly specialised scholars who are unable to compare different projects, or by people with broad scientific qualifications, but without the specific insight required to evaluate the quality of a submitted paper. As a consequence, evaluation bodies increasingly tend to use quantitative methods supposed to be more objective. The range of quantitative

I. Gogolin et al. (eds.), *Assessing Quality in European Educational Research,*
DOI 10.1007/978-3-658-05969-9_6, © Springer Fachmedien Wiesbaden 2014

methods used for research assessment is broad. The best known, but also the most debated methods comes from the field of bibliometrics / scientometrics / webometrics. Indicators like the total number of articles published by an author, the h-index, g-index, the age-weighted citation ratio, the impact factor and many more are used with the hope of reducing error and increasing accuracy of assessment.

Nevertheless, there are a lot of problems facing research quality assessment today. One main problem is the insufficient coverage of Social Science and Humanities research publications in traditional bibliometric databases. Another is the aforementioned lack of a reasonable definition of the concept of research quality in e.g. the field of educational research. The first problem makes the use of conventional bibliometric data sources highly debatable, the second has implications for the trust and fairness of peer judgements, as well as on the question of what actually should be measured. To overcome these problems a new approach based on the analysis of correlations between peer judgements and bibliometric measures was proposed and scrutinized in the EERQI project.

2 Related research activities

Citation analysis, as part of quality assessment tools, is limited by the bibliographic databases where citation data is gathered. This is the main target of criticism of the method. Citations in publications not indexed by these databases are simply lost. That is why new data sources need to be examined regarding their coverage and their usability for impact measures. Several researchers have investigated new quantitative methods for research impact evaluation to enhance traditional citation analysis (Xuemei, Thelwall and Giustini 2011; Kolowich 2010; Burgelmann, Osimo and Bogdanowicz 2010; Priem and Hemminger 2010; Thelwall 2003; 2008; Moed 2005). In the literature, there are two main strategies: one is the examination of WWW usage, and the other citation analysis based on WWW-based data sources. The first mentioned strategy evaluates the impact of a paper or a single researcher through potential readership statistics, e.g. article online views, clicks or downloads. The most ambitious attempt at this is the project MESUR (Bollen, 2010). The project, does not limit itself to one single metric indicator, but utilises a whole range of types and facets of usage metrics.

The second approach mentioned extends traditional citation analysis to the WWW. In an article published in 2001 Blaise Cronin argued that: “Citation analysis is an important piece of the bibliometric research pie; one that will become even more central with the growth of the web and for a very simple reason.

The links (reference citations) provided routinely by authors in their reports and papers are a means of exposing the underlying socio-cognitive structure of science." (Cronin, 2001 p. 2) Making use of the infrastructure of the WWW, today's researchers have a wider range of diverse options to communicate and disseminate their findings than ever before. These options include (open access) repositories, online journals, and Web 2.0 applications such as blogs, wikis, social bookmarking tools, Twitter and online reference management systems. Based on this infrastructure Cronin stated that: "After all, citations and 'sitations' are not merely similar phonetically ... Highly linked sites are the web's equivalent of highly cited papers." (Cronin, 2001 p. 2)
A third new trend occurred with the growth of reference management systems and their combination with social network features. (Xuemei et al., 2011) This third approach overlaps with web citation analysis, but intends to make use of the facilities that reference management systems can provide to track scholarly influence from users.

Regarding the problem of the lacking definition of research quality we think that the meaning and interpretation of research quality is strongly related to the intentions and purposes of the assessing body, as well as the performance objectives, and to the mission of the entity being evaluated. Determining the quality of a piece of research necessitates scrutinizing the research processes and the research outputs. The most often used and best measurable research output is the dissemination of published research in the form of research articles.[20]

David Bridges, Professorial Fellow in the University of Cambridge Faculty of Education and Emeritus Fellow at St Edmund's College, Cambridge was member of the EERQI project team. He argued, that "quality assessment requires a judgement, a form of connoisseurship, based on a widely informed encounter with a situated text rather than anything which can be adequately captured by measurement". (Bridges & Gogolin, 2011) Nevertheless, using metric indicators is exactly what the successor of the British Research Assess Exercise (RAE), The Research Excellence Framework is suggesting for future research assessment. "It is widely expected that the ratings will initially be derived from bibliometric-based indicators rather than peer review. These indicators will need to be linked to other metrics on research funding and on research postgraduate training. In a final stage the various indices will need to be integrated into an algorithm that drives the allocation of funds to institutions." (The use of bibliometrics to measure research quality in UK higher education institutions, 2007 p.2)

[20] We did not take into account others forms of research output like oral contributions to a workshop, or lectures since the EERQI project proposal was aiming at the quality measurement of written research texts solely.

There have been several attempts to seize the relationship between academic impact measured via citations and research quality. (Hornbostel, 1991; Hornbostel, 2001; Norris & Oppenheim, 2003; Smith & Eysenck, 2002) It was found that there is a correlation between assessments of results of research output based on bibliometrics and peer evaluation. If this is also the case in educational research is the topic for this paper.

3 Research Carried Out

3.1. Methodology
Intrinsic and extrinsic research quality indicators

Right from the beginning of the EERQI project it was clear that a stable and commonly agreed definition of the concept of research quality in the field of educational research was needed. The educational experts in the project agreed that the concept of research quality in educational research texts is rather difficult and complex, and for that reason it was decided to distinguish between intrinsic and extrinsic indicators of research quality of education research texts. What is integral to the quality of a text and what inherently constitutes elements of quality? What are the more indirect quality indicators of a research paper? The project team defined the terms as follows: Intrinsic indicators of the quality of a research text were those which were considered to be integral to the quality of that text, which are constitutive of that quality, which are a condition of judging it to be of high quality. Since quality consists e.g. in the coherence and consistency of the argument, and in the validity of the methods employed, the evidence of coherence, consistency or validity can be considered intrinsic indicators of the quality of the writing. Extrinsic indicators are those which do not inherently constitute elements of the quality of the piece, but which have a positive correlation with judgements based upon such elements. Extrinsic indicators correlate with the quality that can independently be discerned in the text. Extrinsic indicators have a "probabilistic" relation with quality.

Within the project, *rigour*, *originality*, *significance*, *integrity*, and *style* were identified as intrinsic indicators of research quality. As a result of later discussions integrity and style were discarded as being too difficult to identify and only the first three indicators were actually retained. Mentions in online reference management systems, usage, and citation information were considered as relevant extrinsic quality indicators.

3.2. New data sources

Besides the traditional databases Web of Science and Scopus we suggested in 2010 the use of additional new data sources to calculate citation based metrics. The aim was to overcome the problem of lacking coverage of educational research published in other languages than English and in other formats than journal articles. Today, citations are no longer the only source of impact metrics and Web of Science is not longer the only database for bibliometric measures: the WWW itself can be mined for impact indicators. Jason Priem, researcher in the field of Information and Library Science and one of the first who investigated the viability of assessing scholarly impact over the social web instead of traditional citation analysis, stated in an article published in 2010: "Just as the early growth of Web–supported webometrics and usage–based metrics, the current emergence of "Web 2.0" presents a new window through which to view the impact of scholarship. These days, scholars who would not cite an article or add it to their Web pages may bookmark, tweet, or blog it. Arguably, these and related activities reflect impact and influence in ways that have until now eluded measurement." (Priem & Hemminger, 2010)

For the above reasons we proposed to work with online reference management tools. "Many scientists now manage the bulk of their bibliographic information electronically, thereby organizing their publications and citations material from digital libraries." (Hull, Pettifer, & Kell, 2008) The use of online reference management systems like Mendeley, CiteULike, and Connotea is increasing continuously. We think that these systems present an opportunity to create new data resources for quantitative measures. Metrics based on a diverse set of e.g. online reference management systems could yield broader, richer, and timelier assessments of current and potential scholarly impact.

Reference management software is a class of applications developed to assist in the process of compiling bibliographies and managing textual bibliographic records in one or more databases. Originally, in the 1980s, these applications were developed to facilitate the task of writing papers by managing the references. Over the last few years, they have evolved significantly, and can now be seen as a tool for the entire management of textual databases. Reference management systems like CiteULike, and Mendeley, have also incorporated social and collaborative features (Duong, 2010). These features enable the users to share a personal library within a private or public group and to decide at what level to collaborate and be found by other researchers working in the same area. Users may also look for citations in the collective library that are similar to those stored in one's own library. By allowing researchers to expand their bibliographic records and eventually interact with other researchers in their field, collabora-

tive reference management systems have a potential of growing into resource discovery environments.

In addition to the citation indicators based on Web 2.0 applications, Google Scholar and Web of science, we also decided to incorporate the analysis of a second group of indicators: web usage. The advantages of usage data as part of impact measures lies in the chance to record interactions for all types of scholarly content, i.e. papers, journals, preprints, but also for blog postings, software etc. Since the measurement of these interactions can start immediately after the publication, it is a very rapid indicator of scholarly trends[21]

3.3. Research design

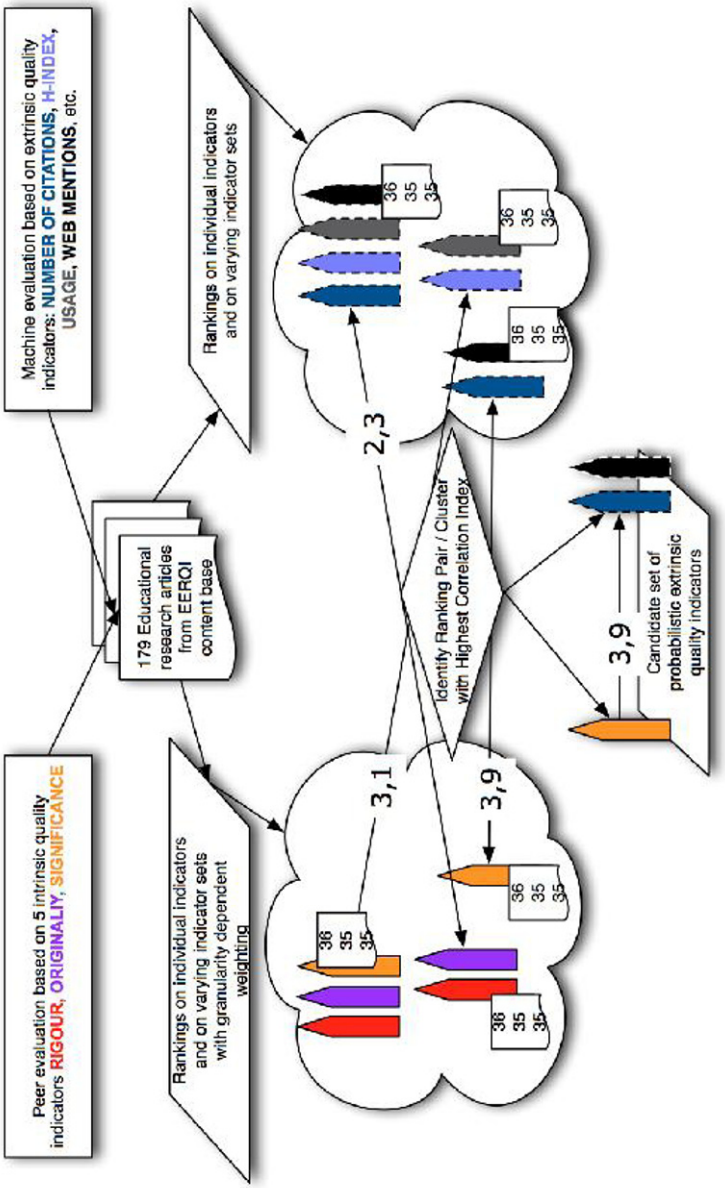
In the course of the EERQI project a proposal for analysing the relation between assessment results based on extrinsic metrics and assessment results based on intrinsic indicators was made. The intrinsic indicators were operationalized and transferred into items of a peer review questionnaire. We intended to do a comparative and weighted analysis of a ranking based on the results of this scaled questionnaire and a ranking obtained from the extrinsic indicators in several iterations.

The underlying assumption is that there is one specific combination of extrinsic indicators correlating the best with one particular intrinsic indicator. By discovering this combination of extrinsic indicators we were aiming at being able to make statements such as: The weighted combination of the extrinsic indicators "mentioning of article in Mendeley", "mentioning of article in Connotea", and "mentioning of article in CiteULike" corresponds best to the intrinsic indicator originality. Or: A ranking based on citations per paper gathered from Google Scholar weighted 2 times corresponds best to a ranking based on the average score on the indicator significance.

This part of the research strategy is illustrated in figure 1 below:

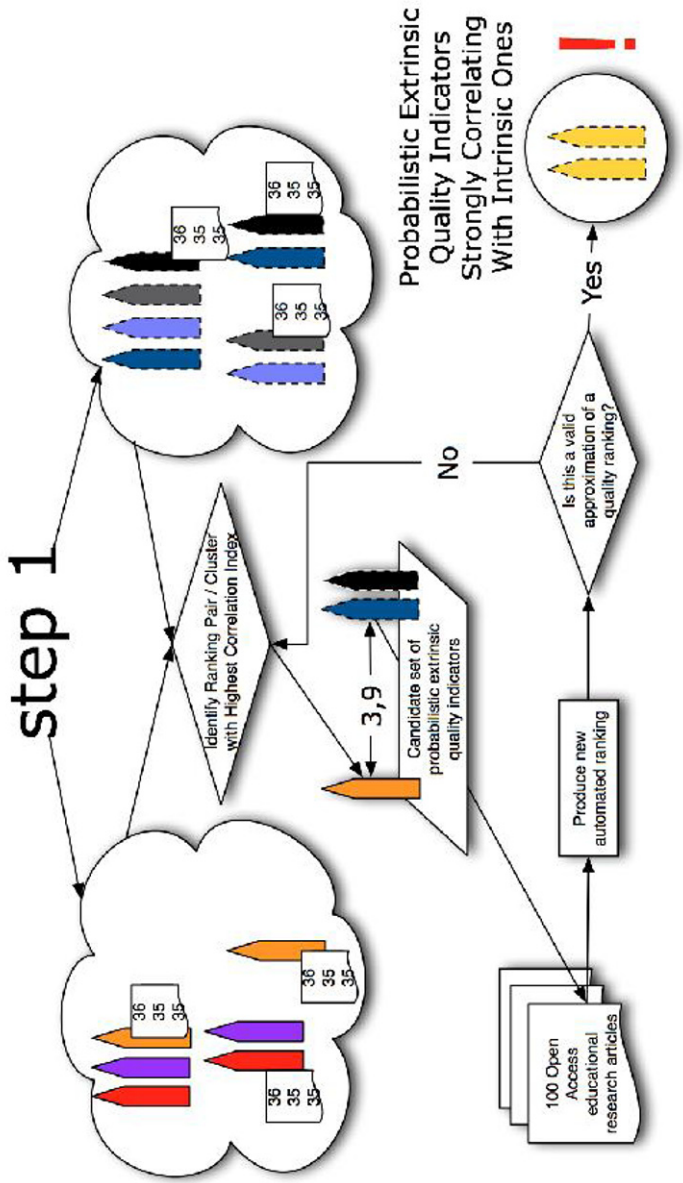
[21] We are aware of the fact, that the need for rapid publication and citation of research information is more characteristic for the STM field than for e.g. the area of educational research.

Figure 1: Correlation Identification Methodology Initial Steps



Originally, the first step of analysis was supposed to be followed by an iterative process as depicted in figure 2 below. It should be noted that these further iterations were never carried out because of various synchronization problems within the project, resulting in important delays in the initial steps. Furthermore, the first results of the initial step analyses did not encourage further investigation.

Figure 2: Further Iterations within the Correlation Identification Methodology



To obtain and compile the actual values for the above mentioned indicators we developed a piece of software called aMeasure. It is a stack of functions to measure extrinsic characteristics of research publications using Google Scholar, Google Web Search, MetaGer, Library Thing, Connotea, Mendeley, and CiteULike (Stoye & Sieber, 2010). In the context of the EERQI project aMeasure was used to collect information about extrinsic characteristics of educational research publications. It consists mainly of 4 parts:

1. a crawler to gather all information from Google Scholar (GS), Google Web Search and the Social Network Services,
2. a database to store the gathered information,
3. a client side application (JAVA-applet), and
4. a web interface to present the results and the content of the database to end users.

The main component of aMeasure is the crawler. For optimal work the crawler needs to be provided with author names. It has turned out that a major challenge in measuring extrinsic characteristics of research publications is the reliable identification of author names in the Social Network Services, GS, Google Web Search, and MetaGer. We have therefore based our attempts on the findings presented by Derek Ruths and Faiyaz Al Zamal in the paper: "A Method for the Automated, Reliable Retrieval of Publication-Citation Records" published in 2010. In this paper they present a series of filters to apply at results returned by an online publication search engine. One of these filters is a so-called name matching filter. Ruths and Zamal conducted several queries and retrieved "that when such a search is performed, the backend algorithm selects publications by applying a lenient filter to author names." (Ruths & Al Zamal, 2010 p.3) They found that slight modifications of the authors name have a significant impact on the initial set of candidate publications returned by the search engine and therefore recommended to use the following query syntax: author: "the first name of the author the initials of the middle names the last name of the author". Using this syntax the crawler queries GS for the authors and all of their papers. This is done via Screen-Scraping. In addition Google Web Search, MetaGer and the Social Network Services are queried to get information about the impact of each author's paper. The process of crawling is done on a central server located at Humboldt-Universität zu Berlin and it is constantly running in the background. As Google has limited the number of requests to an unknown randomly selected amount per IP per day the crawler is subject to this limit too. If this limit is reached and a user intends to search for an author's name which has not been

already stored in the central database, a Java-applet is querying GS instead of the crawler.

All gathered data are stored in a central Mysql database located on the EERQI server to enable various exports via the web interface. GS is used to retrieve information about authors, their papers, and the citations of these papers. Due to the fact that Google does not provide an API aMeasure is required to use a technology called Screen-Scraping. The same technology is used to query MetaGer and the Social Network Services. A more comfortable method is used for retrieving results from Google Web Search and Mendeley, which are providing APIs to their search engines. These web search engines are queried with every single paper and the name of the author, for example: "Sahra Ahmed" + "Disablement following stroke". The results are then presented via a web interface.

Relying on the "name filter" solely is not a suitable, sufficient criterion to discern the publications that belong to a given author. Since many individuals share the same last name, many more share the same first name. Taking this into account we integrated a second filter, which ensures that the publications fall within the time span of an author's career. As we do not see how to get hold of each authors individual curriculum vitae we decided to limit the search results to the last 60 years arguing that an author is unlikely to start publishing before his/her 20th birthday and after his/her 80th year of life.

In addition to the aforementioned strategies, we also take into account the results of a prototype classifier developed by ISN Oldenburg that has been refined and trained for analysing educational research literature. This classifier contains a fingerprint of word shingles (strings of defined length) typical for relevant publications in educational research. The classifier can be queried via an API for the probability of a given publication (identified by its URL) to determine whether it is an educational research publication or not.

We also considered using author affiliations for a more precise author disambiguation, but decided against it for a number of reasons. One problem is the matter of lacking standardization and stability of institutions names. Names of institutions change over time, and different authors use different formats for writing the name of the same institution. Another problem leading us to abandoning author-affiliation matching, even if the aforementioned problems could be reduced by matching author and city of affiliation instead of department or institution, is that authors change affiliation and move to other cities. To be able to calculate e.g. the h-index or citations per paper for an author, we need to retrieve full information on publications from the author. If we delimit the search by filtering the results based on one city, we will lose information on publications by the author written in other cities at earlier or later stages of the career, thus creating a distorted picture of the authors publication activities. For in-

stance, a search for “Stefan Gradmann” + “Berlin” resulted in much fewer hits than “Stefan Gradmann” + “Hamburg”, although we knew from the curriculum vitae that it is one and the same person in both cases.
The following extrinsic characteristics can be retrieved and calculated from GS using aMeasure:

- Number of papers per author.
- Number of citations per author.
- Year – first year of retrieved publication until last year of retrieved publication.
- Citations per year.
- Citations per paper.
- The h-index provides a single-number metric of an academic's impact. A scientist has index h if h of his/her Np papers have at least h citations each, and the other (Np−h) papers have at most h citations each. The h-index is calculated based on the full list of an authors output and the obtained citations. The h-index is robust in the sense that it is not sensitive to papers with few or no citations. However, the h-index is also insensitive to a limited set of papers with outstandingly high citation frequencies, which becomes a drawback when we have authors with relatively few publications but high citation rates.
- To take into account the above mentioned problem with the h-index, we also use the g-index, a development of the h-index giving more weight to highly cited articles. (Egghe, 2006)
- The e-index is aiming to differentiate between scientists with similar h-indices but different citation patterns. (Zhang, 2009)

The following extrinsic characteristics can be retrieved and calculated from Google Web Search and MetaGer using aMeasure:

- Google Web Search hits matching the author’s name.
- MetaGer hits matching the author’s name.
- The following extrinsic characteristics can be retrieved and calculated from Social Network Services using aMeasure:
- CiteULike hits matching the author's name and the articles title.
- Library Thing hits matching the author's name and the articles title.
- Connotea hits matching the author's name and the articles title.
- Mendeley hits matching the author's name and the articles title, readers of article in Mendeley.

Unfortunately, GS and Google Web Search present an estimated result count only, since every user and every API request is not able to retrieve more than the first 1000 results for a specific search request. Regarding Google Web Search the company has shut down their old XML-API which enabled users to get very close to these 1000 results. Currently the Google-AJAX-API is limited to 64 search hits. If the Google Web Search reaches 64 hits, we are using "Screen Scraping" of Google Web Search to get the full list of results.

We learned that a robust method to identify authors is essentially needed as it is the critical step in making it possible to automatically track all the contributions that a researcher has made. This problem is very well known. In 2006 Elsevier launched its service "Scopus author identifier". The author identifier assigns a unique number to the authors who have published articles in journals covered by Scopus. An algorithm distinguishes those with similar or identical names on the basis of their affiliations, publication history, subject areas and co-authors (Qiu, 2008). There are other initiatives for dealing with author identification and disambiguation. In 2007 CrossRef invited a number of people to discuss unique identifiers for researchers, and in 2008, Thomson Reuters launched ResearcherID. In an article in PLoS Comp Biol, Bourne and Fink argue that one solution to this difficulty is OpenID, a standard "That means that an identity can be hosted by a range of services and people can choose between them based on the service provided, personal philosophy, or any other reason. The central idea is that you have a single identity which you can use to sign on to a wide range of sites. There are two major problems with OpenID. The first is that it is poorly supported by big players such as Google and Yahoo. Google and Yahoo will let you use your account with them as an OpenID but they don't accept other OpenID providers. More importantly, people just don't seem to get OpenID" (Bourne & Fink, 2008) The whole issue of identifying and distinguishing between authors is surrounded by problems, both in terms of the development of technical solutions as well as coming to an agreement on standards for dealing with the problems associated with author identities; at the same time as a solution is critical for fair metrics based research assessment.

Currently aMeasure is filtering self citations with the help of GS. By using GS it is possible to search within all citations a paper has received. By subtracting all citations where the author of the original paper is also the author or co-author of the citing paper from the total amount of citations the paper has received we can filter out self citations. This technique prevents us from analyzing all citations manually, which would involve many queries to GS and would reduce the amount of papers and authors we are able to analyze per day. As some authors published a lot of papers receiving many citations, and as there is a daily limit to the number of GS sets per user or IP per day, this solution seems to be

the most effective in terms of returning hits in a reasonable time. From our point of view, tools like CleanPoP do not seem to address the problem of limited Google search returns. A further drawback of CleanPoP is the necessity to manually select author names and possible duplicates. This means that every single citing paper needs to be analyzed.

3.4. Source Data

The Publishing houses Symposium, VS-Verlag, Barbara Budrich Publishing, Taylor and Francis Publishing as well as the DIPF (German Institute for International Educational Research), IRDP (Institut de Recherche et de Documentation Pédagogique) and INRP (Institut National de Recherche Pédagogique) delivered nearly 6000 educational research publications (journal articles and book chapters) in the languages German, French and English and helped building the EERQI content base. Additional 42.000 educational research open access documents were crawled and added to the content base. Since most of the documents were in PDF format without sufficient metadata or XML-based structure, citation analysis within the EERQI content base could not be carried out as originally intended.

3.5. Analysis

For the analysis of correlation between intrinsic and extrinsic indicators a sample of 179 paper assessments based on intrinsic criteria was used in combination with two files of related extrinsic data:

- citation numbers of rated papers obtained with Google Scholar (on March 8, 2011)
- data from search engines and social-network services.

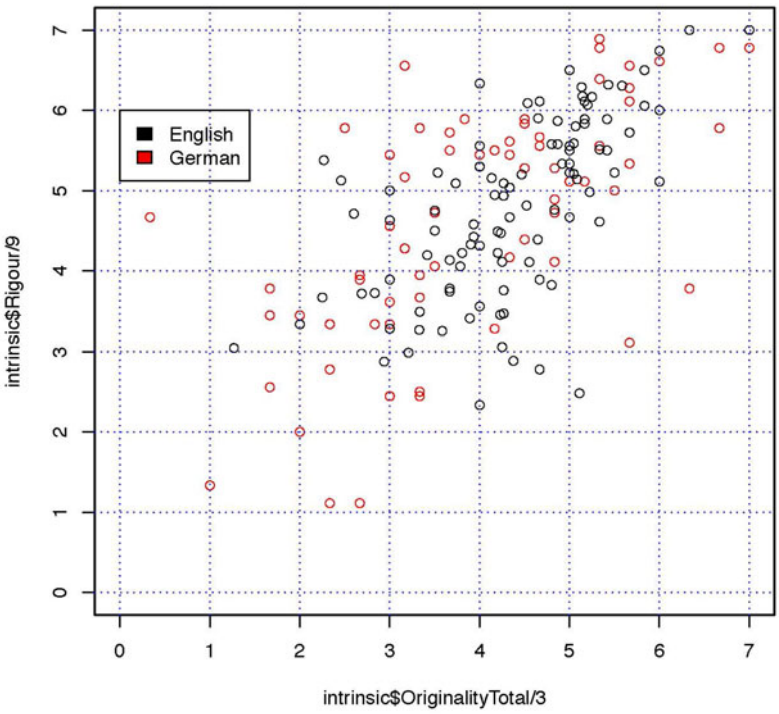
As the extrinsic author data generally suffered from homonymic authors we only used paper attributes. Papers were in English and in German and distributed over three thematic groups:

- Group 1 includes papers about "assessment, evaluation, testing & measurement" (35 / 35)
- group 2 about "comparative and inter-/multicultural education" (33 / 17)
- group 3 about "history and philosophy of education" (34 / 17)

We first had a closer look at the interrelation between the three remaining intrinsic indicators which each received a respective average of nine, three and four ratings of different aspects. This resulted in a combined rating score for each paper: the average ratings of all 16 aspects total score on a scale from 0 to 7.

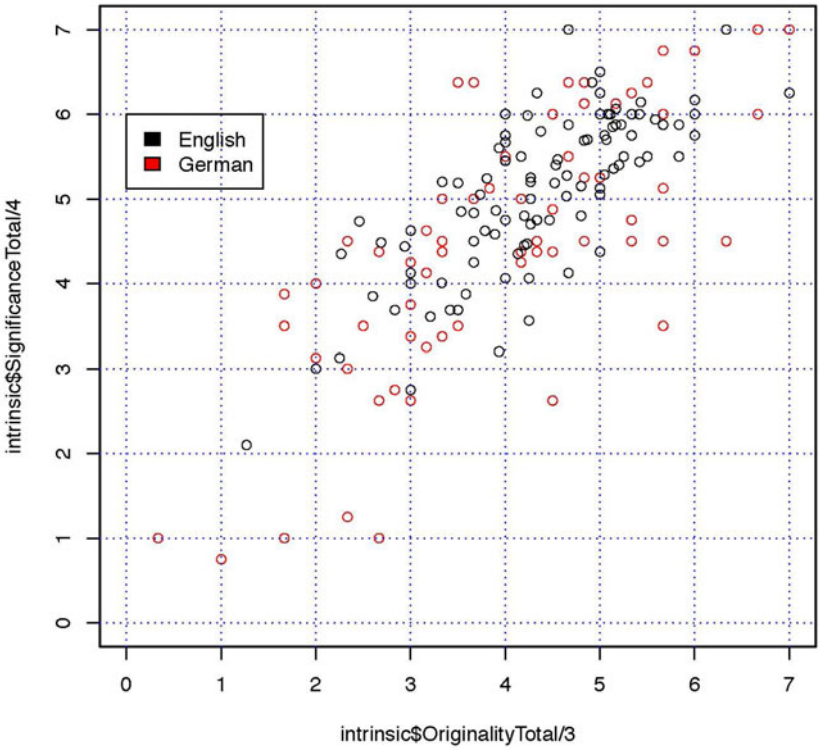
The scatter plots in the three figures of mean scores of rigour, originality, and significance show that the latter two correlate best, especially for English-language papers. This is evident when comparing the low correlation strength in the interrelation of originality and rigour as shown in figure 3 below:

Figure 3: Originality - Rigour Interrelation



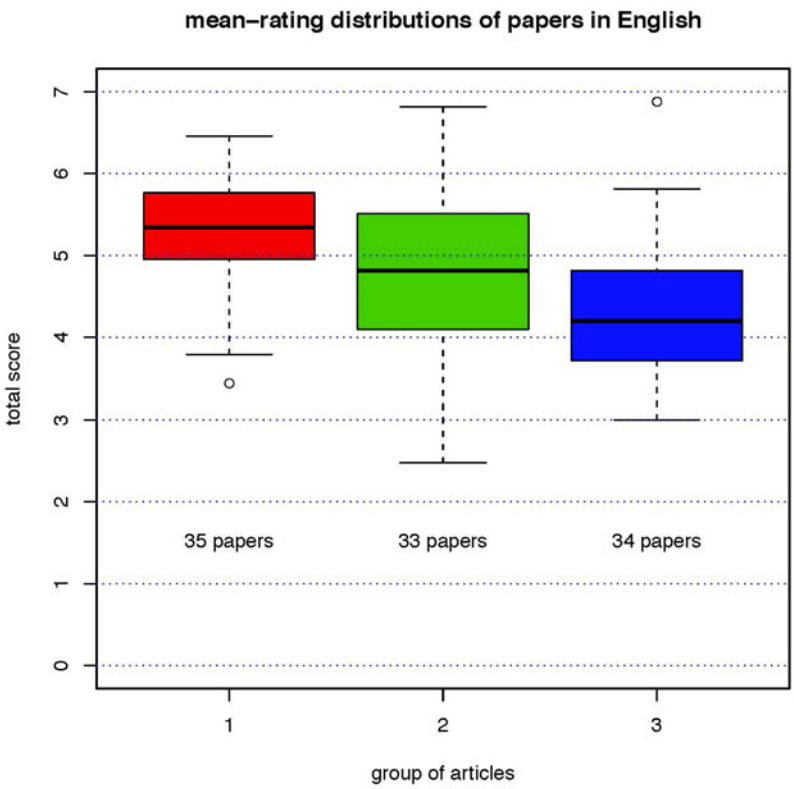
This clearly differs from the relatively high correlation strength in the interrelation of originality and significance as illustrated in the figure 4:

Figure 4: Originality - Significance Interrelation



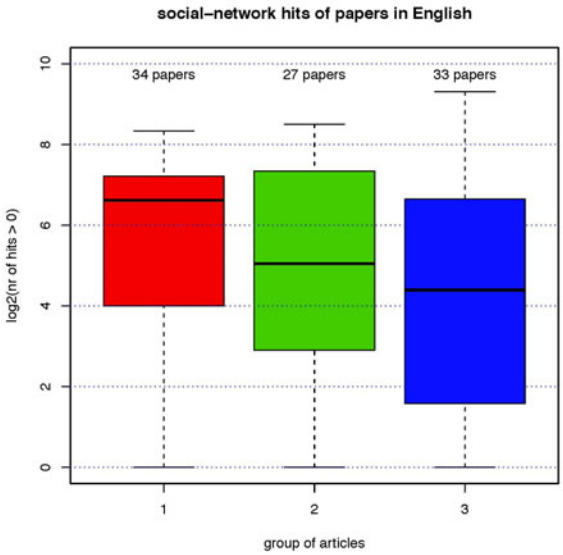
Regarding the relative ratings in the three groups of papers it is interesting to note that the first group is clearly rated best as can be seen in figure 5 below (the values for the German papers do not differ significantly):

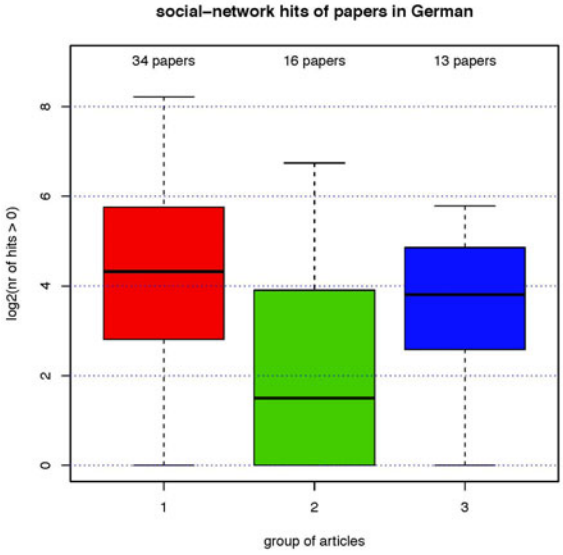
Figure 5: Box plots of Mean Rating Distributions of Papers in English



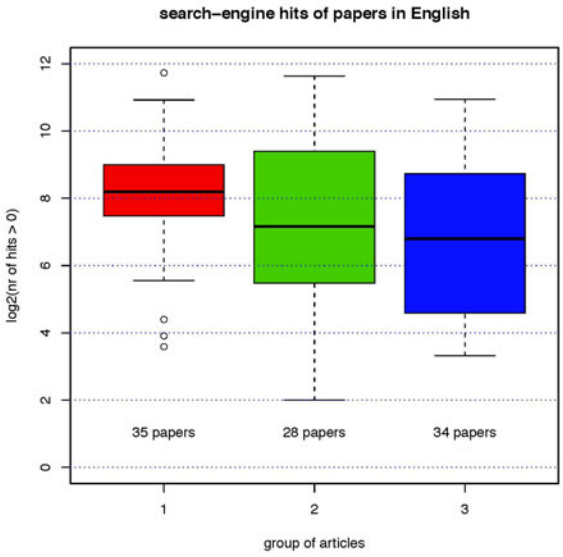
We then looked into the extrinsic paper data from search engines and Social-Network Services. These were extracted from the following sources: CiteULike, Library Thing, MendReader, Google and Metager. Many papers had only hits in one service. To get useful data we therefore applied the in-dubio-pro-reo rule and selected maximum values. We also assumed that zero hits cannot be used as a valid value of an indicator and thus excluded papers without hits from the analysis. Furthermore, the hit distribution of papers with at least one hit was heavily skewed to the left: Many papers had only a few hits and only a few papers had many hits. We therefore used the logarithm of hit numbers as a more adequate representation. We use dual logarithms for all box plot diagrams, i.e. the value of 8 on y-axis corresponds to 256 hits, a value of 10 to 1024 hits. The resulting diagrams show the following results:

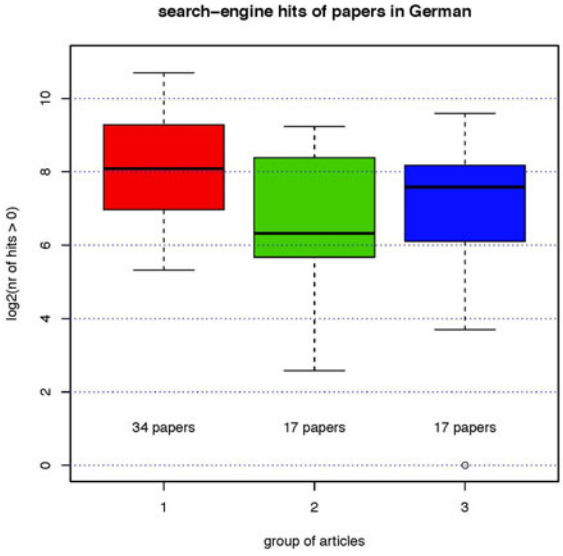
Figures 6 and 7: Social Network Hits of Papers in English and German





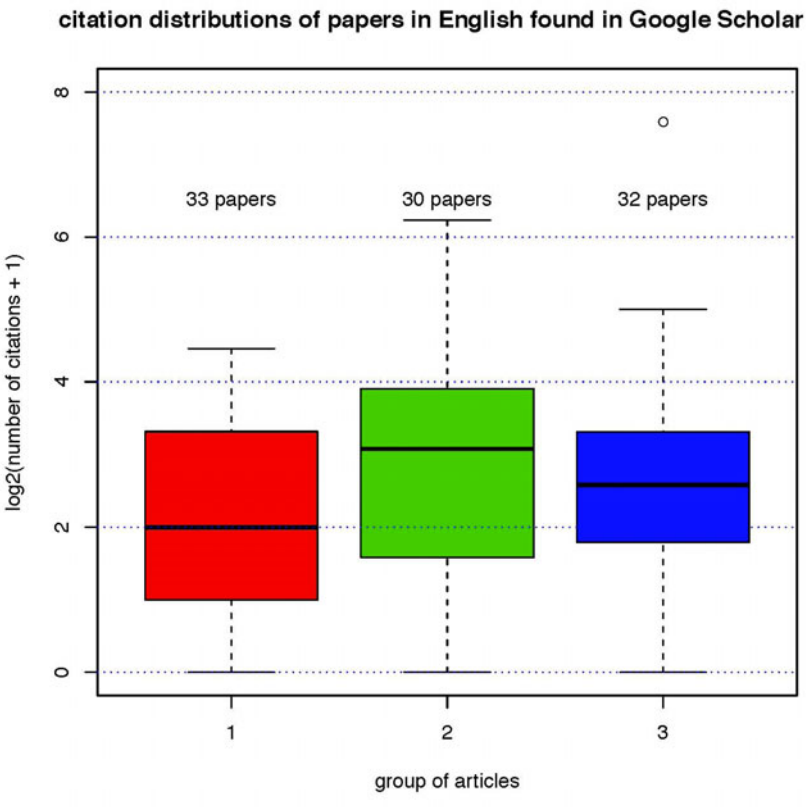
Figures 8 and 9: Search Engine Hits of Papers in English and German





One observation to be made is that all papers with social-network hits also have search engine hits and that both hit numbers correlate quite well in each of the three groups for papers in English, but less well for papers in German.
Finally, we looked into Citations in Google Scholar and analysed the citation distributions for samples of the three groups. Not all papers were listed in Google Scholar and only very few papers in German are in the sample: we decided to omit them. For the graphical representation we used the y-scale of dual logarithms of numbers of citation + 1. The addition of 1 is a usual bibliometric method to include papers without citations into the analysis of log-values. It can be justified with the argument that publishing a new result is its first citation (Figure 10).

Figure 10: citation distributions for samples of the three groups



This diagram is interesting in that the first (red) group receives the highest rating in the peer assessment using intrinsic indicators (see Figure 5), but receives the fewest citations (in contrast to the results for search engines and social-network services, where – at least for papers in English – peer ratings and numbers of search hits seem to correlate when aggregated into thematic groups).

3.6. Results Assessment

Based on the selected articles we found no significant correlations between the extrinsic indicators of research quality and the intrinsic ones - we even found evidence of non-correlation! A first test based on a non-parametric regression model to analyse the correlation between the different indicators had not been successful, either. The measurement model with three intrinsic and two extrinsic latent factors which was conducted by Prof. Ton Mooij at Radboud University, revealed a significant inter-correlation between the extrinsic respectively the intrinsic group of indicators. The results give evidence that the indicators are multi-collinear. However, no significant correlations were found between the intrinsic and the extrinsic factors that were selected for this test. In a second attempt, rank correlations and conducting factor analysis calculations based on 179 articles were carried out. In the third approach, a test of modelling the correlation between the indicators by using different regression models (non- parametric) was not successful either. This first attempts to identify correlations between extrinsic and intrinsic indicators were primarily based on the testing of uni-variate and linear correlations between the two sets of indicators. Correlations between the multivariate elements of each set are most probably non-linear and complex. (EERQI Project Final Report, 2011 p. 19)

In any case, we can conclude that the two sets of indicators are not correlating significantly but that they rather are complementary to each other. In other words, an article that has been judged as of high quality referring the indicator "rigour" may be well presented in online reference management services, even if it was not considered to be 'original'. Extrinsic and intrinsic indicators as defined in the EERQI project can clearly complement, but not possibly replace each other.

9 Limitations

Our study is based on a small sample of documents. All of these were traditional journal articles. Since monographs and anthologies are important media for publishing research in e.g. educational research, and that it is a type of publication which cannot be expected to disappear anytime soon (Wolfe Thompson, 2002) we are currently looking into ways to make monographs measurable. Since the lack of coverage of books in both the Web of Science[22] and in Scopus[23] is well known, we have decided to utilise another group of tools, analysing data from shared cataloguing services like "Library Thing". This, however, will be reported in a separate publication.

Another issue is, that even if the amount of data in the WWW allow us to get around the limitations of Web of Science' and Scopus' coverage, there is still the underlying problem of search engine reliability. Not only is there considerable variation between search engine retrieval performances, but the same search engine will also produce different results for the same search at different times and for different users. The situation is further aggravated by the fact that the coverage of Google is totally unknown up to know. That is why extreme caution is mandatory in using web-derived indicators for assessing research impact. Even more caution is advised if web based indicators are used in evaluation procedures.

The same, by the way, is true for the traditional sources of bibliometric information: the amount of fuzzy or inaccurate data, and the lack of standardisation of attributes, found in the course of our work is astonishing and makes us conclude that any figures derived from these sources needs to be used very cautiously, and any mechanistic trust in their reliability is likely to produce considerable harm!

10 Acknowledgements

Special thanks go to Prof. Ton Mooij for his contribution to the analysis of intrinsic and extrinsic data.

[22] The Book Citation Index by Thomson Reuters was launched in the end of 2011 - after the official end of the EERQI project.

[23] Even if in 2011 325 book series were part of the Scopus database one can hardly mention this as sufficient coverage - even more if this is the number for the book series of all disciplines covered by the database. http://www.info.sciverse.com/scopus/scopus-in-detail/facts

11 References

Bollen, J., 2010. The MESUR project: an overview and update. Available at: http://www.sparceurope.org/news/AAR_JB_MESUR_project_overview_update.pdf/view

Bornman, L., Mutz, R. & Daniel, H.-D., 2007. Gender differences in grant peer review: A meta-analysis. Journal of Infometrics , (1), pp.226-238.

Bornmann, L., 2008. Scientific peer review. An analysis of the peer review process from the perspective of sociology of science theories. Human Architture: Journal of the Sociology of Self-Knowledge, 6(2), pp.23-38.

Bourne, P. & Fink, J., 2008. I Am Not a Scientist, I Am a Number. *PLoS Comput Biol*, 4(12).

Bridges, D., 2009. Research quality assessment in education: impossible science, possible art? *British Educational Research Journal*, 35(4), pp.497-517.

Bridges, D. & Gogolin, I., 2011. The Process of Development of „Intrinsic Indicators". In *EERQI Final Conference, Brussels, 15th–16thMarch 2011*.

Burgelman, J.-C., Osimo, D. & Bogdanowicz, M., 2010. Science 2.0 (change will happen… .). *First Monday*, 15(7). Available at: http://www.uic.edu/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2961/2573.

Cicchetti, D., 1991. The reliability of peer review for manuscript and grant submission. *Behavioral and Brain Sciences*, 1(14), pp.119-186.

Cronin, B., 2001. Bibliometrics and beyond: some thoughts on web-based citation analysis. *Journal of Information Science*, 27(1), pp.1-7. Available at: http://jis.sagepub.com/content/27/1/1.full.pdf+html.

Duong, K., 2010. Rolling out Zotero across campus as a part of a science librarian's outreach efforts. *Science & Technology Libraries*, 29(4), pp.315-324.

EERQI Project Final Report, 2011 Hamburg. Available at: http://eerqi.eu/sites/default/files/Final_Report.pdf#page=9.

Egghe, L., 2006. Theory and practise of the g-index. *Scientometrics*, 69(1), pp.131–152. Available at: http://www.springerlink.com/content/4119257t25h0852w/fulltext.pdf.

Gilmour, R. & Cobus-Kuo, L., 2011. Reference Management Software: A comparative analysis of four products. *Issues in Science and Technology Librarianship*. Available at: http://www.istl.org/11-summer/refereed2.html#9 [Accessed January 23, 2012].

Hornbostel, S., 1991. "Drittmitteleinwerbungen. Ein Indikator für universitäre For-schungsleistungen?" *Beiträge zur Hochschulforschung*, (1), pp.57-84.

Hornbostel, S., 2001. “Third party funding of German universities. An indicator of re-search activity?” *Scientometrics*, 50(3), pp.523-53.

Hull, D., Pettifer, S.R. & Kell, D.B., 2008. Defrosting the digital library: bibliographic tools for the next generation web. J. McEntyre, ed. *PLoS computational biology*, 4(10), p.e1000204. Available at: http://dx.plos.org/10.1371/journal.pcbi.1000204.

Kolowich, S., 2010. New Measures of Scholarly Impact. *Inside Higher Ed*. Available at: http://www.insidehighered.com/news/2010/12/17/scholars_develop_new_metrics_for_journals_impact.

Kousha, K. & Thelwall, M., 2007. Google Scholar citations and Google Web/URL citations: A multi-discipline exploratory analysis. *Journal of the American Society for Information Science and Technology*, 58(7), pp.1055–1065. Available at: http://onlinelibrary.wiley.com/doi/10.1002/asi.20584/full.

Mead, T.L. & Berryman, D.R., 2010. Reference and PDF-manager software: complexities, support and workflow. *Medical Reference Services Quarterly*, 29(4), pp.388-393.

Moed, H., 2005. The future of research evaluation rests with an intelligent combination of advanced metrics and transparent peer review. *Science and Public Policy*, 34(8), pp.575-583.

Norris, M. & Oppenheim, C., 2003. Citation counts and the Research Assessment Exercise V: Archaeology and the 2001 RAE. *Journal of Documentation*, 59(6), pp.709-730.

Priem, J. & Hemminger, B., 2010. Scientometrics 2.0: New metrics of scholarly impact on the social Web. *First Monday*, 15(7). Available at: http://firstmonday.org/htbin/cgiwrap/bin/ojs/index.php/fm/article/view/2874/2570.

Qiu, J., 2008. Scientific publishing: Identity crisis. *Nature*, 451, pp.766-767. Available at: http://www.nature.com/news/2008/080213/full/451766a.html.

Ruths, D. & Al Zamal, F., 2010. A Method for the Automated, Reliable Retrieval of Publication-Citation Records. *PLoS One*, 5(8). Available at: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0012133.

Rynes, S.L. & Brown, K.G., 2011. Where Are We in the “Long March to Legitimacy?” Assessing Scholarship in Management Learning and Education . *Learning and Education*, 10(4), pp.1-55. Available at: http://www.aom.pace.edu/InPress/main.asp?action=preview&art_id=947&p_id=2&p_short=AMLE.

Smith, A. & Eysenck, M., 2002. The correlation between RAE ratings and citation counts in psychology. Available at: http://cogprints.org/2749/.

Stoye, D. & Sieber, J., 2010. *Description of aMeasure: Measuring extrinsic quality indicators in educational research publications EERQI report*, Berlin. Available at: http://edoc.hu-berlin.de/oa/reports/reJ3Xv4PJ82ZM/PDF/29B6vgnyGba6.pdf.

Thelwall, M., 2008. Bibliometrics to webometrics. *Journal of Information Science*, 34(4), pp.605-621.

Thelwall, M., 2003. Web use and peer interconnectivity metrics for academic web sites. *Journal of Information Science*, 29(1), pp.1-10.

The use of bibliometrics to measure research quality in UK higher education institutions, 2007 London. Available at: http://www.universitiesuk.ac.uk/Publications/Documents/bibliometrics.pdf.

Williamson, A., 2003. What will happen to peer review? Learned Publishing, 16(1), pp.15-20. Available at: http://www.ingentaselect.com/rpsv/cgi-bin/cgi?ini=xref&body=linker&reqdoi=10.1087/095315103320995041.

Wolfe Thompson, J., 2002. The Death of the Scholarly Monograph in the Humanities? Citation Patterns in Literary Scholarship. *Libri*, 52, pp.121–136. Available at: http://www.librijournal.org/pdf/2002-3pp121-136.pdf.

Wolinsky, H., 2008. What's in a name? *EMBO reports* , 9, pp.1171 - 117. Available at: http://www.nature.com/embor/journal/v9/n12/full/embor2008217.html.

Xuemei, L., Thelwall, M. & Giustini, D., 2011. Validating online reference managers for scholarly impact measurement. *Scientometrics*, 89(3), pp.1-11.

Zhang, C.-T., 2009. The e-Index, Complementing the h-Index for Excess Citations. *PLoS ONE*, 4(5). Available at: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0005429.

Citation patterns in educational research

Fredrik Åström

Short Summary
The purpose of this study is to investigate citation structures in educational research; and also, to study the visibility of European educational research. Bibliometric analyses are performed on data from Web of Science, the EERQI Content Base and Google Scholar, investigating both characteristics of publications through frequencies and distributions, as well as citation structures through co-citation analyses. The results show fragmented citation patterns presenting little opportunity to detect robust evidence of visibility or impact of contemporary educational research on any level of aggregation other than field level. This should be interpreted considering the diverse nature of educational research, and an organization of the field that differs from a strong norm, not the least in research evaluation programs, of research essentially being a cumulative process.

1 Introduction

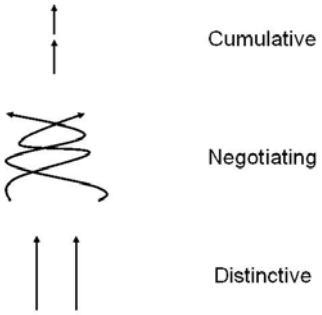
The main purpose of the EERQI project was to promote the visibility of European educational research by developing new indicators and methodologies for analyzing different aspects of quality in educational research publications (Gogolin 2012). In the project, a distinction between intrinsic indicators, or aspects of quality inherent to the document, and extrinsic indicators, i.e. indicators not being inherently linked to quality but being regarded as associated with quality through signs of use or impact, was made. One type of extrinsic indicator that has become increasingly important in the evaluation of journals, research institutions and scholars, is indicators of impact. The use of a document, as reflected in its inclusion in the reference lists of other documents, is assumed to indicate that the cited paper has an impact on the research community, i.e. the other authors citing the paper; and at least to some extent, the impact is assumed to reflect the quality of the document. The use of citation analysis to analyze impact or quality of research as reflected in publications is to a large extent based on assumptions of research essentially being a cumulative process; and also, that the process to a large extent is taking place within a defined research field and within a reasonably close timeframe. We cite the colleagues within our field because we build

I. Gogolin et al. (eds.), *Assessing Quality in European Educational Research,*
DOI 10.1007/978-3-658-05969-9_7,

our own research on the results, methods and theories of scholars who have previously studied the same phenomena.

However, this assumption can be contested. Both the extent to which research is organized and communicated cumulatively, and the extent to which we draw on research by active colleagues within our own field, varies substantially between different research fields. Whereas some fields indeed primarily organize research cumulatively, and using the references to show what previous research they are building their own efforts on, other fields – not the least in the social sciences and the humanities – use references to a large extent in order to debate or negotiate previous interpretations. We also find fields where there is little or no references to other researchers in the same field: the contribution to the field is shown through the uniqueness of the research done and the distinctive position of the publication (Figure 1) (Åström and Sándor 2009, Hellqvist 2010, Hammarfelt 2011).

Figure 1. Modes of scholarly communication (Åström and Sándor 2009, p 13).



Therefore, identifying communication and citation patterns in the specific research field is imperative to determine whether using citation based indicators makes any sense when trying to determine the impact of a document, a journal or an author. Thus, the aim of this paper is to analyze publication characteristics of educational research documents, with a particular focus on references and citations. The questions addressed are:

- What are the publication and citation structures in educational research in the *Web of Science* (WoS) databases?
- What publication and citation structures can be found in European educational research; and do they differ from the field taken as a whole?
- What can we say about the visibility of European educational research in the WoS databases and in *Google Scholar*?

To answer these questions, analyses were performed on five distinct sets of data. Two sets were retrieved using WoS, where one is based on 20 educational research journals indexed in WoS, and the other set is based on citations to articles in the EERQI Content Base from journals indexed in WoS. Two sets were collected using a combined search in the EERQI Content Base[24] and Google Scholar; and one set was gathered by gathering references from articles in journals in the EERQI Content Base.

2 Publication and Citation Structures in Educational research in the Web of Science databases

To get a general idea of citation structures in educational research, articles published over the years of 1998-2007 from 20 journals indexed in the WoS subject category *Education and Educational Research* (Table 1) were analyzed. Analyzing WoS data in a field where much of the research is published in sources outside the WoS databases is obviously something that can be debated, not the least in terms of to what extent the results of the analyses are representative for the field as a whole. It is, however, standard operating procedure for bibliometric mappings of research fields; and it will provide us with a baseline with which we can compare results of analyses on other educational research publications.

Table 1. Educational research journals selected for analysis

American Educational Research Journal	Instructional Science	Journal of the Learning Sciences
American Journal of Education	Journal of Computer Assisted Learning	Learning and Instruction
British Educational Research Journal	Journal of Education Policy	Reading Research Quarterly
Computers & Education	Journal of Educational and Behavioral Statistics	Review of Educational Research
Educational Evaluation and Policy Analysis	Journal of Experimental Education	Scientific Studies of Reading
Elementary School Journal	Journal of Higher Education	Sociology of Education
Harvard Educational Review	Journal of Research in Reading	

[24] The EERQI Content Base contains European educational research literature gathered both from publishers being part of the EERQI project and from open access sources online.

From these 20 journals, WoS data on 4,386 articles were downloaded for a set of analyses, including an author co-citation analysis (White and Griffith 1981), using the *Bibexcel* software (Persson et al 2009) and the Kamada-Kawai (1989) algorithm in *Pajek* (De Nooy et al 2011).

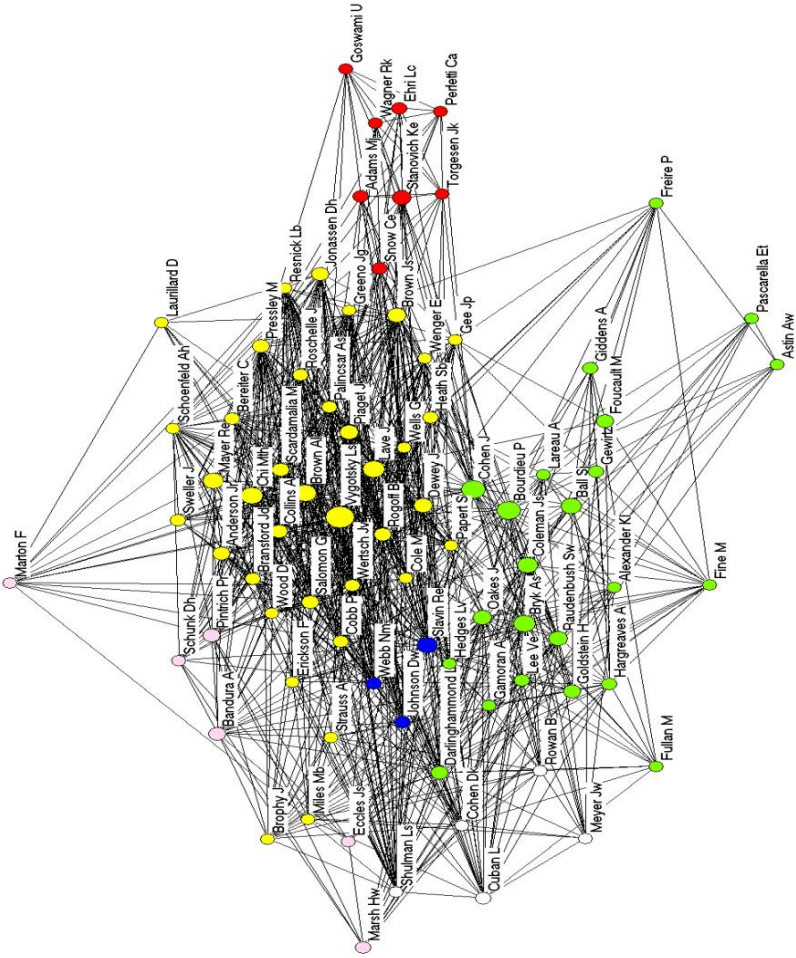
Some descriptive statistics reveal that 48% of the articles selected for analysis are written by authors with a US affiliation, whereas 39% of the authors are European. Of the European authors, 46% have an affiliation in the UK, thus the percentage of authors with an Anglo-American affiliation is 66%. These articles contain a total of 179,240 references to works by 52,648 authors, published over the years 1517-2008. Out of these, 90% of the references are published from 1980 and later; and when looking at the distribution of age of references from 1966 and onwards, 53% are published between the years 1966-1989. The tendency to cite relatively old references is also reflected in the *Citation Half-life* – a WoS measure for analyzing the life-span of articles in a journal by looking at the average number of years it takes for an article in a journal to receive 50% of its citations – for educational research journals in the WoS databases, which on average is 8.3 years. Apart from the large numbers of relatively old references, it is also worth noting that the share of references to journal articles is only 30%, reflecting how much of the scholarly communication in educational research takes place in, and is based on, literature in types of publications not indexed in the WoS databases For the co-citation analysis, the 82 authors with 75 citations or more were selected (Figure 2).

The map reveals little in terms of legible structures that could signify different research areas within the larger field of educational research, and we find a substantial amount of scholars from other fields than educational research among the highly cited authors such as Foucault, Giddens and Bourdieu; and at the same time, a large presence of citation classics in educational research and pedagogics such as Vygotsky, Piaget and Dewey. The lack of structure is partly depending on the presence of links between all authors in the analysis. It should however be noted that the strength of these links is relatively low. When comparing this analysis with an analysis on a field like library and information science (LIS) (Åström 2007) we find that, despite similarities in the distribution of citations per author, the number of co-citation links in educational research is lower both in absolute numbers as well as in terms of the distribution of co-citation links for the most co-cited authors.

Another feature when comparing the results of this analysis to LIS research, which is a field sharing many features with educational research, such as the multidisciplinary nature and the close connection to the field of professional practice, is the relatively low amount of citations within the document set. *Citation Among Documents*-analysis was suggested by Persson (2001) to avoid topic

drift by excluding references that was not also included among the citing documents. When analyzing LIS, the number of citations among documents is 8%; and the results yield a fair representation of the structure of the field as a whole. When doing the same kind of analysis on educational research, the number of citations among documents is only 2% of the total amount of citations; and these seem primarily be limited to educational research oriented towards computer assisted learning and other types of computer oriented educational research.

Figure 2. Author co-citation analysis based on educational research journal articles in WoS, 1998-2007.



Furthermore, when looking at the distribution of citations per author, it is a less skewed distribution than we would normally assume in bibliometric analyses. Typically, almost all distributions we find in bibliometrics follow a Pareto distribution (Bradford 1934, Fairthorne 1969, Merton 1968) with an 80/20 relationship. In the case of educational research as reflected in WoS, 20% of the most cited authors only accumulate 65% of the citations. The distribution is still skewed, but still substantially less so than we would expect.

3 Citation Structures in European Educational Research

To investigate to what extent the analysis of citation patterns in educational research as reflected by the WoS databases is representative for the field as a whole, a similar analysis was performed based on a set of articles from journals in the EERQI Content Base. The references from 375 articles published in nine journals during the year 2008 were collected (Table 2). The limited number of articles from which the references were collected makes the analysis more tentative than the WoS analysis, but was necessitated by the manual collection of data. Another consequence of this is limitations in terms of descriptive statistics on the journal articles providing the references for the co-citation analysis.

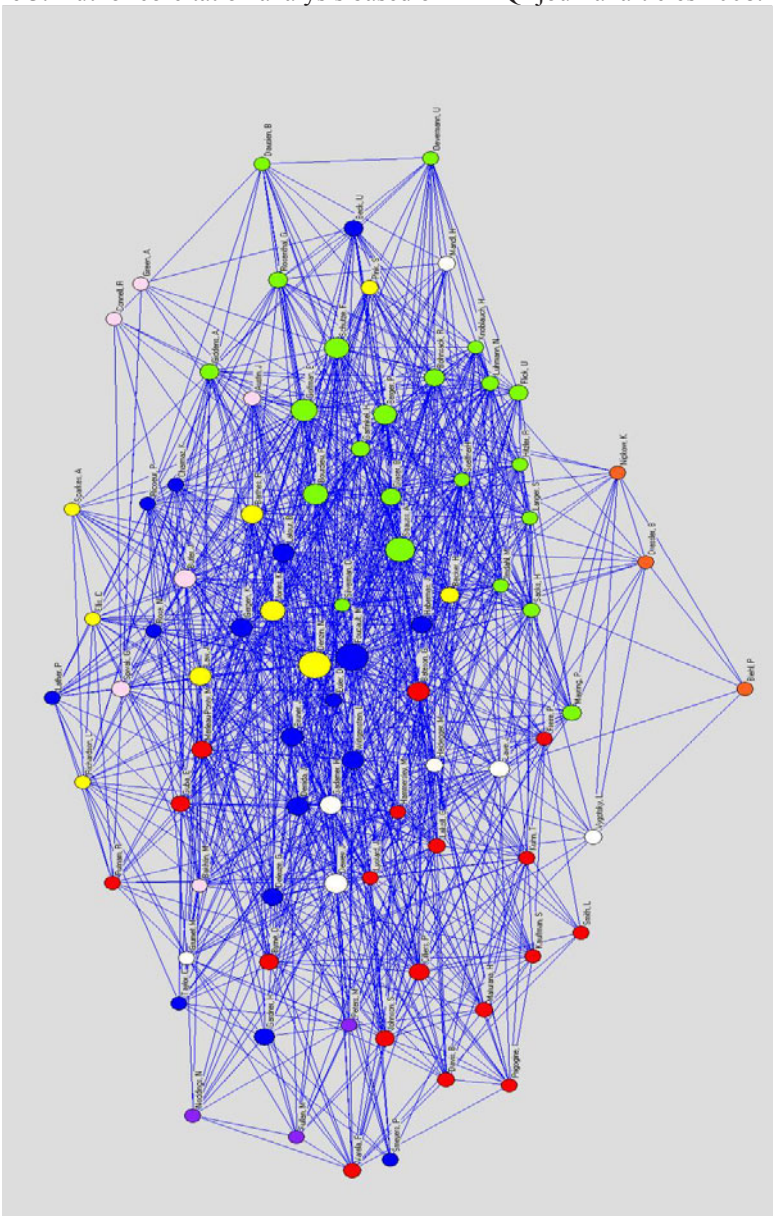
Table 2. European Educational research journals selected for analysis

Journal	Articles	References
Assessment in Education: Principles, Policy & Practice	23	709
Compare: A Journal of Comparative Education	45	1,487
Educational Measurement: Issues and Practice	17	514
Educational Philosophy and Theory	68	2,182
European Journal of Education	10	325
Contemporary Issues in Early Childhood	27	900
Forum : Qualitative Social Research	122	4,858
Berufs- und Wirtschaftspädagogik Online	26	898
Theo-Web	37	1,184
Total	**375**	**13,057**

From the 375 journal articles, 13,057 references by 7,150 authors were collected for an author co-citation analysis (Figure 3). As with the WoS co-citation analysis, the co-citation links are relatively few in relation to citation frequencies among the individual authors, and we find a map with few legible structures signifying different research areas with educational research. There are also similarities in-between the WoS and EERQI maps when looking at the highly

cited authors: in both cases, there is a dominance of on one hand, general theorists from the humanities and social sciences, and on the other, citation classics in educational research and pedagogics. And when looking at the distribution of citations per authors, it is even less skewed than in the WoS analysis. In the WoS dataset, the share of citations for the 20% most cited authors was 65%, in the case of the EERQI dataset the share is 50%.

Figure 3. Author co-citation analysis based on EERQI journal articles 2008.



Since the share of articles and references per journal varies substantially, the same analysis was also done after normalizing the data by adjusting the frequency of citations in relation to the total number of references per journal. This results in a number of identifiable clusters in another way than the raw frequencies used in the map in Figure 3. However, these clusters primarily relate to what journals the citations come from, not necessarily to research areas or specialties within educational research. Not the least since the data set includes both English and German journals; the conclusion to be made is more related to a heterogeneous citation structure related to different journals and national differences, rather than different research areas.

In conclusion of the co-citation analyses, when comparing the analyses of the WoS and EERQI datasets, there are many similarities. The highly cited authors included in the analysis receive a smaller share of the total amount of citation than expected, the number of co-citation links between the authors is relatively low and when visualizing the results of the co-citation analysis, we find little in terms of legible structures signifying different research areas. Of the authors included in the co-citation analyses, circa 20% are present in both sets, but in terms of what kind of authors being cited in educational research in WoS and EERQI respectively, there are substantial similarities, with both sets being dominated by general theorists from the social sciences and the humanities, and by citation classics in educational research and pedagogics. Thus, we have a sparsely populated 'citation universe', where few authors receive any larger amounts of citations and a following lack of identifiable structures when analyzing citations on an aggregated level.

4 The Visibility of European Educational Research in the Web of Science Databases

The sparsely populated citation universe we find when analyzing structures on an aggregated level promises little for using citation analysis to analyze impact and visibility of research in the educational sciences. Few European educational research journals are indexed in the WoS databases; and as we see from the descriptive statistics on author addresses in WoS journal articles, the dominance of Anglo-American educational research in the WoS databases is substantial. Of the journals included in the EERQI Content Base, none are indexed in WoS. However, WoS is not entirely limited to the journals being indexed in the databases. By performing a *Cited Reference Search*, it is possible to find 'non-source items', i.e. documents that are not indexed in the databases but still appear in the citation indexes since they are cited in journals indexed in WoS (Butler and

Visser 2006). This gives us the opportunity to examine what impact the 'EERQI journals' have had as reflected in the WoS, or at least to test whether the EERQI journals are at all visible in WoS.

There are problems with using the Cited Reference Search option in WoS. When searching for cited journals, the full title cannot be used, only the abbreviated title. Abbreviated titles for journals indexed in the WoS databases can easily be retrieved in Journal Citation Reports (JCR), although some articles will not be found since the standardization of cited journal data is not perfect. When looking for cited journals not indexed in WoS and present in JCR, another way of finding the preferred abbreviation is needed. By doing a search on cited author from a set of a 100 articles in the EERQI Content Base, nine EERQI journals receiving citations from WoS indexed articles could be found (Table 3).

Table 3. Journals from the EERQI Content Base cited in WoS journals (data collected in September 2010)

'EERQI' Journals	**Cited Articles***	**Citing Articles****
Contemporary Issues	276	475
E Learning	107	163
Erziehungswissenscha	480	473
European Ed Res J	120	180
Forum QualitativeSo	459	553
Policy Futures Ed	97	132
Res Comp Int Ed	18	19
Rev Francaise Pedago	366	400
Z Didaktik Naturwiss	49	67
Total	**1,972**	**2,444*** **

* Number of articles from EERQI Content Base journals cited in WoS journals
** Number of WoS journal articles citing EERQI Content Base journals
***The total amount of citing articles differs from the sum of citing articles per journal since articles from more than one EERQI journal have been cited by the same article.

On average, few articles have received more than one citation. The average number when calculating number of citing articles per cited article is 1.01; and the variations in-between journals is also low, ranging from 1-1.5. This confirms the sparsely populated citation universe identified in the co-citation analyses, not the least in terms of 'intra-educational research citations'. However, the 2,444 articles citing EERQI journals give us an opportunity to at least look into some macro-level structures, i.e. the citations to educational research as represented by these nine journals.

The impact of EERQI journals was investigated by analyzing three different aspects: in what journals we find articles citing EERQI journal articles, from

which countries citing articles comes from and what WoS subject categories are used to describe the journals in which the citing articles are published. In the first analysis, the article frequency per journal is investigated (Table 4).

Table 4. WoS-indexed journals citing EERQI journal articles (10 articles or more).

Article freq.	Journal name	Article freq.	Journal name
251	Zeitschrift für Pädagogik	16	Oxford Review of Education
91	Pädagogische Rundschau	15	Culture & Psychology
39	Revue Française de Sociologie	15	Higher Education
32	European Journal of Psychology of Education	13	British Educational Research Journal
32	Historische Sozialforschung	13	Perception & Psychophysics
31	International Journal of Science Education	12	Paedagogische Rundschau
28	Journal of Education Policy	11	Teaching and Teacher Education
21	Comparative Education	11	Journal of Computer Assisted Learning
21	Zeitschrift fur Erziehungswissenschaft	10	Kölner Zeitschrift für Soziologie und Sozialpsychologie
20	Année Psychologique	10	Computers & Education
17	British Journal of Sociology of Education	10	Comparative Education Review

Of the 22 journals publishing more than ten articles citing EERQI journals, 11 of them can be identified as European by the title alone; and when looking at the share of articles, 75% of those 719 articles included in the table are published in European journals. Considering this, it is interesting to analyze the national origin of the articles by investigating the author addresses for the articles citing EERQI journals (Table 5).

Table 5. Country of article origin through author affiliation for articles citing EERQI journals (fractionalized counts). 10 articles or more.

Freq	Country	Freq	Country	Freq	Country
379,334	USA	30,486	Netherlands	14,565	Finland
331,185	Germany	27,332	Belgium	14,416	Italy
303,73	UK	25,698	Sweden	12,5	NewZealand
183,728	France	23,5	SouthAfrica	11,499	Brazil
97,995	Canada	19,583	China	11,229	Ireland

81,645	Australia	18,498	Austria	11,166	Denmark
41,889	Switzerland	18,166	Israel	10,4	India
38,198	Spain	14,666	Norway	10	Mexico

Given the large number of articles being published in European journals, it might be surprising that American affiliated authors contribute with the largest share of articles citing EERQI journals on a country by country basis. However, when analyzing the author affiliation on continent level, the North American contribution is 26% whereas the share of European authors is 61%, which in terms of visibility means that 39% of the articles citing EERQI journals are by authors of a non-European affiliation. To some extent, this reflects the Anglo-American dominance in the WoS databases, but it also illustrates the problem of how to classify the origin of research publications, especially on journal level: even though a journal may be published in Europe, it will still publish articles by non-European authors and vice versa (Danell 2001).

When analyzing the WoS subject categories of journals with articles citing EERQI journals, it is not surprising to find the categories 'Education & Educational Research', 'Education, Subject spec.' and 'Psychology, Educational' as top ranked, with 1,084 occurrences or a share of 44%. More interesting to note, however, might be the frequencies of subject categories that are not nominally related to educational research, to get an idea of the use of EERQI journal articles outside the educational sciences (Table 6).

Table 6. Subject categories for journals citing EERQI articles. The analysis only includes frequencies of subject categories not jointly used with 'Educational Research', 'Education, subject spec.' or 'Psychology, Educational'. Fractionalized counts, frequencies of 20 or more.

Freq	Subject Category	Freq	Subject Category	Freq	Subject Category
120,247	Psychology, Multidisc	47,326	Lang & Linguistics	25,914	Psychology, Applied
118,832	Sociology	39,497	Psychology, Clinical	24,964	Publ Env Occ Health
62,878	Computer Science	32,333	Nursing	24,617	Economics
53,863	Psychology, Experim	31,033	Management	22,498	Psychiatry
53,448	Social Sci, Interdisc	30,997	Psychology, Devel	22,163	Communication
53,195	Psychology	29,165	Political Science	21,665	Information & Libr Sci
49,915	Psychology, Social	28	History	20,5	Indust Rel & Labor

There is a wide range of psychology oriented subject categories reaching high frequencies, which is not surprising considering the close ties between educational research and pedagogics on one hand, and psychology, especially developmental, on the other. And taking into account the strong presence of computer oriented educational research in the citations among documents analysis when doing the WoS co-citation analysis, the strong presence of computer science journals citing EERQI journals is also quite understandable. In general, it is striking to notice the wide range of different subject categories citing educational research journals, from sociology and management, over nursing and library and information science, to medicine and engineering; the last two just outside the subject categories with frequencies higher than 20.

5 The Visibility of European Educational Research in Google Scholar

Analyzing the visibility of EERQI journals in the WoS databases still leaves questions regarding what the results actually represent: the field of educational research per se or the (limited) presence of educational research in the WoS databases. Looking at the co-citation analyses, there seem to be structural similarities but at the same time, substantial differences in terms of journals and au-

thors cited, as well as e.g. variations depending on national origin of research. To get a sense of the visibility of EERQI journal articles outside the WoS, analyses were made on documents citing EERQI journals in Google Scholar. Due to the level of manual work involved in collecting data, there are differences in the kind of analyses undertaken; and it was also necessary to collect two different data sets for different analyses.

To get a sense of the life span of EERQI documents, the EERQI Content Base was searched for documents published in the years 2000-2003 (Table 7). To make the analysis more robust, a larger data set over a longer period of time would have been desirable. However, very few documents published before 2000 could be retrieved and considering the average 8 year citation half life of articles in educational research in WoS, the analysis was limited to the year 2000-2003. These documents were searched in Google Scholar to retrieve data on the number of citations to the EERQI documents, as well as the publication year of the citing documents.

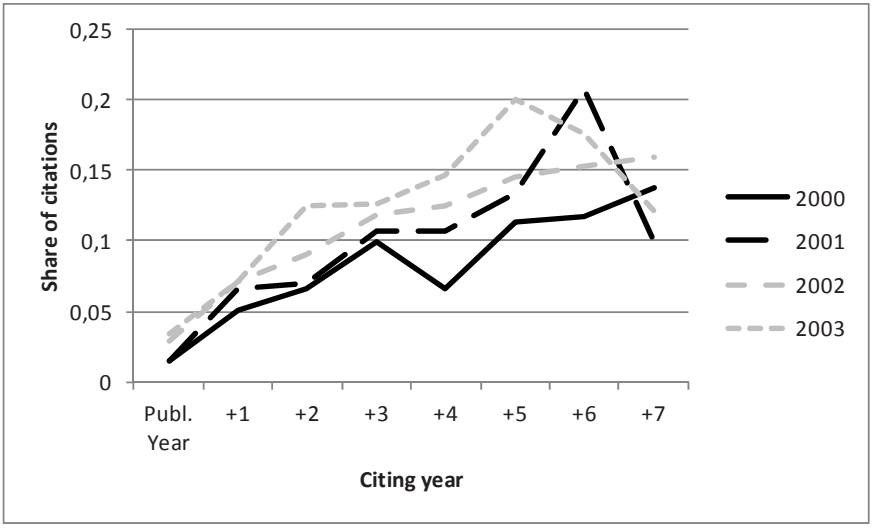
Table 7. EERQI Content Base research articles published 2000-2003 and their citations in Google Scholar (data collected in October 2010).

Publication Year	No. Publications*	No. Citations
2000	47	334
2001	78	485
2002	120	600
2003	165	927

* Number of publications retrievable by searching on publication year; and, only including publications receiving citations in Google Scholar.

The average number of citations per paper is relatively stable, varying between 5-7 citations per paper. To investigate if it is possible to get a sense of how long the EERQI documents were cited, the share of citations per year was calculated on the basis of publication year for the EERQI documents (Figure 4).

Figure 4. The share of citations per year for EERQI documents published 2000-2003



With a few exceptions, there is an increase of citations for each of the seven years following the year of publication, which is in accordance with the eight year citation half life in WoS. There are two exceptions, where we see a decline in the share of citations. For publications from 2001, the decline in the seventh year after publication is to a large extent depending on outliers, where a few articles on the topic of student participation in school reform received an inproportionate number of citations in 2007, causing a spike and a following decline in the relative share of citations. In the case of EERQI documents published in 2003, the decline in the last year analyzed may be explained by the analysis being performed at the end of 2010, thus all citations from that year might not yet have been indexed in Google Scholar.

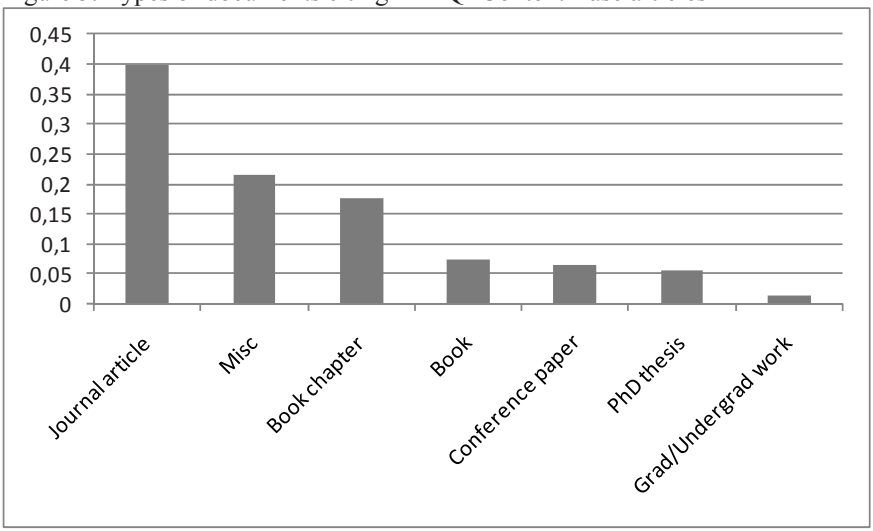
The second set of data from Google Scholar is based on 100 EERQI Content Base documents used in a test of the intrinsic and extrinsic indicators developed within the EERQI project (data collected in October 2010). From these documents, information on author, title and journal was gathered. These documents were searched in Google Scholar and for the documents citing EERQI publications; information was gathered on author, title and source, type of document and language. This made it possible to analyze the number of citations, including controlling for self-citations, in what kind of documents EERQI articles are cited and how the citations are distributed over languages. In total, the

100 articles have received 347 citations, i.e. an average of 3.47 citations per paper (CPP). Since the 100 cited documents are split into 25 each of German and French articles and 50 English articles, separate analyses for each language was also conducted. In terms of general citation patterns, French articles stands out with a low CPP of 1.48 and the most highly cited article attracting only ten citations, whereas the German and English equivalents received around 40 citations each.

The distribution of citations to papers was also analyzed. Generally, this distribution tends to be skewed, with around 20% of the authors in a field receiving 80% of the citations (Merton, 1968). The distribution of citations to the EERQI material is roughly the same, with a 28/80 distribution of percentages for cited authors and citations. But there are some interesting variations to be noted when comparing articles in different languages: for the English language articles, 37% of the cited authors are needed to reach 80% of the citations, whereas for the French articles, the numbers are 20/84. And when looking at the number of articles without any citations, the figures are 32% for the whole material, whereas it ranges from 10% for the English to 64% for the French. By matching already available information on author names for the cited documents and retrieved information on author names for citing documents, it was possible to investigate the level of self-citations; here defined as citation links between documents where at least one author is the same in both the citing and cited document. The variations between languages were very small, and the share of identified self-citations is 20% of the total amount of citations.

An important issue when using citation data from Google Scholar is the lack of control over where the citations are coming from. To investigate this, the citing documents were classified according to document type (Figure 5).

Figure 5. Types of documents citing EERQI Content Base articles



Most citations to the EERQI articles comes from journal articles. The 'Misc' category – being the second most frequently occurring document type – contains documents such as project reports, internal discussion or seminar papers being available online and in one instance, papers/articles in publication archives/repositories, with no identification of source of publication other than the archive itself, and a few powerpoint presentations where the forum for the presentation could not be identified. Also here, we can identify some differences between citing publications in different languages: whereas almost 50% of the citations to English language articles comes from other journal articles, the figures for the French and German articles ranges between 22-27%. Citations from book chapters make up 32% of the German citations, while the figures for citations to English articles in book chapters are 14% and for the French only 3%. More than 50% of the citations to the French articles come from the Misc category, primarily publications in archives such as HAL (http://hal.archives-ouvertes.fr), while 'Misc' citations to English and German articles make up 17-20%.

The analysis of language of the citing documents in relation to the language of the cited document show a large dominance of citations coming from publications in the same language as the cited document (Table 8).

Table 8. Language of documents citing EERQI journal articles

	Language, citing publication			
EERQI publ.	**English**	**French**	**German**	**Other**
English (N=210)	93%	0%	3%	4%
French (N=37)	0%	92%	0%	8%
German (N=100)	10%	0%	90%	0%
All (N=347)	59%	10%	28%	3%

Following the strong tendency towards citations within the language groups, the distribution of citations to the whole set of EERQI documents show a relatively low share of English citations, which is typically held as the international language of academic research. This tendency can also be related to the results of a normalized co-citation analysis of the EERQI journal articles, where the structure of references to a large extent was related to the individual publications and their origin. This can be seen as a reflection of educational research to a large extent being organized on a national level, typical for fields of research in the humanities and the social sciences (Whitley 2000).

6 Conclusions

When looking at the results taken together, there are some patterns that become apparent. One of the more important issues arising is that a large share of the highly cited authors are either general theorists in the social sciences and humanities or citation classics in educational research and pedagogics, that even highly cited authors in the educational sciences attract relatively few citations, and that the links between the highly cited authors are weak. Regardless if we look at co-citations in WoS educational research journals, citations to EERQI journals in Google Scholar or references in EERQI Content Base journals; the citation frequency is low, the links between authors or documents are few, and the distribution of citations is not so much skewed in the way we typically find in bibliometric research but rather to be regarded as scattered in a sparsely populated citation universe. This is further emphasized by findings suggesting that even educational research journals with a general orientation sharing little in terms of similar citation patterns and the flow of citations between publications in different languages is very limited – even in the case of non-English publications citing English research. The few exceptions where we find citation patterns similar to what we would normally expect in bibliometrics – adhering to what is typically seen as the norm for scholarly communication patterns, and the basic assumption for most models for using citation analysis as an indicator on visibility or

impact of published research – are found in educational research with strong ties to e.g. the computer sciences and quantitative research. In terms of visibility of educational research publications, this means we have to reach very high levels of aggregation before we start finding any robust signs of impact. Not even on journal level do we find substantial evidence of impact unless we operate on large time spans, and even then, it only works on a very limited number of journals.

It should be kept in mind that many of the analyses performed here are based on small datasets; and that the analyses therefore should be regarded as tentative. However, in terms of citation patterns on a structural level, the different analyses show considerable similarities; and this is regardless of whether it is a matter of an analysis of more than 3,000 documents in the WoS databases or one on a 100 documents in the EERQI Content Base.

The results are hard to interpret in relation to generally held assumptions related to citation impact analyses, scholarly communication and the organization of research in general. We find little evidence of a cumulatively organized organization of research, with citations to contemporary peers within our field. Instead, the results need to be seen in the light of a scholarly communication structure where much is a matter of negotiating different interpretations from various theoretical viewpoints, developed in different fields in the social sciences and the humanities. Furthermore, the multidisciplinary nature of educational research, with research spanning from philosophy of education to computer assisted learning, also contributes to a fragmented appearance when analyzing the citation patterns of the field. In addition to this, we also have a field with substantial national differences in terms of research traditions, as well as the educational systems that form the basic empirical area of research for the field.

In conclusion, the citation analyses of educational research reflect a field where it is very hard to use citation analysis to detect evidence of impact or visibility of research in the field, at least with any ambitions towards contemporary analyses. Instead, the analyses show a field of great diversity both in terms of research areas and specialties within the field, and in national orientation and organization of research.

7 References

Åström, Fredrik (2007): Changes in the LIS research front. Time-sliced cocitation analyses of LIS journal articles, 1990–2004. Journal of the American Society for Information Science and Technology, 58(7), 947-957.

Åström, Fredrik and Sándor, Ágnes (2009): Models of Scholarly Communication and Citation Analysis. In: B. Larsen, Birger and J. Leta, Jaqueline (Eds.): ISSI 2009: The 12th International Conference of the International Society for Scientometrics and Informetrics. Rio de Janeiro: BIREME/PAHO/WHO & Federal University of Rio de Janeiro, 10-21.

Bradford, Samuel C. (1934): Sources of Information on Specific Subjects. Engineering, 137(Jan 26), 85-86.

Butler, Linda and Visser, Martijn S. (2006): Extending citation analysis to non-source items. Scientometrics, 66(2), 327-343.

Danell, Rickard (2001). Internationalization and homogenization. A bibliometric study of international management research. Umeå: University. Diss.

De Nooy, Wouter, Mrvar, Andrej and Batagelj, Vladimir (2011): Exploratory Social Network Analysis with Pajek. Cambridge: University Press.

Fairthorne, Robert A. (1969): Empirical Hyperbolic Distributions (Bradford-Zipf-Mandelbrot) for Bibliometric Description and Prediction. Journal of Documentation, 25(4), 319-343.

Gogolin, Ingrid (2012): Introduction. In: Gogolin, Ingrid and Åström, Fredrik (Eds.): Assessing quality in European education research. Indicators and approaches. Wiesbaden: VS Verlag, xx-yy.

Hammarfelt, Björn (2011): Interdisciplinarity and the intellectual base of literature studies. Citation analysis of highly cited monographs. Scientometrics, 86(3), 705-725.

Hellqvist, Björn (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.

Kamada, Tomihisa and Kawai, Satoru (1989): An algorithm for drawing general undirected graphs. Information Processing Letters, 31, 7-15.

Merton, Robert K. (1968): The Matthew effect in Science. Science, 159(3810), 56-63.

Persson, Olle (2001): All author co-citations versus first author co-citations. Scientometrics, 50(2), 339–344.

Persson, Olle, Danell, Rickard and Schneider, Jesper W. (2009): How to use a Bibexcel for various types of bibliometric analysis. In: Åström, Fredrik et al (Eds.): Celebrating Scholarly Communication Studies. A Festschrift for Olle Persson at his 60th Birthday. International Society for Scientometrics and Informetrics. Available at: http://www.issi-society.info/ollepersson60/ollepersson60.pdf

White, Howard D. and Griffith, Belver C. (1981): Author co-citation. A literature measure of intellectual structure. Journal of the American Society for Information Science, 32(3), 163-171.

Whitley, Richard (2000): The intellectual and social organization of the sciences. Oxford: University Press.

The EERQI Peer Review Questionnaire – From the development of 'intrinsic indicators' to a tested instrument

Ingrid Gogolin, Verena Stumm

Summary
All areas of research are increasingly confronted with demands for 'accountability', for the implementation of performance measures or other means of ensuring 'value for money'. In particular for research funding, a researcher's demonstration of 'high quality', if not 'excellence' - of his or her individual achievements and working environment - are preconditions *sine qua non*. The European Educational Research Quality Indicators (EERQI) project was motivated by this development and the assumption that it may cause undesired side effects – not least, because many methods and instruments which are applied in order to detect 'quality' seem to lack in quality themselves. Within the framework of the EERQI-project, this assumption was examined from different perspectives. Our contribution presents one of the EERQI-approaches, namely the attempt to develop and evaluate an instrument for peer review purposes. We describe the process from the first attempts to develop a set of criteria, which most likely refer to the quality of an educational research publication, to the final evaluation of a peer review questionnaire with criteria which are widely accepted and shared in the educational research community.

1 Introduction

The European Educational Research Quality Indicators (EERQI) project was motivated by international notions of scientific quality and the fact that funding allocations based on these notions may cause undesired side effects. This can be the case when questions of how quality is interpreted and measured are left open. Current instruments for measuring quality (such as citation counts) can cause such side effects, despite considerable improvements in recent years. Criticisms include the language biases that are present in such instruments. For instance, knowledge in the Social Sciences and Humanities (SSH) is produced in many languages other than English and a greater diversity of document types is produced. In their report on creating a comprehensive database of research outputs in the SSH, Moed et al. (2009, p. 5) state: "Journal articles are only a minor part

I. Gogolin et al. (eds.), *Assessing Quality in European Educational Research,*
DOI 10.1007/978-3-658-05969-9_8, © Springer Fachmedien Wiesbaden 2014

of research output. A substantial part is communicated through books, especially in the humanities. There is less concentration in a limited number of international-scale journals. Much more often than in science, national or regional journals are important." On the same topic, Hicks and Wang (2009, p.2) state: "In the social sciences, humanities or arts it is largely impossible to substantiate statements on research excellence with reliable indicators for international benchmarking of fields and institutions."

The EERQI project was based on analyses of this field. It aimed at developing and testing alternative approaches to detect the quality of research publications, with educational research serving as a case model for the SSH. To be clear, the EERQI approach was not to develop one particular instrument, such as a European Citation Index. This is because the research team did not assume that the risks and shortcomings of the existing methods could be overcome by a single, small-scale research project.[25] Rather, it was EERQI's aim to develop and test intelligent combinations of different methods that complement each other in the assessment process (this included bibliometric methods where appropriate). The 'EERQI Prototype Framework' thus refers to this set of tools that can be used and combined in one way or another to detect different facets of the quality of a research publication. The general assumption of the EERQI project, however, is that final decisions have to be made by a well-informed reader – a 'connoisseur' of the respective area. The tools developed and tested by EERQI may serve the assessment process and assist the reader in that process; but, in the end, they cannot replace him or her.

Part of the EERQI process was to develop, test and improve a set of text-immanent criteria that assist the evaluation of educational research publications. These criteria should function in the sense of signals that signpost expounded aspects of educational research quality – or in other words: as 'indicators' in the sense of social sciences. In an iterative process, the project established a set of generic criteria that can be applied for the determination of quality in peer-review processes. This set has been presented to the (not only educational) research community on several occasions. It was thoroughly tested and revised, positively evaluated and subsequently transferred to what we called the 'EERQI Peer Review Questionnaire'. This instrument was then again tested for reliability and practicality. In the following chapter, we present the process of development and testing of this instrument and the final version of the questionnaire which can now be utilised by the scientific community.

[25] Other initiatives, such as the project "Towards a Bibliometric Database for the Social Sciences and Humanities – A European Scoping Project", actually strive for such developments. See for example Meester 2013.

2 'Intrinsic indicators' – from the idea to the test

The EERQI project began by identifying a state-of-the-art report on existing methods and indicators (Botte & Vorndran, 2008). The report included an overview of widely used methods for quality detection and techniques, some of which are still in development (such as online usage metrics or new retrieval and clustering approaches). Areas to be explored by the EERQI project, such as the role of semantic text analysis, were then considered. On the basis of these preparations, the first EERQI workshop was organized in 2008 as an international event in Leiden, Belgium under the auspices of the "European Association for Research on Learning and Instruction (EARLI)". This assembly of educational researchers from all over Europe decided that the EERQI project had to deal with both 'extrinsic' and 'intrinsic' quality indicators which deliver different but complementary information. Extrinsic indicators – such as the formal features of a text and its producers, the medium of publication, citations – concentrate on the outward appearance of a text and its potential impact which is detected on these grounds. These features can be considered as signals of quality because they indicate the accessibility of a text and its potential contribution to a scientific discourse. Intrinsic indicators, on the other hand, relate less to the context but to the *content* of a publication. Such indicators thus have to be identified within the textual performance of a publication itself (for further clarification see Botte and Mooj in this volume). Whether extrinsic measures do actually carry signs of quality, was subject to debate throughout the EERQI project (see Bridges in this volume and Bridges, 2009). Another source of controversy derived from the different 'national' research traditions which were assembled in the EERQI project. 'Good quality' in educational research publications may refer to different parameters in different national contexts and their research traditions; therefore 'terms' that indicate quality might not be appropriate across borders (Rey, 2006).

The 'EERQI Peer Review Exercise' was established against the backdrop of such controversies. The first aim of the exercise was to test the sheer possibility of agreement among an international, sub-disciplinary complex educational research community on a core set of terms representing 'good' (or 'bad') quality research publications. The second aim of the exercise related to transferring this potential set of terms to a useful tool to be used by researchers. Based on the assumption that it would indeed be possible to identify a set of core terms indicating quality, the EERQI team and consulted experts concluded that a peer review questionnaire that had been tested for reliability and validity would perfectly serve the purpose to denote unambiguous as well as approved terms and their operationalisation.

The 'EERQI Peer Review Exercise' was carried out in an iterative process of consultation, further development and feedback. Firstly, educational researchers and other experts in the field of quality assessment were invited via the relevant research associations (for example, European Association for Research on Learning and Instruction (EARLI); European Educational Research Association (EERA); British Educational Research Association (BERA); German Educational Research Association (GERA/ DGfE); Swiss Society for Research in Education (SSRE/SGBF); see lists of partners, advisory board and cooperating experts on http://www.eerqi.eu/page/who-who) or direct approach to deliver lists of terms to the EERQI team which, from their point of view, represented markers of quality in research publications. Responses were compiled in a list which was delivered back to the participating experts, as well as to others, for comment on the assembled terms (see lists of partners, advisory board and cooperating experts on http://www.eerqi.eu/page/who-who). On the basis of feedback, a revised list of terms and their operationalisation was developed which was again presented to experts in different formats. One was the response format mentioned above. Other formats were presentations and collections of feedback in the framework of Symposia in European and national educational research conferences (e.g. the annual European Conference of Educational Research, 2010ff; the biannual German Conference of Educational Research Mainz 2010; public EERQI workshops in France, Germany and Switzerland). Moreover, the German Educational Research Association established a special think tank to accompany the process of clarification and condensation of the developing lists of terms and their explanations (DGfE-Strukturkommission, see http://www.dgfe.de/wir-ueber-uns/vorstandskommissionen.html). The members of this think-tank served as consultants of the review processes that were applied on the basis of feedback from presentations and consultations. Through this iterative process of consultation and review, the list of terms that were broadly accepted to indicate quality of educational research publications was reduced from around 180 to 14.

This list served as a basis for the first empirical test (pilot) of the EERQI peer review questionnaire. The terms and operationalisations were transferred to scales and items in the questionnaire. The questionnaire was delivered to selected educational research experts from the English, French/ Swiss, German and Swedish educational communities, thus representing the four EERQI languages. The experts were selected by appointees of the national educational research associations representing each language, by the EERA Executive Board and by EERQI's cooperating experts. Each of the selected persons was asked to apply the questionnaire to 15 research papers in his or her research language which had been randomly selected from the EERQI Database (see Gradmann et al. in this volume). Furthermore, the experts were asked for comments on the questionnaire

and the applied procedures. Roughly 30 reviewers were recruited for each project language. The review form included 14 scales, each containing two to four items. An ordinal Likert scale from one (completely agree) to four (completely disagree) was used for each item. An obligatory item for a short summary text and justification of the final judgment was further included, as well as the possibility to include additional comments. The resulting data was analysed by descriptive statistics and qualitative analysis. Specific values for the scales and items were calculated, redundant items were detected and eliminated, and the item scaling was tested.

The quantitative and qualitative analyses revealed that the pilot version of the questionnaire did not meet EERQI's own quality requirements. The convergent discrimination power showed moderate values between .45 and .77; the divergent discrimination power showed even lower values in most cases (.30 to .78). One scale termed 'style' showed high correlations with another scale termed 'rigour'. The scales showed satisfactory values for Cronbach's Alpha (.69 to .85), but the standard deviation for scale sums was very low in all cases. Interpreting values with respect to reliability was thus difficult.

The most important reasons for the unsatisfactory results were identified in the analyses. For instance, the roughly 100 reviewers turned out to be a very heterogeneous group in terms of respective expertise and professional background, the methodologies they preferred, the areas and genres they felt familiar with, and, not least, the experience they had with review processes. Sub-samples could not be analysed because of the small sample sizes and given the diversity of reviewer backgrounds. Moreover, the random selection of articles from the database and assignment to reviewers caused problems. Firstly, a very heterogeneous sample of texts was selected for the process. Secondly, the texts did not meet the reviewers' areas of expertise in a good number of cases. Despite these shortcomings, the pilot test yielded many constructive ideas for improving the design of the peer review exercise and the revision of the questionnaire. The following lessons could be learned from the pilot:

- The main survey would have to integrate a larger number of reviewers and a smaller number of articles, in order to create more data on each article.
- More detail on the professional background of reviewers should be collected in order to carry out deeper analysis of potential impact on the judgements.
- Scales should be longer and the item scaling should be expanded.

As the most important result of the exercise, however, five generic terms were identified as indicating the minimum requirements of quality in educational research publications. These terms were

- Rigour,
- Originality,
- Significance (for other researchers, policy and practice),
- Integrity (including considerations of authenticity, honesty and ethical requirements in the conduct of research) and
- Style (including clarity, communicability, eloquence and elegance).

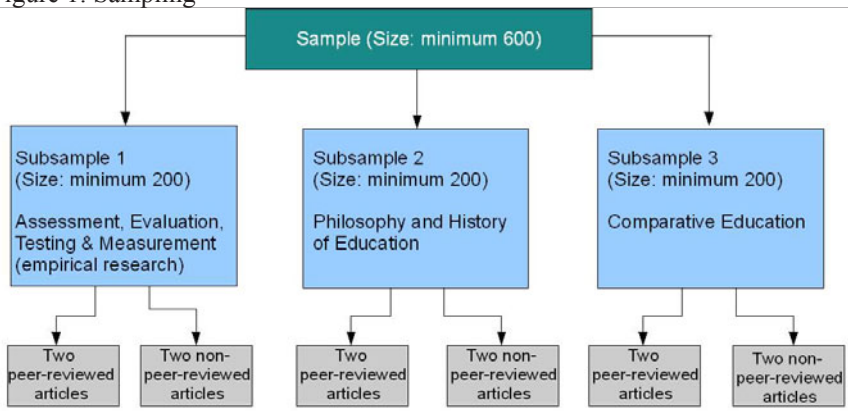
During the pilot phase, the 'integrity' and 'style' indicators proved to be dimensions of the three scales 'rigour', 'originality' and 'significance', rather than independent ones. – On the basis of the pilot results, the final part of the EERQI Peer Review Exercise was carried out. The results are reported in the following section.

3 The EERQI Peer Review Questionnaire: Two-step development

The final stage in developing the EERQI Peer Review Questionnaire comprised a two-step approach of a pre-test followed by a test. In step one, a test version of the final questionnaire based on selected texts was carried out by a number of educational research experts. In step two, a revised version of the instrument was delivered to a broader sample of reviewers. Based on the results of the pilot study, an *ex-ante* definition of the sampling procedure was applied in both steps. In consultation with the experts, three exemplary areas in educational research were identified which, from their point of view, represent the range of approaches in this research area and its sub disciplines. They are: Empirical research (assessment, evaluation, testing, measurement); philosophy and history of education; comparative education. The articles to be selected for testing derived from these three research approaches. As the survey could only be carried out with texts that had already been published and were part of the EERQI database, the team decided to select an equal number of articles that had been peer reviewed and articles that had been published without being subject to peer review. We wanted to test whether any corresponding distinctions could be re-identified in our data.

On the basis of these considerations, the following design was developed for the final two steps of the survey.

Figure 1: Sampling



The item scaling was enlarged from a five-step to a seven-step Likert scale in order to achieve more variation. Each part the questionnaire was expanded and redundant items were revised. Each part included quantitative and qualitative items. As well as calculating item values and reliability, item validity was carried out using procedures developed by Yousfi, Koopmann and Amelang (2005).

Step One

The revised questionnaire that was tested in step one included three different parts:

1. Indicators (3 scales, 38 items)

The revised version of the questionnaire contained three scales - rigour, originality and significance - as indicators of quality. In all earlier steps of the exercise, these terms were unanimously accepted by the participating experts and remained stable in the pilot testing. The 'rigour' scale was divided into five sections, each referring to a specific part of a text (abstract, introduction, methods and approaches, results, discussion). As 'integrity' and 'style' indicators had proved to be dimensions of the three aforementioned scales, rather than independent ones, items representing these aspects were integrated into the three remaining scales.

2. Demographic data

In the revised version of the questionnaire, more demographic data was collected. These data should help to provide a clearer picture of the reviewers' area of educational research expertise and their experiences as reviewers. Furthermore,

exploratory questions were also included in this part, such as "name three main criteria which must be fulfilled by an excellent article".

3. Miscellaneous

This section of the questionnaire consisted of closed and open questions. Three questions referring to 'rigour', 'originality' and 'significance' with ten-step Likert scales invited the reviewer to deliver a comprehensive judgement on a given article. These direct ratings were used in the statistical procedures as an initial examination of the items' validity.

Forty-five reviewers took part the first step of the final part of the peer review exercise. Although this was a smaller number of participants than those in the pilot, the results are more reliable due to the more standardized procedures (i.e. less variation in the articles and the research areas they represent). The following information illustrates the results of this step.

Demographic information: More than 60 per cent of the participating reviewers were female, 22 per cent male and 14 per cent unknown. The reviewers represented a broad age range, from 23 to 72 years. A total of 34 (76 per cent) reviewers worked at a university. Almost half of the participants mentioned that they worked as a professor or senior researcher. In general, the participants had a considerable amount of professional experience with answers revealing a mean of 14 years' working in research, but with a high standard deviation of 9.6 years. 63 per cent of the participants were experienced in reviewing research articles. Most participants worked in applied or basic research. Just one participant mentioned strategic research.

Results: Considering the fact that just a small number of participants took part in this step of the exercise, different tests were carried out in advance of calculating item characteristics and specific values. The correlations between the results for peer-reviewed versus non-peer-reviewed articles were highly significant ($p<0.01$). Also, the correlations between the subgroups were significant. The group 'philosophy and history of education' showed lower – but still significant – correlations with the results of the other groups. Perhaps this was due to the smaller sample size. All in all, no relevant differences between the subgroups could be found. Therefore all data were equally processed for the following analysis.

Item difficulty varied between .60 and .85. Most of the items showed a middle item difficulty. The convergent discrimination power varied between .65 and .85. The divergent discrimination power fluctuated between .10 and .74. The values for Cronbach's Alpha varied between .78 and .96. All values were examined in combination with the scales' length and the scales' standard deviation. Item validity was calculated using the procedure proposed by Yousfi,

Koopmann and Amelang (2005). The items were correlated with the direct ratings for 'rigour', 'originality' and 'significance'. Most values for item validity showed medium to high values (.60 – .80); solely the items related to 'abstract and keywords' (.3 – .6) showed significantly lower values. Thus we assumed that these do not provide any additional information about the overall quality of a text. The values for 'originality' (correlations between .67 – .76) and 'significance' (correlations between .59 – .77) show medium to high correlations. All correlations were statistically significant. Furthermore, the three scales showed high correlations with each other (.73 – .79), subscale 'abstract & keywords' being an exception. A factor analysis showed that the scales loaded on one main factor and two sub-factors.

The qualitative data aggregation (open answers to questions on quality) showed that reviewers predominantly referred to four main criteria when it comes to 'excellent research publications': 'Rigour' (with respect to coherence, clear presentation, results, argument, structure, etc.), 'significance' (for example, innovation and relevance), 'originality' and 'ethical aspects'. As the reviewers had to respond to the open questions before they started working with the questionnaire, we interpreted the high compliance with the terms developed by EERQI as a further signal of the general appropriateness of this instrument for peer review processes in educational research.

Step Two

Step two of the process aimed to optimise the questionnaire on the basis of the results achieved in step one. Important to this step was the reduction in the number of items in the instrument in order to increase its user-friendliness. Multiple iterations of the statistical procedures were carried out in order to reduce the number of items, while retaining a balance between items and item characteristic values. In order to meet this aim, a method of item reduction was carried out in which after the deletion of an item all values were again calculated for characteristics of reliability and validity. This process resulted in a questionnaire that included nine items for the 'rigour' scale, three items for 'originality' and four for 'significance'. The table below gives a detailed overview of the item values:

Table 1: Item values revised EERQI Peer Review Questionnaire

Scale	Subscale	Number of items	Reliability	Mean value for item validity
Rigour		9	.92	.76
	Methods & Approaches	3	.83	.72
	Results	2	.94	.64
	Discussion	4	.90	.82
Originality		3	.91	.78
Signifi-cance		4	.91	.78

The revised questionnaire also included the three items (direct ratings) for a general judgement on the quality of a text, which we again used for testing item validity. Furthermore, the reviewers had to indicate whether the article was related to their own area of expertise. The open questions were slightly reformulated, now asking the reviewers to describe what had to be improved in the reviewed article (and why), and whether there was anything missing in the article. All in all, the revised version of the questionnaire included 16 items concerning the indicators, three direct-ratings and four open questions.

Final Testing

For the final part of the study, articles for review were again selected from the three research areas mentioned above. The allocation procedure, however, was slightly altered. The reviewers could choose both the language and number of texts they wished to review. A further change re-focussed the exercise on just two of the four EERQI languages, namely English and German. The number of Swedish-language texts in the EERQI database was too low for another selection procedure. And due to a restructuring of the Institut National de Recherche Pédagogique, a participating institute from France, its continuance in the study was uncertain and the technical requirements for carrying out the exercise could not be met. In order to support the project under the given conditions, however, a group of French reviewers qualitatively evaluated the questionnaire's appropriateness and applicability and their feedback was included in to final analysis.

The most significant change in this step was the reduction of the EERQI languages to German and English. Moreover, the reviewers were not selected from the respective national contexts. Instead, reviewers from different member associations of the European Educational Research Association took part in the process, provided they had mastery of German and/or English. The main goals of this final part of the exercise were to further analyse the questionnaire's validity and to reach a cautious judgment on whether the questionnaire could function in different national and cultural research settings.

At this stage, the articles were assigned to reviewers in a two-step electronic process. Reviewers indicated to the EERQI research team their readiness to participate via a standardized input screen which contained filter questions relating to expertise, language command, as well as indicating how many articles he or she was willing to review. Texts were then randomly assigned to reviewers on this basis. The reviewers received anonymised versions of the texts (as per the entire process) for blind review. Each text had its own access code and the reviewers received passwords to access the texts that had been allocated to them. Reviewers also received a research ID in order to retain privacy. The assessment took place online.

Figure 2: Standardized input screen for reviewing process

This allocation procedure can be said to have functioned very well as the research areas of the texts were matched satisfactorily with those of the respective reviewers. Anecdotal evidence for this can be drawn from the following correspondence from one reviewer to the research team:

Dear Ms Gogolin,
Unfortunately I am unable to review the article that was 'assigned individually' (by whom?) to me…because I am the original author of the text. I assume that the text cannot be assessed by the author?

***Results*:**
In this stage of the exercise, 106 English and 73 German articles were assessed 653 times. Each article was rated one to ten times by different reviewers.
Statistical analyses were carried out separately for the German and the English subsample, as well as for both of them together. Further analyses were carried out with respect to the three areas of educational research. In addition to descriptive analyses of the scales' characteristics per subsample, nonparametric tests were carried out in pairs. The analysis showed no significant differences between the two subsamples (German or English). The Wilcoxon test shows a value of $p < 0.05$; yet this has to be interpreted carefully on account of the different sample sizes. Also, the analysis for the three areas of research showed no significant differences. The values for Cronbach's Alpha vary between .73 and .96. A comparative analysis between these values and the results of the first step shows no significant differences. The values for item validity show a large range. Most 'originality' items and all pertaining to 'significance' were satisfactory. A detailed inspection of values for the 'rigour' scale showed that the whole scale, but not the single items, lead to an acceptable correlation with the results of the direct rating. The scales explain 50 to 60 per cent of the variance of the direct ratings. Calculations in which the 'not applicable'-option is excluded showed even higher values for the coefficient of determination ($R^2 \approx .70$). In general, the scales' and item characteristics show good values, also for the subsamples of different areas of educational research and the two language groups.

Further analyses of the qualitative data (open answers on the quality of the procedure and instrument) also showed very satisfactory results. The group of French reviewers gave positive feedback. They evaluated the questionnaire as helpful to the review process and stated that the criteria – rigour, originality, significance – were appropriate for the French subsample. With respect to the qualitative feedback from other participants, the questionnaire seems to be adequate for review processes in national and international contexts, as well as in the three examined areas of educational research. The values for item validity, however, advise that the scale for 'rigour' should only be used as a whole.

4 Discussion

The main aims of the EERQI Peer Review Exercise were met. It could be shown that educational researchers across different sub disciplines were able to agree upon a set of generic quality indicators as well as their operationalisation. This would seem to be appropriate for the assessment of quality in research publications. The indicators seem further useful for a broad range of approaches and research traditions which are used in educational research, whether they derive from social sciences or the humanities. Research which is embedded in social scientific traditions, however, seems to relate better to the indicators and items than works which are inspired by the humanities, e.g. for history or philosophy of education. The combination of a small number of scales and items, combined with an open feedback format – as introduced in the final version of the questionnaire – was highly appreciated in the qualitative part of the evaluation. A specific advantage of the questionnaire, according to feedback received, is its brevity in combination with intelligibility of the items. This contributes to its usefulness in light of growing demand in the area of research assessment.

Open questions remain, however. To what extent can applying the peer review questionnaire (be it the EERQI questionnaire or another good quality instrument) solve problems in quality detection? A special question is related to the ongoing debate in the EERQI project on the appropriateness and function of 'extrinsic' and 'intrinsic' indicators in this problem. In approaching this question, two experimental methods were carried out that attempted to identify the relation of both types of indicators. The results of these approaches are presented in the contributions from Mooj and Hilf/ Severiens in this volume. Both approaches show that the goal of detecting direct relations between both types of indicators could not be achieved by the EERQI project – but some fundamental steps forward have been taken in this direction.

5 References

Botte, A., & Vorndran, A. (2008). *Analysis and Evaluation of Existing Methods and Indicators for Scientific Quality Assessment*. Report. http://www.eerqi.eu/sites/default/files/Analysis_and_evaluation_of_existing_methods_and_indicators.pdf. Deutsches Institut für Internationale Pädagogische Forschung (DIPF). Frankfurt.

Bridges, D. (2009). Research quality assessment: impossible science, possible art? *British Educational Research Journal, 35*(4), 497-517.

Hicks, D., & Wang, J. (2009). Towards a Bibliometric Database for the Social Sciences and Humanities – A European Scoping Project. Georgia: Georgia Institute of Technology.

Meester, W. (2013). Value of Bibliometrics.Towards a comprehensive citation index for the Arts & Humanities. *Research Trends,* (32). Retrieved from http://www.researchtrends.com/issue-32-march-2013/towards-a-comprehensive-citation-index-for-the-arts-humanities/ website:

Moed, H. F., Linmans, J., Nederhof, A., Zuccala, A., López Illescas, C., & de Moya Anegón, F. (2009). Options for a Comprehensive Database of Research Outputs in Social Sciences and Humanities. Research report to the Project Board of the Scoping Study "Towards a Bibliometric Database for the Social Sciences and the Humanities" set up by the Standing Committees for the Social Sciences and the Humanities of the European Science Foundation (ESF). Version 6 April 2009. Leiden, Granada: Centre for Science and Technology Studies (CWTS), Leiden University, The Netherlands; SCIMago Research Group, CSIC Madrid and University of Granada, Spain.

Rey, O. (2006). What is 'good' research in education? *Veille scientifique et technologique*. Lyon: Institut National de Recherche Pédagogique.

Yousfi, S., Koopmann, B., & Amelang, M. (2005). *Correlates of item validity. On the eminent importance of global self-ratings.* University of Standford; University of Heidelberg. Heidelberg, Stanford.

A Prototype Empirical Framework of Intrinsic and Extrinsic EERQI Indicators

Ton Mooij

Short Summary
The research question to be answered in the context of the EERQI project is: What do statistical analyses show us about the relationships between intrinsic and extrinsic indicators of quality and what does this mean when constructing a prototype EERQI framework?

The pilot study involved the scoring on both intrinsic and extrinsic indicators for 177 research documents or articles written by 268 authors. Intrinsic data were gathered by peer review and extrinsic data were collected from the Internet. Confirmatory Factor Analysis (CFA) resulted in a measurement model containing three intrinsic and two extrinsic latent factors. There are significant correlations between the intrinsic latent factors and between the extrinsic latent factors; however, no significant correlations have been found between intrinsic and extrinsic factors. This outcome underlines the notion that intrinsic indicators may add specific quality information to an EERQI prototype that consists solely of extrinsic indicators, and vice versa. Testing by means of a structural model revealed that the more a reviewed document is related to the reviewer's own area of research, the higher the score the reviewer gives the document with respect to 1) significance, originality and consistency and 2) methodological adequacy. No relationships were found between the reviewer's own area of research and the extrinsic latent factors. These effects on the two intrinsic latent factors indicate that there may be some subjective evaluation bias in peer reviewing.

The conclusion is that the outcomes of the statistical analyses seem plausible and support the validity of the conceptual framework. An initial prototype EERQI framework has been constructed, which is in line with the main goal of the EERQI project. Although the pilot had some methodological limitations, the present empirical outcomes are promising for future EERQI developmental and research activities, which could, for example, also integrate semantic latent factors and indicators.

I. Gogolin et al. (eds.), *Assessing Quality in European Educational Research,*
DOI 10.1007/978-3-658-05969-9_9,

1 Introduction

Impact indicators in educational research such as those based on data from Social Science Citation Index (SSCI) were long based on measures of citations to documents in specific scientific journals. Nowadays, search engines automatically use various types of 'objective', 'external', or 'extrinsic characteristics', e.g. the bibliometric or semantic characteristics of publications or documents found in many different Web-based sources. The goal of the international 'European Educational Research Quality Indicators' project (EERQI; FP7 # 217549) is to improve citation-only assessments of the quality or impact of educational and other research (cf. Gogolin, 2008; Gradmann, Sieber, & Stoye, 2011). Therefore, in addition to bibliometric and citation indicators, efforts were also made to distinguish specific indicators reflecting the more 'subjective', 'internal', or 'intrinsic' quality of research documents. Intrinsic indicators refer to the content of a publication or research document and are supposed to explicate or describe such aspects as rigour, originality, significance, integrity, and style (cf. Bridges, 2009).

In earlier EERQI papers (Mooij, 2008a, 2008b), I outlined a possible approach to empirically exploring and analysing relationships between sets of intrinsic and extrinsic indicators of the quality of research documents. I also analysed the statistical relationships between both intrinsic and extrinsic indicators in order to construct an initial EERQI prototype framework (Mooij, 2011). In this chapter I concentrate on the main aspects and outcomes of the empirical research involved in constructing this prototype. The research question to be answered is: What do statistical analyses show us about the relationships between intrinsic and extrinsic indicators of quality and what does this mean when constructing a prototype EERQI framework?

2 Intrinsic and extrinsic quality indicators

2.1. Intrinsic indicators

In the EERQI project, intrinsic indicators were chosen to operationalise the concepts: methodology, results, discussion, originality, significance, validity, and miscellaneous. Twenty items were devised to assess these seven quality concepts: see for these concepts and items Table 1. The answer alternatives for each item were: 'not relevant for this text' (=0), 'very poor' (=1), (2), (3), 'average' (=4), (5), (6), and 'excellent' (=7). A final item, item 21, allowed peer reviewers to indicate how closely the document they had evaluated related to their own area of research. Here the answer categories were: 'Very closely' (=1), 'Closely'

(=2), ‘Less closely’ (=3), ‘Not at all’ (=4). The complete dataset resulting from the final pilot in the EERQI project consists of 177 research documents or articles written by a total of 268 authors. Peer reviewers scored these documents with respect to all 21 items. For each document, peer review evaluation scores were aggregated by calculating their mean across reviewers.[26]

[26] The dataset containing both the intrinsic and extrinsic scores of 177 research documents became available on 2 March 2011. The dataset contains scores by peer reviewers who are partners in the EERQI project or attended the European Conference on Educational Research in 2010. Some of the reviewers scored two or more research articles. If available per document, the scores of various reviewers were aggregated. It seems that value 0 (‘not relevant for this text’) was included in these scores, however. This problem could not be avoided because only the aggregated data were available. The 177 documents represent three different European languages. In combination with the small number of reviewers, the actual data structure does not permit assessment of interobserver reliability or multilevel analyses between and within languages and/or reviewers, respectively.

Table 1 – Concepts and items assessing intrinsic quality (n documents=171)

Concept_var.	Description of variable or item	Min.	Max.	M	SD
1 Methods_1	The methods are intelligibly described	.00	7.00	4.02	2.03
2 Methods_2	The method / approach is appropriate	.00	7.00	4.70	1.63
3 Methods_3	The method / approach is accurate	.00	7.00	4.34	1.78
4 Results_1	The results are completely described	.00	7.00	4.51	1.66
5 Results_2	The results are correctly described	.00	7.00	4.53	1.67
6 Discussion_1	The study's method is reflected in an appropriate way	.00	7.00	3.94	1.82
7 Discussion_2	The study's results are reflected in an appropriate way	.00	7.00	4.51	1.69
8 Discussion_3	The pattern of reasoning is consistent	1.00	7.00	5.48	1.10
9 Discussion_4	The discussion shows a critical evaluation of the work	.00	7.00	4.67	1.47
10 Originality_1	The study shows new approaches in its methodological procedures	.00	7.00	3.39	1.63
11 Originality_2	The study shows new approaches in the structure of its argumentation	.00	7.00	4.16	1.35
12 Originality_3	The study contributes innovative ideas for the state-of-art in its research area	.50	7.00	4.52	1.33
13 Significance_1	The study contributes to the development of its research field	1.00	7.00	5.02	1.28
14 Significance_2	The study makes a significant contribution to the latest discussions within the research field	1.00	7.00	4.82	1.30
15 Significance_3	The study makes a significant contribution to the latest discussions within the educational policy field	.00	7.00	4.62	1.52
16 Significance_4	The study makes a significant contribution to the latest discussions within the educational practice field	.00	7.00	4.51	1.57
17 Validity_1	How do you evaluate the article concerning its Rigour?	.00	7.00	4.72	1.38
18 Validity_2	How do you evaluate the article concerning its Originality?	1.00	7.00	4.82	1.05
19 Validity_3	How do you evaluate the article concerning its Significance?	1.00	7.00	5.03	1.22
20 Miscellaneous2	Comparing this article to an article representing good research, where would you place it on a scale from 1 to 7, with 7 being excellent quality and 1 being bad quality?	1.00	7.00	4.61	1.11
21 Miscellaneous1	The reviewed article is related to my own area of research…	1.00	4.00	2.40	0.54

In the univariate analysis using the Statistical Package for the Social Sciences (SPSS, version 17.0), only documents without system-missing values were used, which resulted in item-specific information for 171 documents. Table 1 also

presents the descriptive statistics of these intrinsic items. The means vary from around 4 (average) to 5; standard deviations vary from 1.05 to 2.03.

2.2. Extrinsic indicators

Extrinsic indicators usually measure aspects of research documents such as number or distribution of citations (per author; across authors; per document; hits resulting from search engines for a paper or author/combination of authors, and so forth). The information on extrinsic indicators was provided per author. Because research documents constitute the unit of analysis, the extrinsic information was aggregated per document. When there was more than one author per document, the available information per indicator was aggregated by totalling the scores of the authors per document.[27] The dataset of 2 March 2011 contains information about 12 extrinsic indicators. Five of these were neglected.[28] Information about the remaining seven extrinsic indicators, their range of scores, means and standard deviations is given in Table 2.

Table 2 – Variables assessing extrinsic quality (n documents=171)

Variable name	Description	Min.	Max.	M	SD
1 Cit/paper	Citations per paper without self-citations using full title of the article	.00	804.81	18.48	64.36
2 WebMennAuth	Web mentions of author in search engine BING; number of URLs of pages matching the query submitted	2.00	1791.00	352.23	280.33
3 WebMentTitle	Web mentions of article title in search engine BING; number of URLs of pages matching the query submitted.	.00	1046.00	25.24	131.59
4 GoogleHits	Google Web Search results	.00	3265.00	219.91	448.85
5 MetagerHits	Metager hits	.00	133.00	4.74	16.16
6 CiteULikeHits	Mentions of article CiteULike	.00	486.00	21.32	60.55
7 LibraryThingHits	Mentions of article LibraryThing	.00	651.00	29.34	89.95

[27] Identification of documents and authors is based on the variable 'revID' (named 'CODE' in earlier datasets). Each record starts with the character 'd' or 'e', followed by a number; sometimes another character has been added. Each additional character appears to represent another record in the database, possibly identifying specific authors in multi-author documents.

[28] These are 'ConnoteaHits', 'MendReader', 'Downloads08', 'Downloads09', and 'Downloads10'. The reasons were that scores on all documents were 0 for the first two variables; the Downloads variables had many missing values.

The variable 'number of citations per paper' [Cit/paper] has a very skew distribution to the right.[29] The respective scores were therefore transformed by taking their square roots. The range of the transformed scores is 0.00 – 28.37 with Mean 3.24 and SD 2.83. Principal factor analysis was used to explore the relationships between the seven extrinsic variables listed in Table 2. The variables WebMentTitle and MetagerHits are not related to the other variables or only to a very limited extent. Given the present focus, it was decided to drop these two variables.

The Eigenvalues and percentages of variance of the remaining five variables point to the presence of two underlying factors: see Table 3.

Table 3 – Eigenvalues and % of variance for extracted factors of five extrinsic variables

Factor	Eigenvalue	% of Variance	Cumulative %
1	2.612	52.236	52.236
2	1.152	*23.039*	75.276
3	.606	*12.126*	87.401
4	.444	8.882	96.283
5	.186	3.717	100.00

The loadings of the five variables on the two factors were rotated (oblique, geomin) within the EFA procedure of statistical program MPlus 6.1: see Table 4. The results in Table 4 illustrate that 'Citations per paper (without self-citations)' and 'Web mentions of author in search engine BING' represent factor 1, whereas the second factor represents numbers of hits by three other search engines.

Table 4 – Factor loadings of extrinsic variables after oblique (geomin) rotation

Variable name	Description	Factor	
		1	2
Cit/paper (sqrt)	Citations per paper without self-citations using the full title of the article	0.921	-0.001
WebMennAuth	Web mentions of author in search engine BING; number of URL's of pages matching the query submitted	0.405	0.098
GoogleHits	Google Web Search results	0.023	0.947
CiteULikeHits	Mentions of article CiteULike	0.000	0.689
LibraryThingHits	Mentions of article LibraryThing	-0.112	0.867

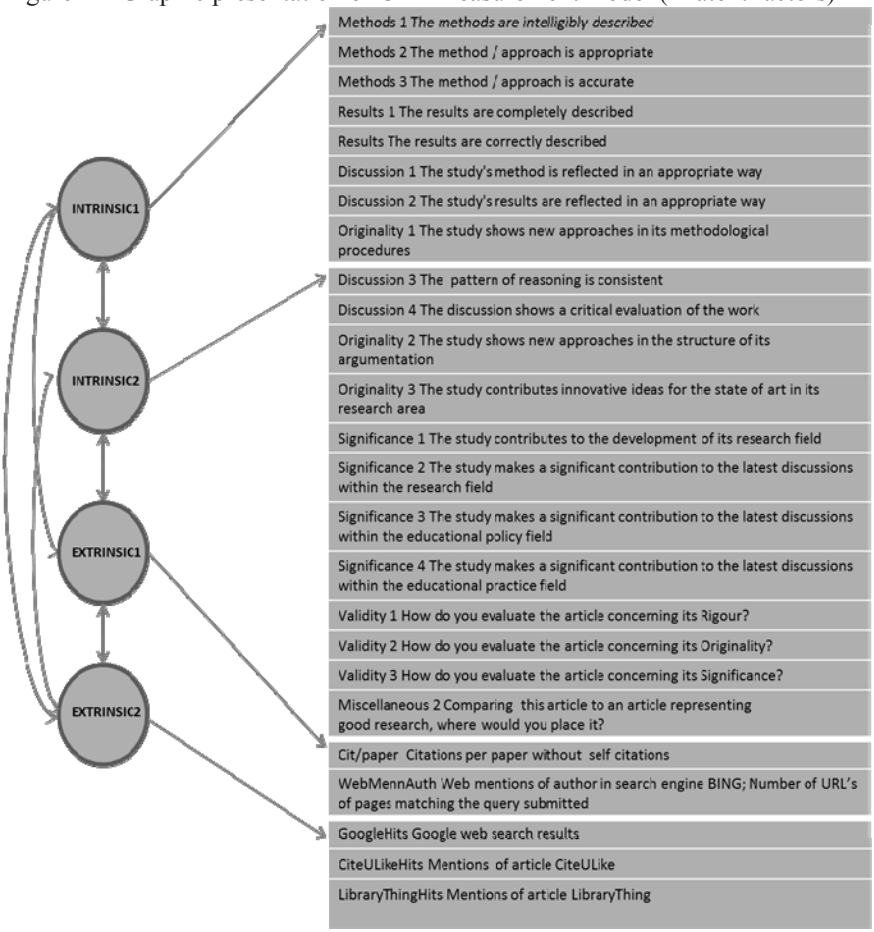
[29] The value '0' may reflect 'missing value' or 'no hits'/'no citations'. In this paper, the latter ('no #') is assumed.

3 Relationships between intrinsic and extrinsic indicators

3.1. Modelling intrinsic and extrinsic latent factors

The study used the results of an earlier factor analysis based on the intrinsic variables of Table 1 to construct a measurement model with two intrinsic factors and two extrinsic factors (cf. Table 4). The model is given in Figure 1.

Figure 1 – Graphic presentation of CFA measurement model (4 latent factors)



In Figure 1, latent factor 'Intrinsic1' represents the intrinsic indicators methodological adequacy, completeness and correctness of reporting results, appropriateness of discussion, and originality with respect to methodological procedures. *Intrinsic1* therefore indicates *methodological adequacy of the document.* Latent factor 'Intrinsic2' stands for logical consistency, critical evaluation, innovation, various types of significance and overall evaluation of the information in a document. *Intrinsic2* thus represents *significance, originality and consistency of the document.* Furthermore, latent factor 'Extrinsic1' refers to number of citations per document without self-citations and Web mentions of author by search engine BING. *Extrinsic1* therefore indicates *number of citations and Web mentions by BING.* Latent factor 'Extrinsic2' rather univocally represents number of hits obtained with search engines Google, CiteULike, and LibraryThing. *Extrinsic2* is thus associated with *number of hits in three specific search engines.*

Figure 1 specifies a 'Confirmatory Factor Analysis' (CFA) to check the relationships between each latent factor and specific indicators or observed variables, while taking account of the correlations between various latent factors.[30] The variance of each observed indicator variable is explained by both the regression on the specific latent factor and specific error variance; error variances between observed indicators may be correlated. The statistical program MPlus (version 6.1) was used to simultaneously check the fit of the measurement model in Figure 1 against the intrinsic scores (Table 1) and the extrinsic scores (Table 2). The outcomes of Maximum Likelihood analysis are given in Table 5.

[30] In the measurement model of Figure 1, the relationships between the four latent factors are standardised to facilitate their interpretation as correlations. Correlations between factors are free to vary. These correlations are represented by the two-way arrows between all pairs of latent factors. The regressions of each of the indicator variables on their respective latent factor are represented by one-way arrows. The total variance of each factor is set to 1.

Table 5 – ML parameter estimates (standardised) of measurement model Fig. 1

Latent factors	Factor loadings				
Indicators	INTRINS1: Methodolog. adequacy	INTRINS 2: Sign./orig./ consist.	EXTRINS1: # cit./Web BING	EXTRINS2: Hits 3 searc. eng.	R^2
Methods_1	0.912**				0.832**
Methods_2	0.826**				0.683**
Methods_3	0.882**				0.777**
Results_1	0.784**				0.615**
Results_2	0.791**				0.626**
Discussion_1	0.881**				0.777**
Discussion_2	0.781**				0.609**
Discussion_3		0.656**			0.430**
Discussion_4		0.612**			0.375**
Originality_1	0.776**				0.603**
Originality_2		0.796**			0.634**
Originality_3		0.873**			0.763**
Significance_1		0.900**			0.809**
Significance_2		0.910**			0.829**
Significance_3		0.809**			0.654**
Significance_4		0.721**			0.520**
Validity_1		0.542**			0.294**
Validity_2		0.785**			0.616**
Validity_3		0.842**			0.708**
Miscellanous2		0.840**			0.706**
Cit/paper (sqrt)			0.592**		0.350**
WebMennAuth			0.685**		0.469**
GoogleHits				0.980**	0.960**
CiteULikeHits				0.674**	0.455**
LibrarThingHits				0.803**	0.645**

	Factor covariances (correlations)		
	INTRINSIC1	INTRINSIC2	EXTRINSIC1
INTRINSIC2	0.631**		
EXTRINSIC1	0.239*	0.148	
EXTRINSIC2	0.147	0.085	0.460**

Fit indices: X^2 (269)=1028.656 (p=0.000); RMSEA=0.129; SRMR=0.072.
* $0.01 \leq p \leq 0.05$; ** $p < 0.01$.

The overall fit of the model is reflected in two statistical indices, the 'Root Mean Square Error of Approximation' (RMSEA) and the 'Standardized Root Mean Square Residual' (SRMR): see the note following Table 5. Both measures are

related to the Chi-Square statistic. Both indices are influenced by the sample size, which implies that a smaller sample results in a less favourable fit. Generally, a value above 0.10 on both indices is considered to indicate a bad fit. With respect to the results in Table 5, it can be seen that RMSEA=0.129 and SRMR=0.072.

Table 5 furthermore demonstrates a strong correlation between the two intrinsic factors (0.631) and a weaker correlation between the two extrinsic factors (0.460). The correlation between Intrinsic1 (methodological adequacy of the document) and Extrinsic1 (number of citations and Web mentions by BING) is also significant (0.239; $p \leq .05$). This outcome illustrates some overlap between intrinsic and extrinsic indicators, a finding that merits more attention for reasons of both EERQI interpretation and modelling.

The other correlations between intrinsic and extrinsic latent factors are not significant statistically. This implies that the use of intrinsic indicators may add quality information to an EERQI that consists solely of extrinsic indicators, or that the introduction of extrinsic indicators may add quality information to an EERQI containing only intrinsic indicators.

Given the data available and the small sample size, the overall results in Table 5 confirm the first empirical check of the validity of the measurement model in Figure 1. Moreover, the confirmatory factor loadings and the variances explained per indicator (R^2) are relatively large. However, inspection of the modification indices reveals that it may be possible to improve Figure 1.

To explore the statistical consequences, some alternative models were constructed and checked against the model presented in Figure 1 and Table 5. An overview of the alternative models and their statistical outcomes is given in Table 6.

Table 6 – Comparison of different CFA models

Alternative measurement models	X^2	df	RMSEA	SRMR
1. Model with 4 latent factors (2 intrins., 2 extrins.; Figure 2)	1028.66	269	0.129	0.072
2a. As Model 1, but with error covariation Result_1 - Result_2	785.23	268	0.106	0.070
2b. Model with 5 latent factors (3 intrinsic, 2 extrinsic; Fig. 4)	758.39	265	0.104	0.077

In Table 6, model 1 is the model given in Figure 1 and Table 5. Model 2a of Table 6 allows correlation between result indicators Results_1 and Results_2. Compared to model 1, model 2a demonstrates a decrease in Chi-Square of 243.424 with a difference of only one degree of freedom (df). This difference between model 1 and model 2a is highly significant: model 2a results in a signif-

icant improvement in model 1. This is also shown in the values of RMSEA (0.106) and SRMR (0.070).

Additional explorative analysis of various parameters suggests combining intrinsic indicators Results_1, Results_2 and Discussion_2. This implies that there are three rather than two intrinsic latent factors, which changes the CFA model of Figure 1 into the CFA model of Figure 2 (see next page).

The statistical outcomes in Table 7 illustrate that, compared to CFA model 1, CFA model 2b (Figure 2) results in a significant improvement in Chi-Square (270.271; df=4; $p<.01$) and acceptable values for both RMSEA (0.104) and SRMR (0.077). Like the outcome of Table 5, this result concerning the relationships between intrinsic and extrinsic latent factors in Table 7 merits more attention for reasons of both interpretation and modelling in the EERQI conceptual framework. Moreover, this empirical outcome again supports the notion that using intrinsic indicators may add specific quality information to an EERQI consisting solely of extrinsic indicators and that introduction of extrinsic indicators may add specific quality information to an EERQI framework containing only intrinsic indicators.

Figure 2 – Graphic presentation of CFA measurement model (5 latent factors)

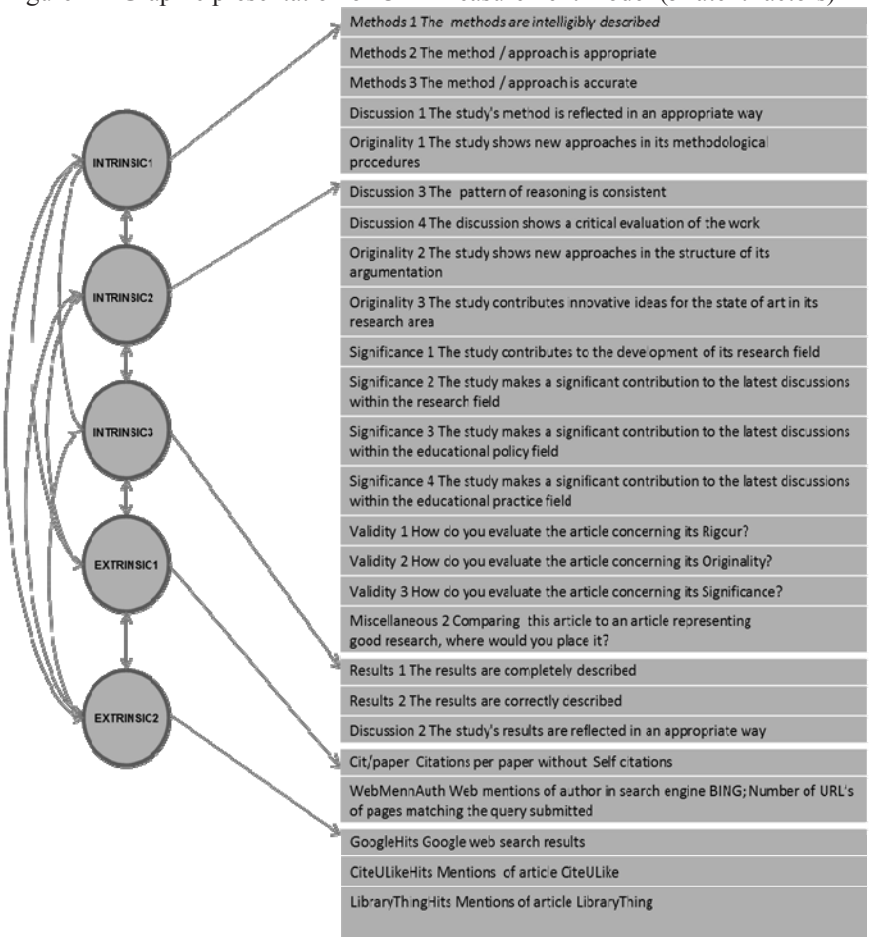


Table 7 – ML parameter estimates (standardised) of measurement model Fig. 2

	Factor loadings					
	INTRINS1: Method-ol. adequacy	**INTRINS2: Sign./orig. / consist.**	**IN-TRINS3: Results**	**EX-TRINS1: # cit./ WebBING**	**EXTRINS2: Hits 3 search**	**R2**
Methods_1	0.907**					0.823*
Methods_2	0.862**					0.743*
Methods_3	0.914**					0.835*
Results_1			0.968**			0.937*
Results_2			0.975**			0.951*
Discussion_1	0.881**					0.776*
Discussion_2			0.787**			0.620*
Discussion_3		0.655**				0.429*
Discussion_4		0.611**				0.374*
Originality_1	0.787**					0.619*
Originality_2		0.796**				0.634*
Originality_3		0.873**				0.763*
Significance_1		0.900**				0.810*
Significance_2		0.911**				0.829*
Significance_3		0.809**				0.655*
Significance_4		0.721**				0.520*
Validity_1		0.542**				0.294*
Validity_2		0.785**				0.617*
Validity_3		0.842**				0.709*
Miscellaneous2		0.840**				0.705*
Cit/paper (sqrt)				0.591**		0.349*
WebMennAuth				0.686**		0.470*
GoogleHits					0.980**	0.960*
CiteULikeHits					0.674**	0.455*
LibraryTh-					0.803**	0.645*

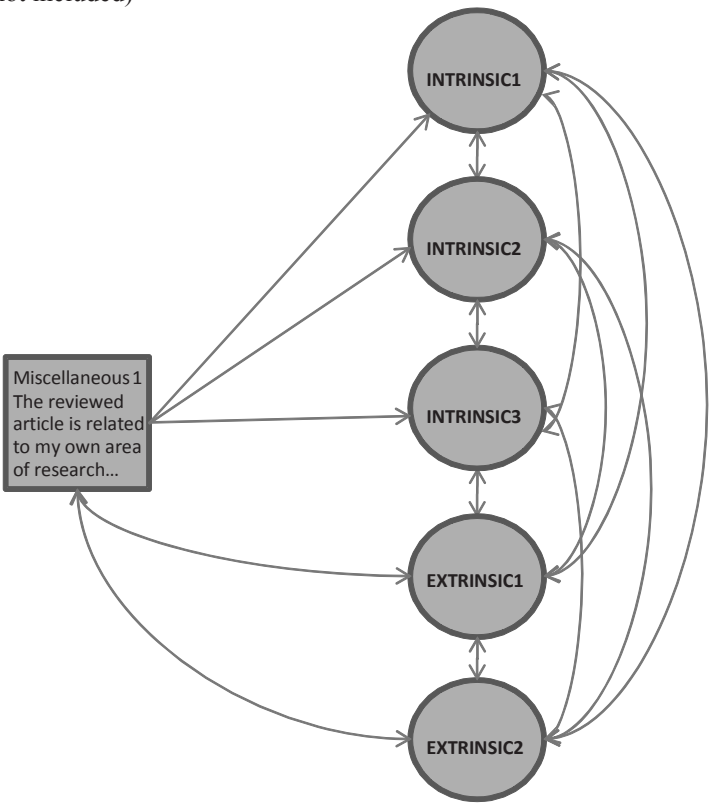
	Factor covariances (correlations)			
	INTRINSIC1	INTRINSIC2	INTRINSIC3	EXTRINSIC1
INTRINSIC2	0.620**			
INTRINSIC3	0.740**	0.476**		
EXTRINSIC1	0.236	0.148	0.188	
EXTRINSIC2	0.146	0.085	0.113	0.460**

Fit indices: X^2 (265)=758.385 (p=0.000); RMSEA=0.104; SRMR= 0.077.
* $0.01 \leq p \leq 0.05$; ** $p < 0.01$.

3.2. Structural model of intrinsic and extrinsic latent factors

A final exploration seeks to explain the latent factors within the CFA model in Figure 2. It is hypothesised that the degree to which the reviewed article or document is related to the reviewer's own area of research (item 21 or Miscellaneous1 in Table 1) influences the scores of the intrinsic latent factors. Inclusion of this explanatory variable in the CFA model of Figure 2 transforms this model into a causal or structural model. The causal relationships are represented by the three one-sided arrows between item 21 and the intrinsic latent factors: see the structural latent factor model in Figure 3.

Figure 3 – Structural model with intrinsic and extrinsic latent factors (indicators not included)



In Figure 3, the specific indicators for the latent factors are the same as those in Figure 2. Moreover, Figure 3 illustrates that the three intrinsic latent factors are regressed on the explanatory item Miscellaneous1 ('The reviewed article is related to my own area of research'). The correlations between the explanatory item and the two extrinsic factors are free to vary. The main results of Maximum Likelihood (ML) analysis using MPlus (version 6.1) are given in Table 8.

Table 8 - ML factor parameter estimates (standardised) of structural model

	Factor covariances (correlations)				
	INTRINSIC1	INTRINSIC2	INTRINSIC3	EXTRINSIC1	
INTRINSIC2	0.604**				
INTRINSIC3	0.735**	0.463**			
EXTRINSIC1	0.247	0.162	0.195		
EXTRINSIC2	0.147	0.091	0.113	0.461**	
	Direct effects				
	INTRINSIC1	INTRINSIC2	INTRINSIC3	EXTRINSIC1	EXTRINSIC2
Miscellaneous1	-0.176*	-0.247**	-0.128		
	Correlations				
Miscellaneous1				0.029	0.020

Fit indices: X^2 (284)=779.559 (p=0.000); RMSEA=0.101; SRMR= 0.075.
* $0.01 \leq p \leq 0.05$; ** $p < 0.01$.

Miscellaneous1 has significant effects on latent factors Intrinsic2 (-0.247) and Intrinsic1 (-0.176); the effect on Intrinsic3 (-0.128) is non-significant. This means that the more closely the reviewed document is related to the reviewer's own area of research, the higher the reviewer's evaluation scores with respect to significance, originality and consistency (Intrinsic2) and methodological adequacy (Intrinsic1). The two effects seem to reflect some subjective evaluation bias that may occur in peer reviewing.

Furthermore, Table 8 shows no significant statistical relationships between Miscellaneous1 and the extrinsic factors. The correlation between intrinsic factors – and not extrinsic factors – and the relevance of the reviewed document to the reviewer's own area of research supports the validity of the conceptual framework.

4 Conclusions

4.1 An initial prototype framework of EERQI indicators

A pilot was conducted to collect data with respect to 'intrinsic' and 'extrinsic' indicators of research documents. The research question was: What do statistical analyses show us about the relationships between the two types of indicator and what does this mean when constructing a prototype EERQI framework? To answer this question, peer review scores or intrinsic indicators were used to assess aspects of research articles or documents related to methodology, results, discussion, originality, significance, validity, and miscellaneous matters. Extrinsic indicators for the same documents were Web-based.

Some consecutive measurement models and their empirical results confirmed the potential relevance and functionality of intrinsic latent factors, extrinsic latent factors, and their indicators. A final check was whether the degree to which a reviewed article is related to the reviewer's own area of research influences the scores of the intrinsic latent factors. Empirical testing in a causal structural model indeed revealed that the more the reviewed document is related to the reviewer's own area of research, the higher reviewer's evaluation scores with respect to 1) significance, originality and consistency, and 2) methodological adequacy. There are no significant relations between the reviewer's own area of research and the extrinsic factors.

The differentiated relationships and outcomes of this pilot support the validity of both the conceptual framework and the empirical research. The conclusion is that an initial prototype EERQI framework has been constructed. The relevant conceptual framework is presented in Figure 3 and Figure 2 combined. It is possible that other types of information, for example semantic indicators and factors, can be integrated into these figures and follow-up research.

Given the statistical outcomes related to Figures 3 and 2 in Tables 5 – 8, a further conclusion is that an initial empirical test of the conceptual EERQI framework has been successful. The main goal of the EERQI project – to improve citation-only assessments of the quality or impact of educational and other research – has been supported. It is, however, important to describe some limitations of the study.

4.2 Methodological limitations

From a methodological point of view, the present pilot has a number of limitations which any follow-up analyses or research should seek to eliminate. These include:

- the exact operationalisation and assessment of both extrinsic and intrinsic indicators need careful consideration, for example for reasons of validity and representativeness;
- the pilot covered only a fairly small number of documents and reviewers;
- the ratio of number of documents to number of indicators (171:25) is relatively low;
- the distribution characteristics of the variables and their consequences for statistical analysis merit greater attention;
- the interobserver reliability of the reviewing information also merits greater attention;
- possible effects of language differences have not been taken into account;
- multilevel analysis was not applied because of the small number of documents/reviewers.

4.3 Future steps

The EERQI project has many different sides to it and considerable potential. In the future, more of the project partners and other parties may be convinced of its significance, originality and consistency (latent factor Intrinsic 2). Exploiting its potential and improving the focus on both methodological adequacy (latent factor Intrinsic 1) and semantic indicators and latent factors should optimise the steps to the further construction and use of EERQI.

5 References

Bridges, D. (2009): Research quality assessment: impossible science, possible art? British Educational Research Journal, 35(4), 497-517.

EERQI project (2010): State of the art reports on EERQI project parts. Preparatory meeting for the second EERQI workshop, 18-19 March. EERQI Project Report.

Gogolin, I. (2008): European Educational Research Quality Indicators (EERQI). (Project 217549). FP7 Collaborative project. Hamburg, Universität Hamburg.

Gradmann, S., Sieber, J., & Stoye, D. (2011): Extrinsic indicators used in EERQI. Berlin: Humboldt-Universität, Institut für Bibliotheks- und Informationswissenschaft.

Mooij, T. (2008a): Suggestions for a first conceptual framework to construct EERQI. Contribution to the international project 'European Educational Research Quality Indicators' (Project 217549; FP7). Nijmegen, The Netherlands: Radboud University, ITS.

Mooij, T. (2008b): Intermediate conceptual framework and procedures to construct EERQI. Contribution to the international project 'European Educational Research Quality Indicators' (Project 217549; FP7). Nijmegen, The Netherlands: Radboud University, ITS.

Mooij, T. (2011): European Educational Research Quality Indicators (EERQI): A first prototype framework of intrinsic and extrinsic indicators. Paper for the final conference in Brussels, University Foundation, 15-16 March 2011 of the international collaborative project 'European Educational Research Quality Indicators' (EU Project 217549; FP7). Nijmegen, The Netherlands: Radboud University, ITS.

Nolin, J., & Åström, F. (2010): Turning weakness into strength: Strategies for future LIS. Journal of Documentation, 66(1), 7-27.

Sándor, Á., & Vorndran, A. (2009): Detecting key sentences for automatic assistance in peer reviewing research articles in educational sciences. EERQI Project Report.

Sieber, J., & Stoye, D. (2011): Description of a Measure. Berlin: Humboldt-Universität, Institut für Bibliotheks- und Informationswissenschaft.

A scientific editor's support tool: Design, analysis and value

Thomas Severiens, Eberhard R. Hilf

Short Summary
This article addresses the requests for an online tool to aid publishers of scientific journals in their work to pre-filter the mass of incoming papers for scientific quality. As a first step the Hamburg EERQI group[31] has analyzed the process of scientific refereeing by developing a detailed questionnaire and asking a group of referees to re-referee a large, preselected stack of previously published scientific papers in educational research (see Gogolin/Stumm in this volume). We proceeded to extract information about the refereeing process from the wealth of statistical data gathered in their survey: We used a linear weighted combination of the detailed accumulated answers, here called *intrinsic* parameters, to determine the best fit to the overall judgment of the referees on a paper. That is, we try to compare the contributions of the individual aspects to the whole.

In a second line of research, the Berlin EERQI group[32] developed a tool called *aMeasure* to estimate the scientific footprint of the selected papers and their authors in Internet archives and services by recording the number of hits to a specific set of queries, here called *extrinsic* parameters. We analyzed the results and extracted information on the correlation between web-footprints such as the number of citations of a paper, as compared to the best correlated combination of the intrinsic parameters. One question was: how much does the overall judgment of the referees correlate with which type of footprint in the web of the same paper? Our analysis shows that researchers citing articles assign different weights to content quality criteria than the average referee does; in the field of educational research referees place the greatest weight on the criteria of validity (62percent), but only 38percent on significance. In the subfield 'history and philosophy of education' of educational research only, the number of citations correlates with the overall judgment of the referees. In general, our analysis aims

[31] The Hamburg EERQI group lead by Prof. Dr. Ingrid Gogolin at Hamburg University, Germany, http://www.ingrid-gogolin.eu

[32] The Berlin EERQI group lead by Prof. Dr. Stefan Gradmann, Humboldt University Berlin, Germany http://www.ibi.hu-berlin.de/institut/personen/gradmann; see the article by Gradmann et al. in this volume.

I. Gogolin et al. (eds.), *Assessing Quality in European Educational Research,*
DOI 10.1007/978-3-658-05969-9_10, © Springer Fachmedien Wiesbaden 2014

to support editors of scientific journals to evaluate their present publication policies. Furthermore, we present an outlook on a possible future scenario.

1 Intention and Outline

The EERQI (European Educational Research Quality Indicators)[33] project is unique in that it brought together commercial scientific publishers, and a large group of referees as nominated by different national and international educational research associations, especially by EERA, the *European Education Research Association*. Furthermore, internationally renowned professional research groups in education as well as scientific information management institutions were involved in the process. One vision was to understand the refereeing process. This should be helpful for the development of a future pre-filtering algorithm for selecting scientific documents, in order to warrant forwarding them to peer reviewers for further analysis of their appropriateness for publication.

EERQI followed two parallel strategies: analyzing the refereeing process, and, independently thereof, looking for machine-extractable footprints of documents and their authors on the Internet as possible Quality Indicators. The refereeing process as such was analyzed by the Hamburg group by identifying a large and distinct set of indicators for quality aspects of scientific papers. These aspects were cast into a questionnaire and made quantifiable by asking referees for a rating (from one to seven) for each entry. The results of this procedure are called 'intrinsic' parameter values. The general question was: Which characteristics does the document itself reveal by assessing the inside process of refereeing based on the answers on a questionnaire completed by many referees?

The other approach of EERQI, carried out by the Berlin group, was the collection of indirect information on the quality of a paper by asking about the impact of the paper on the Internet, and about the footprint of the document's author on the Internet as a possible indirect indication of the quality of her/his current paper. The program called *aMeasure* was developed and applied by the Berlin group to those authors whose papers were assessed via the EERQI questionnaire. The program collects the replies to queries sent to Web search engines, as well as document and citation repositories. The resulting counts provide the values of the so-called '*extrinsic'* parameters, that is, the footprints

[33] EERQI *European Educational Research Quality Indicators*; funded by the European Commission; Project Coordination and Management: Ingrid Gogolin, University Hamburg, Germany; Funding period: 1. April 2008 to 31. March 2011 see the chapters 'research results, project description, reports, partners, publications' on the project homepage http://www.eerqi.eu

left by the author on the web, in the outside world. We are grateful to the Hamburg and the Berlin groups for letting us analyze their full amount of data.

Specifically in Section 2, we will present the '*intrinsic'* parameters chosen by the Hamburg group and analyze the wealth of data gained by their survey. In order to extract knowledge from the data, we will, for example, extract the weighted (linear) combination of ratings for specific questions which correlates best to their overall quality judgment on a paper. This is to be read as an indication of how much specific ratings of a paper by a referee (e. g. 'are the methods good?', 'are there significant results?') contribute to his/her overall decision on a paper to be published.

In Section 3, we will study the *extrinsic* parameters chosen by the Berlin group to measure the author and document rating in repositories and web-services.

In Section 4, we outline the importance, concepts and needs of the emerging *author identification services,* necessary to remove the ambiguities of the author-paper identification hampering the interpretation of those extrinsic parameters which rely on the assumed uniqueness of author names.

In Section 5, we perform a quantitative comparison of linearly combined weighted measures derived from the intrinsic parameters with the number of citations which the papers in question have gathered up to now since their (much earlier) initial publication. A time frame of five years was chosen in order to acquire enough data. This may be an underestimate, since especially in a field of humanities like educational research, where citations are generally few and scattered by source, there may be a major share of citations coming up beyond those first years, but it is considered as an early response period by experts in the field. The EERQI's data provides the unique situation that scientific papers, refereed and published years ago, are afterwards (e. g. five years later) re-refereed within EERQI. We can thus compare the results of this refereeing with the past impact of the paper since its original publication, i.e. a test of the refereeing process itself.

In Section 6 we argue that the future application of extrinsic parameter based measures may be used in a scenario of widespread, openly available scientific documents, that is in Open Access.

Section 7 presents an outlook on the future potential for the scientific publishing market arising from emerging technologies. Powerful new services which are much better suited to support the general concept for the document management market can be envisioned. The question how the EERQI tools will fit into this landscape is explored.

The ongoing need of researchers to receive and provide scientific information for and about their research may lead to radically new services for

them. Such services could be developed in cooperation of publishers and specialized commercial or non-commercial technical institutes, and eventually pave the way to a promising new market for a broad range of commercial service providers.

2 Referee questionnaires

One of the important results of the EERQI project is the creation of an extensive, suitable and detailed data set describing facets of the process of refereeing by renowned scientists. The Hamburg EERQI Group designed, organized and managed a detailed questionnaire with 31 items sent to 284 referees to re-review 180 educational research papers as chosen by the EERQI partners. They received a total of 653 referee reports, and from this extracted an enormous amount of quantitative data on the 31 items (intrinsic parameters) from the process of refereeing, a real gold mine of raw meta-information on the complex process of refereeing.

One advantage of this activity was that the referees could judge 'in hindsight', without the stress of actually deciding whether a paper should be published, and thus avoiding possible conflicts with their own scientific work in progress. By chance, this situation is akin to the case of a paper presented to its intended audience before being submitted for review. The activity of the referees thus mimics the case of journals with a prepublication (open access preprint) policy, where experts discuss publications in the process of their development according to the principle 'publish first, referee later'. The EERQI referees were asked by the Hamburg group questionnaire independent of any specific journal profile and policy. Their cumulated data are now exploited here for extracting possibly useful information for editors and publishers. Different concepts and policies for individual scientific journals could be accounted for by performing the data analysis with different weights for the intrinsic parameters. An example: if one wants to launch a journal with a focus on scientific methods, the intrinsic parameter 'scientific method' would be given a larger weight, thus preferring papers containing powerful methods, even if they do not have many results.

Here we will analyze the data and develop a (linearly weighted) combination of those intrinsic parameters as a measure which correlates best with the intrinsic parameter of the 'overall quality' of a paper. This is seen as an indication of the specific aspects, referees consider important for their own overall decision.

2.1. Intrinsic parameter definitions

The EERQI questionnaire of the Hamburg group asked the referees to make judgments (mostly on a scale from 1 to 7, with 7 being the best rating) for a rich set of carefully debated and finally selected topical questions which were assumed to be facets of the nearest to non-explicit aspects of the decision process of a hypothetical referee, or as close as possible to questions which an editor or publisher would like to have answered by referees in order to decide on the acceptance of a paper. The EERQI parameters of a 'good publication' have been discussed at several occasions and were condensed to a final set to represent a general concept of scientific quality.

These parameterized questions were named and shorthanded by capital letters and the results come as values for the following *intrinsic parameters*:

- *Identification entries* such as a *running number* (named A) in the list of papers, *submission date* (B), *language* (C), *article group* (three subgroups of the extremely wide research area of educational research were defined) (D, E), *membership* in one of the subgroups of the research field of *education* such as (1): *assessment*, *evaluation, testing and measurement*; (2): *comparative and inter-/multicultural education*; (3): *history and philosophy of education*; (4): none of these;
- *Review-identifier* (F); *Article identifier* (G);
- *Some information on the referee*: his/her experience as seen by him/herself (most referees answered: 'yes, I am experienced in the respective area of research...');
- *Some textual comments* [results which were appreciated by the referee (AD), suggestions from the referee for the author (AE)];
- Six scales (general topical questions, with several sub-questions) for the referee to rate a paper, named V1 to V6:
- V1: *methods* (H) to (J);
- V2: *results* (K) to (L);
- V3: *discussion* (M) to (P);
- V4: *originality* (Q) to (S);
- V5: *significance* (T) to (W);
- V6: *validity* (X) to (Z);

- (AC): A general question asking for an *overall judgment* ("Comparing this article to an article representing good research, where would you place it from 1 (low) to 7 (excellent) quality?").

Here we will focus on the six topical questions (V1 – V6) answered by the referees with regard to the quality of a given paper. We will not analyze the two to four subgroups of each intrinsic parameter (e. g. the *method* comes in three sub-queries: is it *intelligible* (H), is it *appropriate* (I), is it *accurate* (J)) in depth, and restrict ourselves to averages over all subgroups. Unfolding these averages will become important if a specific journal with a unique profile desires a more detailed analysis.

2.2. Comparison of the intrinsic parameters with the overall referee judgment

We can now analyze which arguments guide referees towards their overall judgment (intrinsic parameter AC) by varying the weights in a linear combination of all intrinsic parameters, to find the best correlation to the overall judgment. This can also serve as a test for methods, which could later be applied to compare 'intrinsic' to 'extrinsic' measures in as much as this can be made work. The parameter AC indicates 'overall judgement' which we interpret here as the 'individual referee's final summary weight' on the scientific value of a paper, as an overall measure for the decision to publish a paper or not.

We can thus analyze which specific criteria influenced the referee's decision by comparing AC to the set of the six individual intrinsic (averaged) parameters. We computed the mean value V1 to V6 for each category of parameters, with equal weight for the registered answers to its 2 to 4 sub-questions. We restrict the present analysis to global *linear measure definitions*, that is, for each referee report *k* on a paper we sum the 6 intrinsic parameters $V_i(k)$, each with an adjustable weight a_i.

$$X_k := \Sigma_{i=1...6}\, a_i \cdot V_i(k).$$

The a_i are globally adjustable coefficients of 'worthwhileness of each intrinsic parameter' as seen by a referee for the aim of defining a global quantitative 'quality'-parameter X_k for the paper k.

We then ask how much of the referees' topical judgments V1,... V6 enter the *overall judgment AC*. For this we searched for a best fit of the sum of weighted V1, .. V6, that is: we minimize the quantity Y by varying the parameters $a_1 \ldots a_6$:

$$Y := \Sigma_{k=1,\,...653}\, |X_k - AC_k| \text{ with the constraint } 1 = \Sigma_{i=1\,...6}\, a_i\,.$$

We choose the absolute value because we want to measure the average deviation of X from the *overall judgment AC*, independent of whether it is above or below.

The best result for the fit of $a_{1...6}$ (given in percent and rounded to full percent) for the respective parameters is:
methods 1percent; *results* 3percent; *discussion* 7percent; *originality* 13percent; *significance* 5percent; *validity* 72percent.

The spread of the distribution is about one unit within the scale from one to seven, which still means a reasonable result. This combination of the topical judgments mimics best the overall decision (AC). That means, the overall judgment of the referees is based mainly (by 72percent) on their opinion of the *validity* of a paper, with a slight admixture of 13percent *originality*. The minimum average 'spread' of the deviations is 0.6 points on the scale of 1 to 7, which confirms the broad spectrum of personal preferences in rating articles by human referees. This individuality of each referee increases the chance of an author to be accepted if authors try to submit their articles repeatedly (to different journals). It also reflects that in educational research, validity is esteemed to be the principal quality criterion. Scientific readers, we assume, think similarly to referees in judging the value of a paper. Thus, the best routine policy of a scientific publisher for the acceptance of their journal would be to focus on valid and original papers, which represent what the referees think that the scientific readers expect. But we will demonstrate in Section 5 that this is not really the case.

Finally, one may infer that launching a new journal with a specific profile, e. g. focusing on *methods*, on *results* and/or on *discussion,* may necessitate additional effort to sharpen the policy, to train the referees, to perform marketing, to invite and encourage authors.

We had assumed that *significance* is the most important criterion. But it seems that there are few significant papers on the market (or in the sample which was included in the project), to begin with. This situation is supported by the present culture of publishing in the scientific community, where the sheer number of published papers seems to be the most profitable factor for the career of an author. Not too few scientists thus tend to divide important results into smaller chunks of marginal scientific progress. Furthermore, it seems to be attractive for many researchers to work with methods and on topics which allow for rather easy publication within a limited amount of time. This version of ‘attractiveness’ for authors, encouraging them to strive for publicizing 'marginal results' in order to achieve publications in a short time, has been addressed most energetically and analyzed by the famous scientist Jean Zinn-Justin (J. Zinn-Justin, 1998), a French theoretical Physicist who, among other relevant experience, was the editor in chief of the ‘Journal de Physique’.

With the large amount of data on intrinsic parameters collected in the EERQI-project and the adequate tools for analyzing them, the next step should be a more in depth analysis of the existing data together with publishers to help them individually to sharpen their specific journal publication policies with regard to the type of papers they intend to include in their journal. Technically this is done by choosing, in accord to their chosen journal policy, the respective weights, that are: the parameter coefficients.

3 What the Web says about authors and their papers

Part of the activities of the EERQI-project was to analyze the 'footprints', i.e. the impact on the Web of the 309 authors who were included in the experiment. The footprints were harvested in the EERQI project by the Berlin group. The basic assumption driving this activity was: a tool which measures various types of the web-presence of an author might help editors or publishers to pre-filter the incoming quantity of articles and/ or provide an indication to the referees of the author's scientific standing. This is based on the hypothesis that there is a correlation between the reviewer's decision before publication and the measurable impact after publication, and that the quality of other papers of an author are in the same quality range, even if authors are changing their field of research from time to time.

We will later compare the usefulness (is there a correlation?) of some extrinsic parameters with the evaluated intrinsic parameters gained by the peer referees from the document itself. Building on the EERQI set of papers, the Berlin group produced a set of measures of the authors' presence on the World Wide Web as a possible indirect indicator for the quality of a particular scientific paper. These measures can be identified automatically (by a machine program). The aim was to create a kind of automated pre-filter for the reviewing process.

There are two possible strategies to achieve these aims. The first possible approach is the design of a filter with the ability to perform semantic text-analysis. This approach was explored by one of the EERQI-partners (see the contribution by Sandor et al. in this volume). In the Social Sciences and Humanities – as represented by educational research – the architecture, rhetoric and style of research publications cover a broad spectrum. This makes them less accessible for automated semantic analysis than the highly standardized form of texts which can be found in the 'hard sciences'. Thus, the Berlin group in EERQI developed a tool for measuring the web-impact of authors and their documents with the hope that these can act as an indirect indicator of an individual paper's quality – the assumption was that there are (more or less strong) correlations

with the intrinsic parameter of overall scientific quality.

The advantage of this assumption is that most researchers and authors leave traces on the web in their professional life which can be identified by their tool called *aMeasure*. It can be applied to Web-repositories, services and archives in order to extract specific values for the web visibility of authors and their documents which potentially may indicate quality. The question addressed here is the following: Which (linear) weighted combination of 'extrinsic parameters' are suitable to serve as a global measure that can be used as a proxy for the identification of potentially good quality papers?

3.1. The extrinsic parameter definitions and data set

Hereinafter, we give some comments on the chosen extrinsic parameters and try to extract some results from the available data harvested by the Berlin group on the papers used as a test set in EERQI and made available to the EERQI consortium. The parameter name in the data of *aMeasure*[34] is given in quotation marks. According to our understanding, the tool harvests information by automated queries from web services and archives and comprises it into the following selected parameters:

- The parameter 'G' provides information on the paper in question itself: G indicates the *number of mentions of the article title in the search engine BING*; data are gathered with the *LexiURL* searcher. *BING* is a relatively new search engine operated by *Microsoft*, with a yet uneven coverage of scientific publications.
- The variable 'D' represents '*author name*', the name of the first author of a paper. 'D' works as anchor for many follow-up queries which check the impact of the author whose name was identified on the web and in repositories.

'D' does not necessarily refer to a real author or person, but to the spelling of an author's name on the paper in question. With the ever-enlarging Web and a growing global population it becomes more and more likely that there are several authors with identical names. These names thus appear when searching in *Google Scholar* and other respective services in a joint display. In many current search engines, citations are retrieved and added to the cumulative result for all

[34] The source code of aMeasure was not available to us and could thus not be applied directly. It is to be linked from the EERQI server on the results page. http://www.eerqi.eu/page/research-results-eerqi-prototype-framework or on some other Open Source server such as SourceForge http://sourceforge.net/

different authors but of the same name. In order to illustrate this we present an example from the data: a paper shows 'David Johnson' as author's name. At first sight, the impact of this person, as retrieved from the web-queries by *aMeasure* are impressive: 996 scientific papers published, 184,139 citations, 584 publication years, 184 citations per paper, and an h-index of 168. It is obvious that the author of this paper is one of a number of real persons who carry the same name. While one of them may be working in educational research in the UK, others may work all around the world – or even at the same University. A search engine like *Google Scholar* combines all of these 'David Johnsons' into one name, effectively mimicking one super-hero, whereas it is the combined power of many different persons with the same name. This problem of *author-name ambiguity* is known with respect to the database of *Google Scholar* and other web services, not least because they are expanding their database, which amplifies the problem.

Some of the parameters of *aMeasure* also depend on the field 'author-name', for example:

- F: how often the *author name* is found by the search engine *BING*, using *LexiURL*;
- I: *papers per author name* gathered by *Google Scholar*;
- J: *citations per author name* and additional information for excluding self-citations as gathered from *Google Scholar*;
- K: *years.* This parameter seems to indicate the number of years in which the author was active as a writer of scientific papers (defined by the Berlin group as: 'year of most recent paper minus year of oldest paper'). Due to the accumulation of different authors of the same name, *aMeasure* produces some very long author biographies, such as a person T.M. (name abbreviated) (7,000 years), U.F. (more than 1,000 years), or a D.N. with 2,700 years of publication.
- L: *citations per year*;
- M: *citations per paper*.

Another element of the tool *aMeasure* is the search in a set of three citation indexes:

- N: *h-index;*
- O: *g-Index*;
- P: *e-index*;

The data show that despite their different definitions and databases used, these three well-known citation indexes correlate nicely, i. e. they produce very similar results.[35]

aMeasure also studies three interesting parameters: *delta-h*, *delta-g, delta-e*, which measure the time in years, until a given index (h, g, e) rises by one unit. This could reflect whether an author has recently gained in popularity. However, we do not make use of these parameters in this study, although it would be interesting to follow the change of these widespread used citation-indexes over time because it could provide information on the de- or increase of an author's impact as a function of time. An interesting phenomenon of the web-age is that, due to the generally increased availability and accessibility of scientific papers online – and the so much easier copy and paste of citations – the h-index (as do the others) of authors increases with time, even if the author ceased publishing long ago. This phenomenon is exacerbated by the fact that aggregators, such as citation repositories, continually expand their databases and include an increasing amount of non-commercial, and also 'grey' literature such as CERN-reports. Hence, the publication indexes h, g, and e increase accordingly. However, for a publisher or editor – in order to get a feeling of the scientific value of an individual paper – a large citation index (summed up over all of his papers) may state that the author is an established scientist, but it must not necessarily imply that the individual paper is better than average, or a substantial contribution to science by any means.

There are prominent examples[36] of revolutionary papers of extremely high scientific content written by an author with no, or negligible citation index measures. One example is a paper (in Physics) proposing the existence of that one essential particle, a Boson, to be the final missing elementary particle needed to predict how particles gain mass. The author, Peter Higgs, has not published much since then or before. His h-index is in the order of 3. But mankind has currently invested a billion dollars to prove this prediction. If the Higgs-Boson did not exist, Physics would have to be changed radically at a fundamental level. In 2012, the existence of the Higgs-Boson was for the first time proven by experiment. In 2013, Peter Higgs and his cooperation partner at CERN, Francois Englert, received the Nobel Price for their prediction. Another leading High-Energy Theoretical Physicist, Gerard t'Hooft from Utrecht – also a Nobel Price-

[35] Definition of the h-index: papers of an author listed according to number of citations, starting with the most cited one; identification of the list number which equals the number of citations: an author X has published h papers which are equally or more often cited than h times. We do not at this point want to go deeper into the often discussed topic of possibilities for cheating via citation indices.

[36] We apologize for taking examples from physics, not educational research, but we are both physicists.

winner – has stopped long ago to let anyone referee his papers, but distributes them exclusively via a public preprint Open Access repository (namely the arXiv) or on his personal website [37].

3.2 Towards a useful filter

These stories point to the fact that a useful pre-filter for the quality of documents will need a suitably weighted combination of many extrinsic parameters, including the number of citations (which is huge in the case of Higgs). The tool *aMeasure*, which is in focus here, includes the following further extrinsic parameters:

- Results from two search engines:
  - Q: *Google hits*; R: *Metager hits*;
- A set of aggregating repositories:
  - S: *citeulike hits* (author-name);
  - T: *libraryThing hits* (author-name);
  - U: *Connotea hits*;[38]
  - V: *Mendeley hits;* [39]
  - W: *Mesur hits.*[40]

The various hit numbers by citation repositories (parameters S to W) depend partially on whether or not the author actively uploaded her/his publications. But most authors do not even know about the respective repository, as exhibited by the data of the *aMeasure* application in EERQI by the Berlin group. Of 93

[37] http://www.staff.science.uu.nl/~hooft101/

[38] *Connotea* was a free online reference management service for scientists, created in 2004 and discontinued in 2013.

[39] *Mendeley* was a web program for managing and sharing research papers and for online research collaboration. It was founded in 2007. Since 2013 it has been owned by Elsevier Inc., which potentially endangers the open access policy of the program.

[40] MESUR: *Metrics from Scholarly Usage of Resources* was a research project based in the *Los Alamos National Laboratory,* USA. The major objective was the development of a toolkit for the assessment of the impact of scholarly communication items with metrics that derive from usage data. The project was funded from 2006 to 2008.

authors tested by *aMeasure*, 48 did not have a paper at *citeUlike*[41], 41 at *LibraryThing*, and none had registered at *Connotea*. The latter was a free online service to manage own and other references[42]. *LibraryThing* is essentially a book reference managing service.

The quality of some of the mentioned citation database services is pretty questionable:
citeUlike does have a huge amount of duplicates. We tested this by using our own publications and found out that one of us has 182 articles cited there, but the first already comes in 14 doublets; moreover, some papers listed there are not written by the same author. Moreover, false citations (e. g. stating a town as the author) are common, notably for exam works such as theses, as well as wrong assignments of authors to texts (the author identification problem). As can be illustrated by the 'survival-rate' of the services – see footnotes 8 to 12 – another weakness of these tools is their lack of sustainability. On the other hand, what can be observed here is a highly dynamic field of development with a great potential for a future usage in processes of quality detection.

We follow this pathway on the basis of our analysis. For the purpose of developing a future tool which can provide substantial basic information about the scientific standing of an author we propose a set of six parameters which each measures the number of hits in various repositories. According to our analysis, the most promising candidates appear to be:

W1: *early citations of the paper* (or one of its preprint versions);

W2: *Google hits of the author*;

W3: *Google Scholar hits*;

W4: *citations without self-citations*;

W5: *citations per year*;

W6: *h-index*.

W1 reflects the resonance of the most relevant international expert colleagues, as is then reflected in W3; W2 measures the web-presence of the author in general, while W5, and thus W6 measure the scientific output of the author in the past. We propose the omission of the parameters of measuring other citation indexes (since the g- and e-index are mostly quite parallel to the h-index value), the

[41] *CiteUlike* is a free citation service, originally designed and served by Richard Cameron. Although not mentioned on its homepage, *CiteUlike* has been incorporated by Springer Publishing in 2008.
[42] *Connotea* discontinued service on March 12, 2013.

output of smaller search engines, and the response of citation databases - the latter because of the insufficient number of authors registered there.

4 Author identification

The usefulness of any tool such as *aMeasure,* resting on extrinsic parameters, depends on the complete and reliable linkage of an author to his/her text. To assure that one collects only data associated with a single author or text, one either needs an author-identification from one of the early existing individual author-identification services such as the nonproprietary *authorclaim*[43], or the commercial *researcher-ID*[44], or one has to wait until the international initiative ORCID becomes fully operational. ORCID[45] is an international initiative with the policy of serving as a unifying umbrella for the emerging multitude of the proprietary or individual author-identification services, that is to collect a copy of their data bases, eliminate duplicates, and identify author names by their papers, registered at different ID-services, which is possible whenever there is at least one paper in common. The ORCID initiative has been joined by a very broad spectrum of publishers, service providers, institutions, research institutes and by academic organizations. The intended services of ORCID are still being developed, and were in too early a state for the EERQI project. In the meantime, the ORCID author identification service is operational; however, it is not yet coupled with the existing author-identification services such as *Google-scholar* or *authorclaim,* and thus pretty incomplete. In order to illustrate this, we again refer to a self-experiment: for all the scientific papers by one of us, the author-registries yield the following results: ORCID (16), authorclaim (86), Researcher-ID (35), inSPIRE (8), arXiv (8), Microsoft Academic Search (86), *Google--scholar* (163). Only the data of *authorclaim* are free of duplicates, false titles or false authorship assignments. Only *authorclaim*, arXiv and inSPIRE are author-endorsed. There is still a lot of developmental work to be carried out – and by the way: the authors of this paper have themselves been members of the technical committee of ORCID from the very beginning.
The scope of author-identification services is to serve a registry that relates the scientific publications to their authors, identifying them unequivocally and

[43] The only non-commercial Author Identification Service *authorclaim,* developed and served by Thomas Krichel, Long Island University, USA: http://www.authorclaim.org
[44] *ResearcherID*; author-identification service http://www.researcherid.com
[45] ORCID *Open Researcher and Contributor ID* http://www.orcid.org, an open Initiative to unite and map the multitude of the emerging distributed author-identification services.

asking them for their approval to be included in the respective database (endorsement). The subtle but important differences to a passport-identification style of author-identification have been analyzed in detail elsewhere (E.Hilf et al. 2008; T.Severiens,2008).

We expect ORCID to cover a larger part of the scientific documents in the years to come, after the data of other author-identification services will have been incorporated, and after more authors become aware of the author-identification problem and register to one of the services. It is then that the second step of an EERQI-like analysis can be undertaken: looking for the web-footprint of individual authors and thus fully exploiting the true strength of a program like *aMeasure*, and its successors respectively.

5 Comparing intrinsic with extrinsic measures

On the basis of our considerations and relying on the available data, we settled on a two-step strategy:

1. For a given paper we chose those two extrinsic parameters which measure the number of citations in *Google scholar* and the number of hits for the paper's title in *BING* (data partially provided by *aMeasure*). Some bias is certain in this approach, since the papers in the EERQI set are from different years. Anyhow, the two extrinsic parameters chosen are free of the author-name problem. All chosen papers of the EERQI document data set are about five years old. All of them had already been accepted for publication by a respective journal. Several of them were also published in institutional repositories. They met a wide variety of refereeing standards.

2. We compared these two extrinsic parameters to various combinations of intrinsic parameters of the same paper, searching for the best correlation. If there is one, then the extrinsic parameters could be of some use as a pre-filter.

Our task here is mathematically interesting, since the intrinsic weighting of articles involves six parameters, while there are only two extrinsic parameters for now. We look for the best linear combination of extrinsic parameters with the best correlation to an optimized linear combination of intrinsic parameters. If there was one, it would tell us which intrinsic criteria count for a footprint on the web. In order to support publishers or editors in their decision of whether a paper

is eligible or not, one needs an algorithm with a sensible set of weights for these parameters. The simplest case would be a linear combination of the parameters. It would then be the task of an analysis to find the optimum weights of the parameters. If there was a 'true' measure of a set of papers where extrinsic parameters have been measured and a final 'yes/no' judgment has been extracted, this could be used to perform a cluster analysis in order to find the optimum mixture of intrinsic parameters. However, such a true extrinsic measure of a set of papers does not exist. Instead we are given a set of extrinsic parameters, which are also in need of an optimum weighting. Given a linear weighting, we could likewise perform an 'extrinsic cluster analysis', if a true decision of what is intrinsically acceptable were available. But again: it is not.

The solution could be to take a set of papers which have been accepted (or which have been chosen by us to act as 'accepted papers'), then optimize the weights for the intrinsic and the extrinsic parameter sets separately by respective cluster analysis. Here, without training with a set of 'accepted' documents, we try to fit on 'moving ground': that is to take both sets of parameters, intrinsic and extrinsic, and vary them simultaneously to find an optimum with regard to a defined 'discrepancy distance'. For this we would pick the absolute difference between the global intrinsic, and the global extrinsic decision variable, summed over a set of papers each calculated by adding the respective parameters with their calculated weights.

But to come to a useful pre-filter of the quality of documents by using machine generated extrinsic parameters, we suggest that the potentially powerful program *aMeasure* should be adapted under consideration of the following principles:

- use only repositories where author-identification is executed;
- use only repositories where de-duplication of papers is in effect;
- restrict to papers endorsed by the author (e.g. by the homepage/publication list) and extracted by a machine program (C. Schöne, 2013);
- count the number of papers there as an extrinsic parameter.

Clearly for this, we have to wait at least until ORCID is in full operation with a rich database, and until services such as publication-list-analysis, and de-duplication of paper references at repositories are available. Thus, we turned to the next best option:

5.1 Intrinsic parameters versus Google-Scholar number of citations of a paper

We analyzed each paper for the 'number of citations in *Google scholar' (*entries by winter 2012*).* We then decided to count each pair of (referee, paper) separately as independent judgments, thus a paper refereed by three different referees enters the statistics three times. We then casted seven citation-categories of impact as measured by *Google Scholar* using the same principle as before: to have approximately equal numbers of papers in each category. We ended up with the following schema:

Table 1: Categories for citations numbers per paper.

Number of Citations	Category	Number of papers in this Category
>20	7	92
13-20	6	105
7-12	5	96
4-6	4	82
3	3	59
1-2	2	138
0	1	70
Self-citations are not subtracted.		

We then compared the citation-category of a paper with the intrinsic parameter of *overall judgment* of the referees as collected by the answers from the questionnaire sent to the referees. For this we looked for the best fit of the contributions of the six intrinsic parameters (*methods*, *results*, *discussion*,

originality, *significance*, *validity*) by looking for the minimum sum of the absolute values of the difference to the citation-category. The result we get is that the best relative contributions of the intrinsic parameters to correlate to citation numbers are *methods*: 38percent, *results*: 0percent, *discussion*: 0percent, *originality*: 20percent, *significance*: 18percent, *validity*: 24percent.

This sounds interesting, because in essence it tells us that other true experts in the field who found a paper useful and cited it, look more for methods, originality, significance and validity than for results and discussion. This is in some contrast to the overall weighting of the referees who seemed to focus almost on validity alone. The true value of such an analysis of the data could be useful to support publishers who want to design specific scientific journal profiles and who could then define this by giving the intrinsic parameters preferred weights. The analysis here shows for the best fit a large deviation from the average by 1 category unit (trivial, in that it says, these papers had been already accepted five years ago, and thus should not rank too bad here), and a pretty flat minimum, that of almost two category-points (1.94) of the absolute value of the difference of *Google Scholar* minus best mixture of intrinsic parameters.

An equal weight to all intrinsic parameters would give 1.99 as average deviation. Treating each intrinsic parameter separately as an assumed stand-alone parameter, only the intrinsic *results*-parameter appears not to exercise much influence on the colleagues (2.5). Apparently those who cite a paper look for methods which are new to them, and then proceed to present their own results. This is in sharp contrast to the intrinsic internal fit, which stressed the validity of the results and does not put a large weight on the methods, as seen by the referees. In other words: The correlation of the decision of experts to cite a paper, and its esteem as rated by the referees is pretty small here (Table 2). This is a disturbing finding as it shows that scientific refereeing is not too much correlated to the later impact of the paper on the web.

Table 2: Number of papers which were rated 'overall value' (A) by the referees versus their number of citations (B) as extracted from *Google Scholar*:

	A = 7	6	5	4	3	2	1
B = 7	10	26	31	14	8	4	0
6	12	25	29	25	7	5	1

5	4	20	31	21	9	8	3
4	4	19	26	20	8	3	2
3	3	13	17	11	11	3	0
2	5	19	21	13	6	4	4
1	6	23	42	31	22	13	5

The rating correlates somewhat with the number of citations earned, which means the referees are able to 'anticipate' the scientific future value of a paper. Most often papers (which were already published) are rated as moderate (5), for any number of citations. Papers cited never or just once dominate for any rating, except for the highest ratings. This may reflect that the scientific value of some papers is only appreciated after some years have passed; the early citations may not always reflect the future scientific value of a paper. But the disturbing finding is that the correlation of the overall rating of the referees (A) and the (future) citations of a paper is small.

5.2 Intrinsic parameters versus BING number of citations of the paper

The value G of *aMeasure* delivers the number of 'Web mentions of an article *title* in search engine *BING*, that is the number of URLs of pages matching the query submitted, data gathered with *LexiURL searcher*' . This is a valuable and easy to measure web-footprint of any paper. Some data for *BING* were gained by the Berlin group in 2011 using the *aMeasure* software. We decided however to redo and complete the Bing data (in March 2012) to have them collected in about the same time as the *Google Scholar* data given. Such web-footprints of an article grow with time: the retrieval services (*Google Scholar* and *BING*) increase their databases, which will result in a growing number of citations delivered by them. In parallel, authors get an easier entrée to open access publications over time, resulting in a growing number of citations per article. Thus it would be very interesting to repeat the study as a function of time in a future project.

In contrast to *Google Scholar*, which provides data about the number of citation counts in their database, *BING* is a more general search engine which collects data from the web, and we counted here the number of mentions of the article's full title occurring anywhere, not just as citations in another scientific paper. With the extrinsic parameters *Google Scholar-citations* (X) and *BING-hits* (Y) we can now check which mixture of these two is best suited to fit any mixture of intrinsic parameters. The original dream of EERQI can thus be tested in a nutshell, with two extrinsic parameters, and the set of six intrinsic parameters. Some examples of our research outcomes are:

The 155 papers in our data set fall into three groups, as decided by the referees with regard to their content-type, let them be noted as

A. 63 papers in *educational assessment, evaluation, testing and measurement*;

B. 42 papers in *comparative and inter-/multicultural education*;

C. 50 papers in *history and philosophy of education*.

Questions that can be answered now, are:

1. What is the best (linear) combination of the *BING*-rating and the *Google Scholar* rating to mimic the referee ratings for the papers in these groups? We get for the *BING*-part 100percent for A, 67percent for B, and 33percent for C respectively. The interpretation would be that mostly only group C has a sizeable number of papers where the scientific citation as found by *Google Scholar* counts more than just a more general web presence.

2. Are English language papers better found in *BING*? Yes, by 17percent. Are the German language paper ratings closer to the *BING* ones than the English ones? Yes, the respective alignment is 100percent for the German language ones and 72percent for the English ones, almost independent of which intrinsic parameter is chosen.

3. Which of the six intrinsic parameters alone is more reflected by *BING* as compared to *Google Scholar*-citations? *methods* 84percent; *results* 100percent; *discussion* 92percent; *originality* 95percent; *significance* 100percent; *validity* 100percent.

The interpretation of such numbers is yet something different. Our interpretation is that *methods* are of more value for researchers to be cited, because they want to use them for their own research. It would have been best to have a final

'yes/no'-decision from the referees on the publication of each paper. Lacking this we could still ask: which combination of the two extrinsic parameters would lead to a 'best decision profile' for a given admixture of the six intrinsic parameters?

In order to approach an answer we calculated

$$WW := \Sigma s_{k=1...654} \; |\Sigma_{i=1...6} (a_i \cdot V_{ik}) - \Sigma_{j=1...2} (b_j \cdot W_{jk})|$$

and varied the a_i (admixture of intrinsic), b_j (admixture of the two extrinsic) parameters to get WW to at least a wide local minimum. The result for group A is a rather shallow minimum:
methods: 0percent, *results*: 0percent, discussion: 0percent, *originality*: 0percent, *significance*: 38percent, *validity*: 62percent. This may reflect that *BING* mirrors a somewhat broader web presence than just citations in other scientific papers, and that *significance* and *validity* are the best guidance of the intrinsic parameters to estimate a future success in *BING*, that is the Web in general.

6 Application of quality measures to already published documents

As the results of the extrinsic parameters show, the citation rate depends crucially on the digital visibility of a publication. Currently however, most of the research output is being published in subscription journals without even an author's copy in his/her own institutional repository. But the fraction of parallel open access published articles ("OA-green") is growing, which gives services like *Google-Scholar* and *BING* a growing relevance in the academic publication process. Also, at present, many publishers create new OA-journals or publish some papers Open Access in a toll-access journal. This ("OA-gold") is growing rapidly, but at present comes with a broad spectrum of conditions and business models. Due to rising prices, publications in subscription journals are losing their relevance for the exchange of scientific knowledge. Currently, most publishers do not allow OA-green without restrictions. This means that they are forbidding the author to re-use her/his own presentation of a research outcome. Many articles are more or less 'de-published', by printing them in very expensive subscription journals, which only a few may read. In order to give authors more rights as the owners of intellectual property, and to save them from buyout contracts, European copyright laws need to be adapted to the digital age. The ENCES[46] association is one of several European initiatives which tries to

[46] *European Network for Copyright in Support of Education and Science* (ENCES) e.V. http://www.ences.eu

influence this legal process for education and research. For the purpose of developing more precise and automatic measurements of quality of scientific documents, a representative (large) fraction of the publications in any field in a digital and openly readable form will be necessary to train the respective tools.
Currently (2013), European legislation is heading in the opposite direction, as it is being discussed to establish text- and data-mining (*tdm*) as a unique form of use, separate from reading. Until now, *tdm* is and was considered to be a form of use equal to reading. While in contrary to this development, there will be a open research-data pilot in Horizon 2020 programme.

7 Future Strategies and Services

The EERQI study focused on using up to date tools to improve services that intend to support the traditional scientific publication process. Furthermore, it looked for intelligent combinations of tools which support the decision process in selecting scientific documents worth publishing. Now let us address the task beginning with its general requirements.
'Eternally' stable general requirements of science for the Information Management of scientific information may be seen as:

- maximal distribution; no barrier access by any scientist in the world;
- easy re-use of information e.g. via download and subsequent reprocessing (e.g. numerical data or mathematical formulas);
- long term availability and readability;easy access for the community for discussion, further information etc.

A sustainable (business) model is required in order to strive towards achieving these goals by exploiting the available techniques to the fullest, and continuing to design innovative services incorporating new and upcoming technical means suited to serving the goals. The field for the design of innovative service concepts is rich. For example:

- open access to scientific documents;
- multiple storage of copies across the world;
- long term archiving in open formats by public libraries at multiple locations;
- connection of documents to their full information (data measured, information collected data bases, mathematical formulas, etc.);
- embedding, that is, connections to supporting information such as to

reviews from the field, or to related information in other papers. This means both: embedding by linking to earlier publications (backward embedding) and to later ones (forward embedding). And it may even include methods for keeping old texts intelligible by explaining outdated notations etc. The latter must be kept in mind if scientific papers are to remain understandable by future generations;

- community tools for the experts in that specific field;
- abstracting services for quick and easy information of the experts in the field[47];
- inclusion of interactive, dynamic, living documents;
- online open author communities (wiki-type);
- first refereeing in an OA publishing mode, discussing on the web, and deciding on final publishing (long term version and availability) later; this allows competing referee-services on the same paper such as common in the scientific awards scene;
- inclusion of new types of documents: snippets, blogs, remarks, discussion pieces, et cetera.

It is most probable that the future role and market of referees will change – including the now important role of refereed contributions to journals which were central to the EERQI project. The current procedures might well be complemented by other ways of communicating and commenting scientific findings and their 'quality' more directly. Still, even if the analyzed tools with their intrinsic or extrinsic parameters faded from the market, EERQI's general idea of inventing tools for text mining and their intelligent combination will remain. An example for probable future usage is 'trend-scouting'[48], the machine-supported search for emerging new fields of research in a large set of new publications of different types and sources, refereed or not.

8 Closing remarks

The data of EERQI are unique and innovative in their richness of information: a detailed response from referees on how they see quality aspects of a large set of scientific papers; and the results from a tool used to harvest the footprints of authors as a possible indirect indication of the quality of a paper. Within the EERQI we were in charge of the general technical support (of the server etc.).

[47] As for example provided by http://www.papercore.org

[48] e.g. the studies for a tool *e-scout* of the *Institute for Science Networking* http://www.isn-oldenburg.de

For us, it was a challenge extracting as much insight and information of interest to readers and publishers as possible from the large amount of data from both sources. The large amount of detailed information from referees about their ways of evaluating a paper could lend scientific publishers a hand in refining their policy definitions and their decision processes of accepting papers to be published. The wide range of opinions from different referees on the same paper, as could be inferred from the data, will add to the caution – and the need for a policy – on how to decide. A tool for measuring extrinsic parameters in detail could give the editor or publisher a varied and independent piece of information, mainly focusing on the author's past standing, visibility, and impact. Thus, a tool in the spirit of *aMeasure* could be helpful for the identification of an author's general, visible scientific profile. Anyhow, this would not 'measure' the scientific quality of any given unpublished paper. Moreover, unless we can make use of author-identification systems like ORCID, the potential to be a valuable tool, not only for editors, but for hiring committees granting scientific positions, etc., remains fairly limited.
Important results from analyzing the data of the EERQI peer review-questionnaire and of queries to web-engines, including early citations of the EERQI-set of papers, are:

- experts in the field who found a paper useful and cited it, look more for the indicators pooled under the headings 'methods', 'originality', 'significance' and 'validity' than for 'results' and 'discussion'. This is in some contrast to the overall weighting by the referees who seemed to focus almost on validity alone;
- The correlation of the decision of experts to cite a paper, and its esteem as rated by the referees is pretty small. Scientific refereeing in education research does not seem too strongly correlated to the later impact of the paper, as it can be identified on the web.
- Only for the sector of papers from the area of 'history and philosophy' did we find a stronger correlation to citation numbers than to a general web presence.
- Although all papers chosen had been accepted by a refereeing process of a truly existent scientific journal about five years prior to our analysis, the rate of rejected papers in the EERQI-experiment were considerable, and independent of the measured number of citations.

Here, the primary research subjects were (trained) humans - referees and authors with their wide variety of publication habits, abilities to perform and write. These habits are not stable over time. New technical tools and services appear,

new generations of scientists grow up, and new policies are enacted. The necessity to judge on a paper however, rests on human decisions which are not solely, but for considerable parts embedded in traditions. There is no imaginable way to avoid this problem in a 'scientific market' that relies on the traditional concept for refereeing *before* publication. .

A possible solution in this situation may be to make full use of the new techniques for maximizing distribution, availability, and re-use of scientific information. From our point of view, a promising future development should start with the principle: *publish first, review later.* This principle could easily be realized (and has been in some cases[49]) by the publishers if they allow for a preceding OA discussion time for a paper prior to the refereeing and publishing process ("gold-OA"). In this format, the authors retain their copyright for the versions they open to scientific discussion. It is only in a subsequent step that a publisher or editor of a journal may identify a paper as eligible for publication in his journal, and then forward it to a peer review process. If a paper is rejected by the referees at this stage, this will not harm the author's scientific output; but selection for publication in a journal then functions as an additional award for the author, and serves the public with a marker for quality and relevance.

We argue in favour of this concept because we assume that it is less susceptible to misuse than the current procedures are (E.Hilf 2001; IUPAP 2001). Moreover, it can allow delivering information about the potential quality of a paper to the scientific community without temporal delay. It will furthermore allow an author to send the paper to multiple publishers or journals in parallel. The publication in a journal will then function in a similar way as the application for an award. This in turn may change the publication market by increasing competition. Moreover, the concept may optimize possibilities of long-term archiving because it would open up the freedom to store the document in multiple archives across the globe. Long-term archiving could be bolstered and become independent of the 'survival' of a publisher or journal.

Summarizing, the emerging concept in the digital age would be: publish online first, copyright stays with the author, multiple storage of copies abroad, open access, competing journals picking papers after discussion.

[49] The journal *Physical Review* (by APS *American Physical Society*), the most prestigious journal in Physics, uses the preprint server arXiv. It is recommended that the author posts a preprint in arXiv before the application for publication with its start of a refereeing process in the conventional sense.

9 Acknowledgements

Special thanks go to Ingrid Gogolin for her relentless interest, motivation, and patience during the joint work. Information from Stefan Gradmann on data collected with his program a*Measure*, and from Verena Stumm on data from the inquiries of the large set of referees, is much appreciated.

For thorough proofreading including the English language we are grateful to Christian Schöne.

10 References

Hilf, Eberhard R., Kappenberg, Bernd and Roosendaal, Hans E, (2008): Author Identification: The benefit of being able to identify researchers uniquely; volume 5, page 5-8, 2008; The Euroscientist http://www.euroscience.org/author-identification,28115,en.html

Hilf, Eberhard R. (2003): Report on the IUPAP Workshop on Scientific Misconduct and the role of Physics Journals in its investigation and Prevention; at: EPS European Physical Society, Mulhouse, FR http://www.isn-oldenburg.de/~hilf/vortraege/london03/report2EPS.pdf

Hilf, Eberhard R. (2003): Report on the IUPAP Workshop on Scientific Misconduct and the role of Physics in its Investigation and Prevention; Oldenburg; http://www.isn-oldenburg.de/~hilf/vortraege/london03/london03-isntalk-print.html

IUPAP (2003): Outcomes: International Guidelines for Ethical Conduct in Scientific Publishing. International Union of Pure and Applied Physics; http://www.iupap.org/wg/communications/ethics/outcomes.pdf

Schöne, Christian, Bernhardt, Eike et al.(2013): Analysis tools for extracting citations from publication lists; in progress

Severiens, Thomas (2008): Requirements for Author Registries. The Euroscientist, volume. 5, 2008, pages 4 – 5; http://www.euroscience.org/author-identification,28115,en.html

Zinn-Justin, Jean (1998); Peer review and electronic publishing 1997; In: The Impact of Electronic Publishing on the Academic Community, Session 3; The content and quality of academic communication, Peer review and ele tronic publishing; International Workshop; Academia Europaea and the Wenner-Gren Foundation; Wenner-Gren Center, Stockholm April 1997; Wenner Gren International Series V. 73 (Portland Press 1998); Butterworth, I. (Ed.); http://www.portlandpress.com/pp/books/online/tiepac/session3/ch3.html

Guidelines for Transfer of the EERQI Prototype Framework to other Social and Economic Sciences and Humanities

Angela Vorndran

Short Summary
The tools constituting the EERQI framework were developed within the research field of educational science: A peer review exercise was applied involving educational scientists and document evaluation procedures for educational research texts. To enable adaptation of the framework to different disciplinary contexts, a transferability exercise was part of the work. This chapter examines the possibilities of transferring the EERQI framework to the research field of political science taking into consideration the similarities and differences in publication cultures of both fields and developing guidelines for the transfer.

1 Transferability Testing

The EERQI project developed an evidence-based prototype framework for the detection of quality in educational research publications. Approaching the question of quality assessment with a mixed methodology of qualitative and quantitative techniques, the framework comprises different aspects of research publications and national, disciplinary and publication type-related particularities. Based on an introspection of the specialties of educational research publications and the field´s publication culture, a tool was designed for this research field. The instrument allows for integrating different characteristics of documents and their reception in the quality assessment of a publication by developing a calibration of certain indicators with their assigned weightings. As a follow-up on this research a more general application of the framework shall be envisioned by testing the transferability of the methodology to another research field from social sciences, economics or humanities: For this study we chose the field of political science.

As the framework in itself is a flexible construction which can be adjusted to different use cases and application scenarios, its transferability to another research field is an important aspect of its design. The quality indicators which

I. Gogolin et al. (eds.), *Assessing Quality in European Educational Research,*
DOI 10.1007/978-3-658-05969-9_11,

are part of the prototype framework have been developed for educational science but are in themselves largely independent of disciplinary specialties. Only their calibration creates a discipline-tailored instrument. Thus the transferability of the EERQI framework to other social science and humanities fields is expected to be unproblematic.

2 Publication assessment in educational science and political science

The preconditions for testing the transferability of the EERQI framework to another social science field are nested in the publication cultures, assessment procedures and field-related practices of both research fields, i.e. educational science and political science. Publication cultures in the social sciences in Europe are in many ways similar. Political science and educational science share the characteristics that they are chiefly nationally oriented i.e. national research communities interacting strongly within relatively closed circles and publications in national European languages are common (cf. Hicks 1999; Nederhof 2006; Norris & Crewe 1997).

Publications in political science as well as educational science cover a wide variety of publication types. In contrast to the strong focus on journal articles in the natural sciences, the social sciences and humanities in general publish a great share of their research in books or book chapters. This state of publication cultures in the social sciences and humanities also affects the means of assessment of publications. One of the most challenging problems is the overestimation of citation counts in research evaluation. The most common instrument for these analyses, the Thomson Reuters Web of Science (WoS), formerly the citation databases of the Institute of Scientific Information (the ISI databases), primarily indexes a selection of peer reviewed journals, whereas other forms of publications such as conference proceedings are indexed to a limited extent and book only recently and to a very limited extent has started to be indexed. The huge shares of research publications in the social sciences which are not published in journals covered by the WoS databases and in other forms of publications such as anthologies and monographs are therefore neglected in a large number of quantitative assessments of research productivity and impact based on WoS data.

The distribution of publication types in European educational and political science research has not been analyzed to any greater extent, especially not by using data based on other sources than the WoS databases. A German study conducted for educational science showed a dissemination of publication types of 49% for book chapters, 33% for journal articles, 15% for books and 5% for others (Dees 2008). In a survey of German political scientists, Faas and Schmitt-

Beck (2008) showed that articles in peer reviewed journals are considered as the most influential publication type, followed by monographs. Editing volumes, authoring book chapters and writing articles in non-peer reviewed journals are considered relatively equal, albeit on a lower level than the monographs and peer reviewed journal articles; and grey literature gained the lowest reputation, although scientists also considered the grey literature underestimated in assessments based on publication statistics. The participating researchers themselves published on average 2.8 monographs, 3.3 edited volumes, 7.6 book chapters, 4.6 peer reviewed journal articles, 5.8 non-peer reviewed journal articles and 6 other publications.

Regarding the international context, Huang and Chang (2008) analyse the distribution of publication types at the University of Hong Kong in the categories of 'Politics and Public Administration' and 'Education', showing a distribution of 60% and 37% respectively for journal articles, 3% and 8% for books and monographs, 37% and 35% for book chapters and a distribution of 0% and 20% for conference and working papers. For Australian universities´ publications, Butler (2006) reports a distribution of books for 'Politics and policy' and 'Education' of 5.8% resp. 2.5%, book chapters 37.3% resp. 19.3%, journal articles 46.1% resp. 54.5%, conference papers 10.8% resp. 23.6%. This shows that not only peer reviewed journal articles which are solely indexed in Web of Science, but also other publication types play an important role in political science and educational science but are not represented by WoS indicators.

Using publications indexed in WoS as a basis for analysis, Katz (1999) states that the shares of papers from political science/public administration and education account for 0.6% resp. 0.5% of the total content of the ISI/WoS databases from 1981 to 1998. This reflects the strong focus on other research fields than the humanities and the social sciences, in favour of medicine and the natural sciences, in the ISI/WoS databases. Although the distributions of publication types in the two research fields, as well as the different contexts of the studies presented here, do not show a uniform picture, it becomes clear that journal articles are not the only medium for research communication, but other publication types also play an important role.

The appraisal of references in the bibliographies of WoS-indexed content in the two research fields also shows the importance of sources not indexed by WoS. In political and educational science, 70-80% of the references cite texts published in books and other non-journal material (van Leeuwen 2006). In a similar analysis Moed (2005) shows that only 42% of the references in WoS education journals are citing journal articles while the remainder refers to other document types. For political science, the proportion of references to journal articles is even lower at 32%. In a comparison between a number of European

countries and Harvard University, Plümper (2003) analysed political science as represented in the WoS databases. The results show that Harvard not only surpasses all European countries in terms of the total number of publications, but also in terms of shares of publications in highly ranked journals. This gives an indication of the fact that the WoS databases are not representative for European publications in political science.

In the Journal Citation Report (JCR) of 2010, published by Web of Science, 255 educational research journals are listed in the categories "Education & Educational Research", "Education, Special", and "Psychology, Educational"; whereas the corresponding number for political science is 174 journals in the categories "Political Science" and "Public Administration". To compare the two fields, analyses were made using a number of indicators offered by WoS/JCR: total cites, Impact Factor, 5-year Impact Factor, Immediacy Index, Cited Half-life, Eigenfactor Score and Article Influence Score (cf. tables 1.1 and 1.2). Within each category of indicator, the results for the top 20 educational and political science journals respectively were compared. In general, the results show substantial similarities between the two research fields. Small differences can be observed, but they are minor in comparison to e.g. big science journals like Nature and Science.

Table 1.1 - Indicators derived from Journal Citation Index 2010

JCR 2010	Impact Factor P	Impact Factor E	Total cites P	Total Cites E	Immediacy Index P	Immediacy Index E	5-year IF P	5-year IF E
Mean	2.08	2.69	2359.5	3644.4	0.832	1.107	2.89	3.84
median	1.94	2.51	1663	2393.5	0.579	0.986	2.44	3.47

P=Political Science, E=Educational Science, IF=Impact Factor

The numbers displayed in table 1.1 show that educational science journals have a higher Impact Factor (0.6 points on average) than political science journals (cf. 36.104 for Nature). Also, the 5-year Impact Factor, taking into account citations five years after publication instead of the two years counted for the regular Impact Factor, shows the same tendency. Total cites also describe an advantage of educational journals over political science journals: This is most evident in the first rank where total cites of "Child development" (19231) more than double cites of "American Political Science Review" (7459) for 2010 (cf. 511,248 for Nature). Nevertheless it should be noted that "Child development" represents an outlier in terms of total cites for educational research journals. Excluding that

journal, the mean values for total cites in educational research and political science journals are quite similar. Also in regard to the rapidness of articles being cited, the numbers for educational science journals are slightly higher on average as well as in all 20 journals considered.

Additional indicators calculated by WoS are presented in table 1.2 and show higher numbers for political science journals. In case of the Eigenfactor, an indicator measuring how many times a journal giving a citation is cited itself, slightly higher measures for educational science in the top ranks can be perceived but the average numbers show that political science journals receive higher numbers overall. However, the numbers in both research fields are generally quite low (cf. Nature 1.74). This is probably due to the fact that educational science and political science research largely interacts within the field, where few journals with high impact factors exist, which would increase Eigenfactor values. The Article Influence Score, a measure derived from the Eigenfactor, is the only measure where political science surpasses educational science in all top 20 journals. It describes the influence of an article over a five year period, taking into consideration the Eigenfactor and the number of publications in the journal.

Table 1.2 - Additional Indicators derived from Journal Citation Index 2010

JCR 2010	Eigenfactor P	Eigenfactor E	Article Influence Score P	Article Influence Score E
Mean	0.00874	0.00626	2.146	1.604
median	0.00711	0.00453	1.864	1.529

P=Political Science, E=Educational Science, IF=Impact Factor

Both educational science and political science show a characteristic typical of the social sciences and the humanities: the Cited Half-life is very high. All top 20 journals in both research fields show a Cited Half-life of more than 10 years, which proves that the reception process here extends over many years; and half of the citations to articles in a journal in educational research or political science are being made later than 10 years after its publication date.

When comparing these measures based on citation counts in WoS we have to bare in mind that citation counts very often are skewed distributions which do not necessarily allow for dependable judgments based on averages. Taking into consideration the median values, too, is one way to get a clearer picture of the real distribution of citations among the journals analysed.

An overview of countries and publication languages represented in the Journal Citation Report 2010 (JCR) in the two research fields of educational science and political science shows a predominance of Anglo-American journals and an even stronger one of English-language journals (cf. tables 2.1, 2.2, 3.1

and 3.2). However, in recent years, the inclusion of journals from other countries has strongly increased so that presently, a relatively large number of Spanish-language journals, as well as other countries and languages are included. Both fields show a similar coverage of journals in the JCR in terms of national and language coverage: circa 75% of the journals originate from the US and England; and more than 85% of the journals are in English. In comparison to the 2006 JCR, where 89% journals in the educational research categories were from the US or England, and 95% of the journals were in English, there is a clear development towards internationalisation of the journals indexed in the WoS and JCR databases. Apart from the US and England, there are journals from 21 different countries listed in the educational science categories; and in the political science categories, journals from 24 different countries in the 2010 JCR edition. In both fields, the shares of journals of European providence, excluding British journals, are nearly 15%. Considering European national language publications apart from English, the coverage amounts to 9.4% for educational science and 7.5% for political science.

Table 2.1 - National providence of political science journals in JCR 2010

Country	**Number of Journals**	**Shares (%)**
United States of America	70	40.2
England	60	34.5
Netherlands	5	2.9
Germany	4	2.3
Australia	3	1.7
Canada	3	1.7
France	3	1.7
Norway	3	1.7
Austria	2	1.1
Chile	2	1.1
Mexico	2	1.1
Romania	2	1.1
Spain	2	1.1
Brazil	1	0.6
China	1	0.6
Colombia	1	0.6
Czech Republic	1	0.6
Hungary	1	0.6
New Zealand	1	0.6
Philippines	1	0.6
Russia	1	0.6

Slovenia	1	0.6
South Africa	1	0.6
Taiwan	1	0.6
Turkey	1	0.6
Venezuela	1	0.6
Total	**174**	

Table 2.2 - Language of political science journals in JCR 2010

Language	**Number of Journals**	**Shares (%)**
English	149	85.6
Spanish	8	4.6
German	6	3.4
Multilingual	5	2.9
French	3	1.7
Hungarian	1	0.6
Portuguese	1	0.6
Turkish	1	0.6
Total	**174**	

Journal Citation Report, Political Science Journals 2010 (JCR Social Science, Categories: Political Science, Public Administration)

Table 3.1 - National providence of educational research journals in JCR 2010

Country	**Number of Journals**	**Shares (%)**
United States of America	123	48.2
England	71	27.8
Netherlands	10	3.9
Spain	9	3.5
Australia	8	3.1
Germany	6	2.35
Turkey	5	1.9
New Zealand	3	1.2
South Africa	3	1.2
Brazil	2	0.8
Portugal	2	0.8
South Korea	2	0.8
Belgium	1	0.4
Croatia	1	0.4
Italy	1	0.4
Japan	1	0.4
Lithuania	1	0.4
Mexico	1	0.4

Nigeria	1	0.4
Philippines	1	0.4
Poland	1	0.4
Russia	1	0.4
Slovenia	1	0.4
Total	**255**	

Table 3.2 - Language of educational research journals in JCR 2010

Language	**Number of Journals**	**Shares (%)**
English	224	87.8
Spanish	10	3.9
German	5	1.9
Multilingual	4	1.6
Turkish	4	1.6
Portuguese	3	1.2
Croatian	1	0,5
Italian	1	0,5
Japanese	1	0,5
Russian	1	0,5
Slovenian	1	0,5
Total	**255**	

Journal Citation Report, Educational Science Journals 2010 (JCR Social Sciences, Categories: Education & Educational Research; Education, Special; Psychology, Educational)

This overview of publication cultures and reference and citation practices in educational research and political science shows great similarities between the two fields, as well as with other fields in the humanities and the social sciences. The similarities are not only extended to the publication and reference and citation practices per se, but also to issues of coverage in terms of publication types, languages and geographic origin of journals in traditional citation databases, having an effect on the usability of these indexes as data sources for quantitative research assessments through the use of bibliometric indicators. Thus, the solutions suggested in the EERQI framework tackle a general problem, encountered by most social sciences and humanities, and should therefore be likely to be valid for other fields within the humanities and the social sciences. The fact that the EERQI framework shows a high degree of flexibility and adaptability underlines this notion.

3 Methodology for determining the transferability of the EERQI Prototype Framework to political science

The EERQI Prototype Framework in itself was designed by making use of two different approaches to assessing publication quality. On one hand, the peer review exercise provided guidelines used in expert judgment and indications on peer review evaluation criteria. On the other hand, metric indicators of documents were collected including citation data. These went beyond traditional citation measures by also including a variety of measures derived from the analysis of web mentions and other types of usage data. These two approaches were analysed together to find a correspondence of so-called intrinsic and extrinsic quality indicators. The extrinsic indicators should be combined and weighted in correspondence to those judgments made by the peer reviewers.

The methodology applied to determining the transferability of the EERQI Prototype Framework to political science was developed by reproducing various parts of the EERQI project activities and was designed to follow the steps listed below:

1. Set up a separate database to store political science documents
2. Adapt crawler to political science
3. Train classifier for political science
4. Collect data for extrinsic indicators
5. Evaluate extrinsic indicator data and generate extrinsic determined sequences
6. Apply calibration result from educational science to the data obtained in 5.
7. If necessary: adapt calibration formula weights
8. Verify results according to the procedure applied for educational science i.e. present two documents each which are rated best, worst and medium to evaluators to see if the judgment can be reproduced
9. Compare the verification results obtained from educational science and political science
10. Develop guidelines for transfer

These steps should guide the transferability testing procedures of the EERQI framework to political science. In addition, the process of appraising the extrinsic document characteristics might lead to adjustments in indicator weightings to take into account the different publication cultures of the research fields.

3.1 Adaption of EERQI tools to political science

To ensure the functionality of the tools being part of the EERQI framework for an application in political science, some adjustments had to be made in the context of technologies developed especially for the disciplinary conditions of educational research. As a proof-of-concept example, adjustments were only carried out for the German language.

A separate section of the EERQI database was created for the storage of political science documents harvested from the Internet.

The crawler, constructed to harvest documents from the World Wide Web to be stored in the EERQI database, works with underlying search terms and seed URLs leading to sources of relevant documents. These had to be adapted to political science terminology and relevant document sources. Appropriate political science terminology was assembled from thesauri and other sources and websites holding scientific political science content were selected. The crawler was adapted to political science using these resources.

The classifier supporting the crawler in the selection of relevant documents for the research field, which represented an integral part of aMeasure by automatically identifying documents pertaining to the field of educational science, had to be trained for political science, too. Accordingly, more than 150 German-language documents were evaluated by EERQI members and rated as "yes" (pertaining to political science) or "no" (not pertaining to political science). In addition, documents from other research fields were used as negative examples for the classifier training. By extracting word shingles, i.e. groups of three consecutive letters in the text, underlying text characteristics could be identified. Each group of field-specific documents could be identified automatically by rules derived from these characteristics and thus the classifier was trained on content in political science documents. Through this procedure the classifier was enabled to discern political science documents from documents from other research fields. On one hand, this serves as an additional way of restricting web documents crawled for the content base to those actually belonging to political science, on the other, it was aimed at supporting aMeasure in collecting author-based measures in a more reliable way by identifying authors by their disciplinary affiliation.

3.2 Procedures applied to examining possible transfer

In the peer review process for educational science, evaluations of publications were conducted to define quality ratings of a selected number of research docu-

ments. These ratings were based on quality indicators developed within the EERQI project: rigour, originality and significance. To combine these intrinsic indicators with extrinsic indicators collected in the project, a calibration of metric quality indicators representing the judgments made by the expert peer review was targeted. The metric values considered for this assessment were:

- Papers per author
- Citations per author gathered from Google Scholar data
- Citations per year
- Citations per paper
- H-index
- G-Index
- E-Index
- Google-Hits (author)
- MetaGer-Hits (author)
- CiteULike-Hits (author)
- LibraryThing-Hits (author)
- Connotea-Hits (author)

The calibration procedure should combine these extrinsic document measures and weight them individually to resemble peer review judgments. In the case of educational science, the testing procedures did not lead to a useful calibration due to an insufficient data base so that the transfer of these results was not possible.

However, in the procedures for transferability testing, the goal remained to appraise the metric indicators for this research field and investigate if adjustments would be necessary to account for special characteristics of political science as a research field and its publication culture. As a basis for testing the transferability of the EERQI framework, 36 political science research articles were collected. These articles were selected exemplarily from research journals and web resources in German language. For these documents the extrinsic indicators from Google Scholar data, search engines and the social bookmarking services CiteULike, LibraryThing and Connotea were collected using aMeasure.

4 Results of transferability testing for political science

A comparison of publication cultures in educational science and political science

was conducted to identify publication structures and referencing behaviour; and the effect these behaviours have on the possibility of employing metrics based indicators. The similarities and differences of disciplinary habits have to be reflected in the weighting of the various extrinsic quality indicators. In order to achieve this, the document collections established for testing purposes in educational science and political science were analysed and compared.

The 36 selected political science documents are authored by 44 authors in total. Most of the indicators calculated with aMeasure refer to author-based measures. Looking at the data, known problems with author name disambiguation seem likely in this set as well[50]. As data provided by Google Scholar were used for most calculations, well-known shortcomings of these data could not be entirely resolved (cf. Jacsó 2010). The average number of papers per author is high, at 70.1 (mdn=43), and with a very high standard deviation of 72.1. It seems likely that in cases where one author is assigned a very high number of papers, it also includes papers by other authors with the same name and initials. This also applies to the number of citations calculated here with a mean of 857 (mdn=185.5) and a standard deviation of 1335. The other measures take these numbers as a basis when calculating citations per paper and year, and h-, g- and e-indexes. It has to be taken into consideration that looking at the median values and their differences to the mean values it shows that the distribution of papers per author attributed by Google Scholar is very skewed. Some authors are assigned a very large number of papers which accordingly accumulate even higher numbers of citations whereas other authors only amass few papers. These numbers can be influenced by Google Scholar´s techniques of data gathering which might not cover all sources for publication records, some authors might not list all their publications in sources indexed by the search engine. On the other hand, publications authored by other persons might be included because of data processing problems. As a consequence the standard deviation of the calculated measures also exceeds expectations because of these problems. Some strong outliers of authors with more than 800 papers could be observed in the sample.

An alternative indicator used was internet popularity measures. Using search engines author names were searched and hits were counted as ‘web mentions’. This analysis showed that most hits were retrieved by Google with an average of 296 (mdn=178) web mentions per author. The very high standard deviation of 348 shows large differences to which different authors are indexed in Google. In comparison to Google, MetaGer search engine hits were calculated as well. Here, only 11 (mdn=3) web mentions were retrieved per author with a standard devia-

[50] As reported in other project documentation and corresponding literature (e.g. Ruths and Zamal 2010), the unambiguous identification of author names in web resources and most notably in Google Scholar is difficult and the problem cannot be fully resolved.

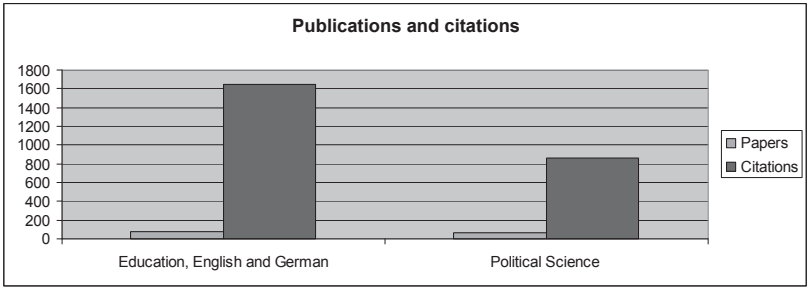
tion of 17. The results even included 19 authors with no web mentions at all, representing a distribution common to popularity measures where lots of counts are gathered by single units whereas many others remain without any counts at all.

Social bookmarking tools served as the third type of sources for data analysis. The authors were most frequently encountered in the web service CiteULike with 28 bookmarks on average (mdn=3.5, SD=55). LibraryThing contained only two bookmarks per author on average (mdn=0, SD=0) and lacks data for 19 out of 36 authors. Connotea does not contain bookmarks for any of the selected authors.

The exercise conducted for the 297 documents exemplarily selected for testing purposes in educational science included 307 authors of documents in English and German. The average number of papers per author is 79 (mdn=40, SD=120). Citations amount to 1650 (mdn= 187) per author with a standard deviation of 10.951 which again raises doubt about the reliability of these measures. The results from search engine queries show an average of 136 web mentions per author in Google (mdn= 0, SD=361) and 3.7 in MetaGer (mdn=0, SD=14). Results from searching social bookmarking tools amount to 13 in CiteULike (mdn=0, SD=48), 17 in LibraryThing (mdn=0, SD= 69) and no hits in Connotea either.

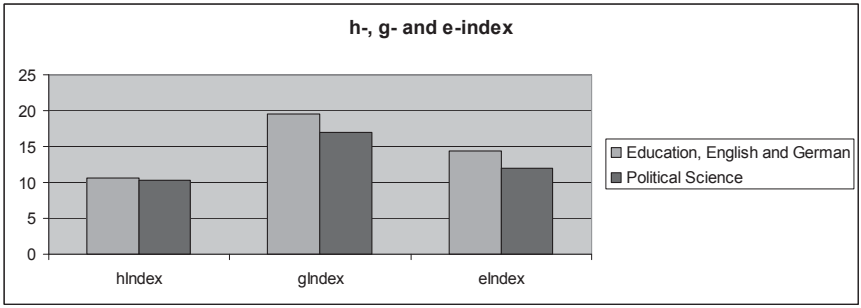
Comparing the two document collections, it is evident that mean publications per author in educational science and political science lie at a comparable level of 79 (mdn=40) resp. 70 (mdn=43). This convergence might indicate that, even when containing some strong outliers, the numbers might be reliable to a degree that they allow for a comparison between the different research fields, although reliability in general is difficult to determine. The number of citations differs strongly with educational science authors amassing almost twice as many citations as political science authors on average (1650 resp. 857) (cf. fig. 1).

Figure 1 - Average publications and citations per author in educational science and political science



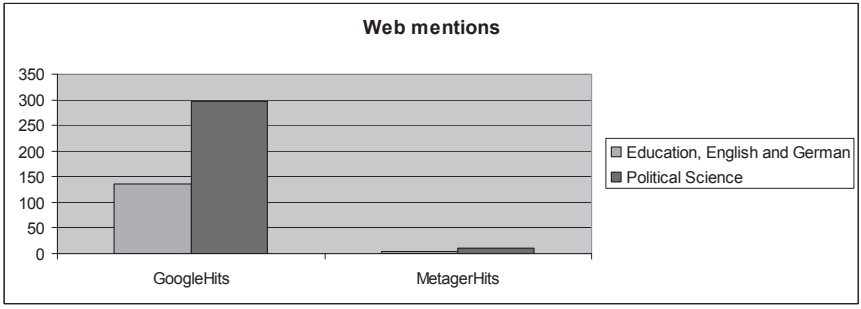
Nonetheless the indicators derived from citation counts converge to a high degree. H-index accounts for 10.6 resp.10.2 for educational science authors and political science authors. g-index amounts to 20 resp. 17 and e-index is 14 resp. 12 (cf. fig. 2).

Figure 2 - h-, g- and e-Index of selected authors



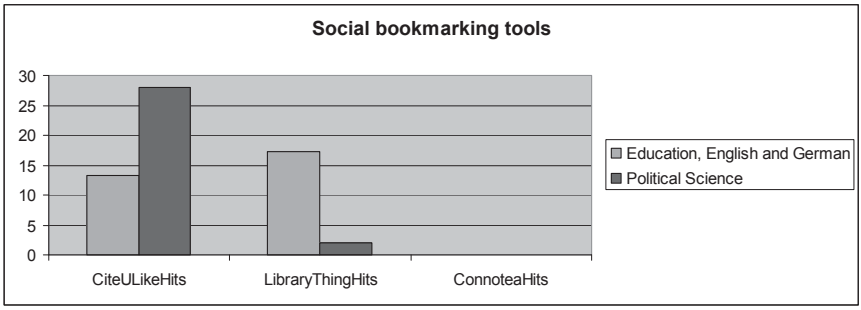
The measures derived from search engine hits show different results. Google web mentions are decidedly lower for educational science (136) than for political science (296), the same is true for MetaGer hits (3.7 resp. 11) (cf. fig. 3).

Figure 3 - Web mentions of authors in educational science and political science



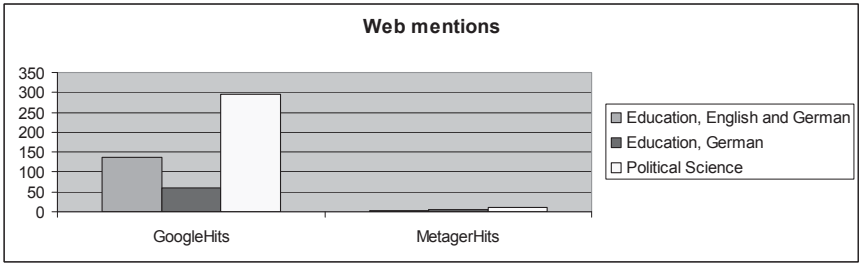
The evaluation of bookmarks for the authors analysed also gives mixed results. In educational science, CiteULike lists significantly lower numbers of bookmarks for the selected authors than in political science (13 resp. 28) whereas in LibraryThing the relations are reversed and educational science surpasses political science in average bookmarkings of authors (17 resp. 2) (cf. fig. 4).

Figure 4 - Average numbers of bookmarks in social bookmarking tools



It is conspicuous that the share of authors not indexed in web-based tools differs strongly. Educational research authors with no web mentions in Google amount to 54% whereas all political science authors gather at least one Google web mention. A similar picture evolves for MetaGer, where 85% of the educational research authors, and 43% of the political scientists receive no web mentions. The same thing can be found in the analysis of the social bookmarking tools, where in both CiteULike and Library Thing, three quarters of the educational research authors in the document set are not bookmarked at all, whereas the corresponding numbers for political science authors are 20% in CiteULike and 43% in LibraryThing. On the other hand, LibraryThing, is more strongly focused on monograph literature which seems to appeal more to educational scientists although book publications are popular in both fields.

Figure 5 - Web mentions of authors in educational science (English and German documents), educational science (German documents) and political science



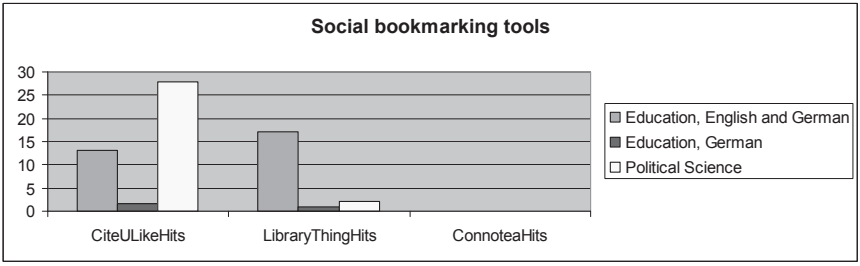
When limiting the set of educational science documents to German language documents to increase the comparability with the German-language document set in political science, the set sizes amount to 143 resp. 44 authors. The tenden-

cies described above for the complete set of educational research authors are even more apparent here.

Compared to 296 average Google mentions in political science, educational science now decreases from 136 to 60. MetaGer web mentions increase from 3.7 to a mean of 6 probably due to the fact that MetaGer is a Germany-based search engine (cf. fig.5).

Bookmarks for German authors in social bookmarking tools are considerably less frequent than before with only 1.7 instead of 13 bookmarks per author in CiteULike and 0.8 instead of 17 bookmarks per author in LibraryThing (cf. fig. 6). A majority of authors is not bookmarked at all in CiteULike (92%) or LibraryThing (96%). The rates of 28 and 2 bookmarks per author for political science are not very high either, but they clearly show a stronger interest of political science readers in web-based tools.

Figure 6 - Average mentions of authors in social bookmarking tools for educational science (English and German documents), educational science (German documents) and political science



It thus becomes clear that using the measures attained with the first version of the tools which are part of the EERQI framework, the results have to be interpreted carefully. Underlying problems like author name disambiguation could not be fully resolved. In a future version, where these problems are solved, the citation counts are likely to be much more reliable. Furthermore, the document sets were not selected systematically but partly aleatoric or following compliance; and because of the differing sizes and allocation to subdisciplines cannot be claimed as a representative sample. Conclusions which can be drawn from the two data sets suggest a direction rather than being reliable measures but can point at interesting parallels and differences in publication cultures in the two research fields. The studies summarized above show similar publication cultures in both fields. Nevertheless those studies primarily focus on print publications and citation analyses of journal articles. The measures developed within the

EERQI framework go beyond these measures and also try to derive quality indicators from web-based tools. Compared to the WoS databases, the number of papers and citations retrieved from Google Scholar suggests an advantage for educational science of using Google Scholar, since publications from the field are more strongly represented online and citations can also be detected to a higher degree. The results from other web-based resources indicate a stronger tendency of political science authors to present themselves and their research online, visible in the higher number of web mentions. Also, political science researchers are more visible in web-based social bookmarking tools as can be seen from the frequency of bookmarks in the tools considered.

5 Guidelines for Transfer of EERQI Prototype Framework to other research fields

The EERQI framework in itself is very flexible and targeted at different usage scenarios. It can be adapted to various situations where quality assessment of research publications plays an important role. It includes numerous indicators based on quantitative analyses of document characteristics which comprise different aspects of a document and its reception in the research community. In the research framework, each of the indicators can be endowed with special weights resembling the importance of each individual document characteristic to document assessment. In general, the indicators do not show a high specificity for educational science and thus facilitate transfer to other research fields. Nevertheless, the framework has to be adjusted to the documents assessed, the research culture of the respective field and publication habits. Also intra-disciplinary differences in these respects might have to be considered. Some guidelines for this procedure are given here.

Although the actual process of transferability could not be conducted to full extent, some important features of the process have been identified and play a role in transferring the framework to another research field. It should be taken into consideration how the publication culture of the field affects the availability of data; and to what extent a certain characteristic of a document have any meaning when quantified into an indicator. Dependent on the aim of publication assessment, each category can be assigned different weights and thus represent major or minor parts of the complete assessment. As the comparison of extrinsic document characteristics of educational science and political science shows, publication cultures and research cultures can differ among fields. The analysis showed that in this sample, political science researchers are more likely to be present on the web and visible in the online bookmarking tools. The weighting of

these web-based components in the indicator framework might thus be more appropriate than in educational science, as more data are available and more reliable calculations are likely.

Another aspect of quality assessment which should be taken into consideration is the different perception of research quality in different fields. For example, while the reception period of a research publication in social sciences is generally very long, in the natural sciences the rapidness of publication after conducting the research is more important in quality assessment.

In general, differences in publication cultures which should be considered when adapting the EERQI framework to another research field can be listed as follows:

- Some fields are more strongly oriented towards English-language publications in journals, thus a lot more data will be available from large indexing services like Web of Science and Scopus.
- The willingness to publish open access affects some measures e.g. usage statistics.
- The extent to which information about a publication can be found online relates to measures based on online mentions.
- The extent to which the research community makes use of online tools relates to the evaluation of online bookmarking tools.
- Discipline-specific or intra-disciplinary focus on research quality might differ.

6 References

Butler, Linda (2006): RQF Pilot Study Project – History and Political Science Methodology for Citation Analysis. http://www.chass.org.au/papers/pdf/PAP20061102LB.pdf [07.02.2011].

Dees, Werner (2008) : Innovative Scientometric Methods for a Continuous Monitoring of Research Activities in Educational Science. In: Kretschmer, Hildrun et al. (Eds.) (2008): Proceedings of WIS 2008, Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting.
http://www.collnet.de/Berlin-2008/DeesWIS2008ism.pdf [07.02.2011].

Faas, Thorsten & Schmitt-Beck, Rüdiger (2008): Die PVS und die deutsche Politikwissenschaft. Kurzbericht zur Umfrage unter den Mitgliedern der DVPW. DVPW-Rundbrief, 139, 166-176.

Hicks, Diana (1999): The difficulty of achieving full coverage of international social science literature and the bibliometric consequences. Scientometrics, 44(2), 193-215.

Huang, Mu-hsuan & Chang, Yu-wei (2008): Characteristics of research output in social sciences and humanities. From a research evaluation perspective. Journal of the American Society for Information Science and Technology, 59(11), 1819-1828.

Jacsó, Péter (2010): Metadata mega mess in Google Scholar. Online Information Review, 34(1), 175-191.

Katz, J. Sylvain (1999):. Bibliometric indicators and the social sciences. Report prepared for UK Economic and Social Research Council. http://dlist.sir.arizona.edu/94/ [14.05.2008].

Moed, Henk F. (2005): Citation Analysis in Research Evaluation. Heidelberg: Springer.

Nederhof, Anthony J. (2006): Bibliometric monitoring of research performance in the Social Sciences and the Humanities. A Review. Scientometrics, 66(1), 81-100.

Norris, Pippa & Crewe, Ivor (1997): Towards a more cosmopolitan political science? European Journal of Political Research. 31(1-2), 17-34.

Plümper, Thomas (2003): Politikwissenschaft in internationalen Fachzeitschriften, 1990-2002. Eine bibliometrischer Analyse der Veröffentlichungsleistung deutscher politikwissenschaftlicher Fachbereiche und Institute. Politische Vierteljahresschrift, 44(4), 529-544.

Ruths, Derek & Zamal, Faiyaz A. (2010): A Method for the Automated, Reliable Retrieval of Publication-Citation Records. PLoS One, 5(8), e12133.

Van Leeuwen, Thed (2006): The application of bibliometric analyses in the evaluation of social science research. Who benefits from it, and why it is still feasible. Scientometrics, 66(1), 133-154.

The relevance of the EERQI framework in the light of future perspectives: Enhancing the visibility and detection of European research publications

Alexander Botte

Short Summary
Against the background of an integrative view on the contributions of the EERQI Framework to support the visibility and detection of quality research publications in the web, some current approaches and developments in the field of web research are presented and discussed. In the light of probable trends in the scholarly publication and communication culture, these approaches are estimated in terms of promising information infrastructures for the enhancement and assessment of educational research.

1 Introduction

The desire for a reliable assessment of research quality is challenged by the fact that available concepts of quality comprise very different aspects and no "official" delineation exists[51]. Nevertheless, a range of indicators normally used in peer review processes are broadly accepted. The dispute starts when criteria (like *originality* or *significance*) shall be quantified and measured in scalable metrics. In this situation, other criteria like *impact* might be considered as there are better chances to quantify impact and to develop tools which can support this measurement on the basis of standardized algorithms[52].

In fact the view at quality is different in different situations where there is a demand for quality ranking. As a starting point for coping with this situation, EERQI chose a variable, situation-based concept of 'quality' in the EERQI framework – exemplified by so-called use cases, specific application contexts.

Secondly, we have to realize that certain procedures of assessing quality are meanwhile established – and they are so for good reasons. Peer review as the

[51] The fundamental discussion of scientific quality dates back at least to the 1980s, e.g.: Marcel Chotkowski La Follette, 1982. or: Weingart/Winterhager, 1984.
52 An interesting view on the relationship between the three concepts in psychology publications is offered by Haslam/Laham, 2010. 216–220.
A German contribution to the same topic: Ilg/Boothe, 2010

I. Gogolin et al. (eds.), *Assessing Quality in European Educational Research,*
DOI 10.1007/978-3-658-05969-9_12,

most common tool of assessment offers the advantage of using different indicators of performance which can be selected according to the respective target of assessment. In certain ways, bibliometric indicators work in competition to peer review, but they can also be integrated into the peer review process ("informed peer review") which is increasingly done.

EERQI developed and tested intrinsic and extrinsic indicators. The evaluation results showed that intrinsic and extrinsic assessments accentuate different aspects of a research document. Their assessment is complementary, as they shed light on different aspects of quality. On high aggregation level, they converge.

Looking at the situation of European educational researchers and their publications from a perspective of international awareness, the problem of quality is first of all a problem of visibility. Highly respected scholars in their own country do not find estimation abroad when they only publish in their native language and in national books or journals. Multilingual Europe does not avail itself of a multilingual access to research publications in the social sciences.

In view of these interdependent conditions underlying the lack of awareness for Europe's multilingual publication culture, EERQI tried to tackle the problems of research visibility, quality and evaluation in a modular framework of approaches and tools. This framework offers very practical solutions on quite different levels, which make considerable use of digital techniques.

The aim of this article is to offer an integrative view on the contributions of the EERQI Framework to the delineation of characteristics/indicators of quality and procedures for enhancing the visibility and detection of quality research publications in the web, and an estimation of the results from a perspective of near- future developments of the educational research publication culture.

2 The reference of the EERQI framework to practical evaluation needs

As denoted above, the view at the quality of publications is different in different scenarios or use cases. Generally, two major (pars pro toto) types of use cases can be separated:

2.1. Support and improvement of the peer review process

If a new paper is to be reviewed there is no problem of visibility and no indicators of reception or popularity are available. The article needs to be read thoroughly and compared to other publications by an expert reviewer. If many papers

have to be reviewed by many reviewers, this is a time-consuming procedure that can hardly be standardized. The EERQI framework supports the peer review procedure

- by a peer review questionnaire which helps reviewers to operate on the basis of standardized review indicators and
- by a specifically trained semantic tool which highlights key phrases in order to make the reading process more efficient and comparable by doubling the intellectual process faced by every single peer reviewer through an automated standardized process.

The idea behind both tools is that they can work corrective on subjective limitations which are often criticized in peer review processes.

2.2. Selection and ranking of a high number of specific publications

For the evaluation of universities or big research institutes, for comparisons on the level of countries, entire disciplines or sub-disciplines, a huge amount of publications has to be examined. Only in some cases the collection of relevant publications is known from the start, if relevant publication lists or comprehensive databases are available. The non-availability of a complete set of relevant publications is generally the case if educational researchers or administrators are looking for a representative and up-to-date insight into topically or geographically defined publication areas.

In these use cases we meet at first the need for reliable selection and then for a possible ranking of documents. The EERQI framework provides new approaches to both tasks, which can be shown by the following application scenario:

- A targeted search in the EERQI search engine delivers a high number of relevant research documents in three or four languages.
- The sorting of retrieved documents (number of query hits) can be done on the basis of classical term frequency ranking or on the basis of an aggregated counting of web receptions as well as results of traditional bibliometric instruments (aMeasure).
- The classical term frequency sorting can also be enhanced by semantic pre-selection (highlighted key sentences).
- The highlighted key sentences can be used to facilitate scanning (reading) of personally selected documents.

This quite practical scenario illustrates the benefits of EERQI to promote the visibility of European educational research publications. In fact, the integrative functionality and usability of the EERQI instruments have not yet reached a stage that would deliver a publishable, comprehensive product: . There is still a way to go regarding technical and expansion development. However, the shape of prototypes suggests that such a scenario might be implemented within a developmental project. Such instruments will gain relevance in the future as the need for research assessment will grow in the area of innovative and internationalized governance of science and education!

The EERQI approach on extrinsic indicators is based on the conviction that bibliometric and webometric indicators are not sufficient as a stand-alone instrument of assessment, but that they are important as correctives and as possible indicators for ranking web documents. Metrics of citations, referencing (links), and usage are appropriate to indicate different levels of awareness of publications, authors, institutes and networks in the scientific community. They reflect unspecified components of quality in a modified form. They have made their way into many scholarly assessment scenarios, however vague and debatable their validity may be. It is in the interest of European social sciences researchers that the coverage of bibliometric instruments will be enlarged and optimized.

It is a major concern of this article to show that the future will provide for good conditions to better the coverage of digital information technologies, grounded in the observation that research publications and the corresponding referencing services will more and more be available in the internet.

3 The change of publication culture and scholarly communication

Scientific publication culture and scholarly communication have changed impressively in the digital age, especially since the emergence of the internet as the omnipresent medium, and this process has undoubtedly not come to an end. In the context of EERQI, the following tendencies of scientific publishing are interesting.

(1) The impressive growth of the internet is a truism. The number of internet sources and their corresponding contents increases continuously: publisher servers53, university repositories54, and of course Google Scholar and Google books. The question is if these sources provide a critical mass of relevant content

53 Statistics from the Association of American Publishers show that in 2011, e-books saw a 153% growth again in July, compared to the same time last year. http://paidcontent.org/article/419-new-stats-e-book-revs-up-153-over-last-year-digital-audio-growing-too/

54 See http://repositories.webometrics.info/toprep.asp

which can be relied on if representative observation and assessment are intended. After all, the fast-growing amount of digitally available content has not yet reached a stage where – as many wish to say– "everything is in the internet". Crucial analyses have repeatedly shown that a lot of important content still cannot be found e.g. in Google Scholar, at least not in full-text format and with broader open accessibility55. On the other hand, there is some evidence that the absence of scientific publications in Google Scholar56 is simply a consequence of the absence of digital availability in the internet, meaning that the growing effort of publishers to present their publications (also) in digital format will gradually change this situation.

Parallel, the venue for immediate and simultaneous scholarly interchange about projects and results is no longer confined to conferences, but can also take place in the internet. New sources of scholarly communication have developed and they gradually gain importance: internet conferences, scientific forums and social bookmarking services and so on. Even though the latter forms of scientific communication do not play a major role in most disciplines, they complement the prospective that scholarly publishing and communication goes digital and will take place more and more in the internet.

(2) For several years, this trend has strongly been supported by scientific politics and administration, as there is a clear and definite willingness to foster the internet as a scientific communication medium. Official political statements and corresponding funding activities on European as well as on national level which advance the digitization and free access to scientific online publications and data are ubiquitous. Elaborated expectations of e-infrastructures "that radically transform the process of scientific and engineering research" 57 form the fundamental strategy of these efforts. The call for e-science is consistently connected with a call for "open access", which is strongly supported by scientific administration. Meanwhile, there is not only a general move of publishers towards online publishing, but also into the direction of open access, mostly following the green road (parallel publication with time lag) but in some disciplines, even the golden road is preferred (mostly author pays model).

For peer reviewed journals, a Scandinavian group of authors analyzed the situation in 2009 and found out that a remarkable proportion of 20.4% of all

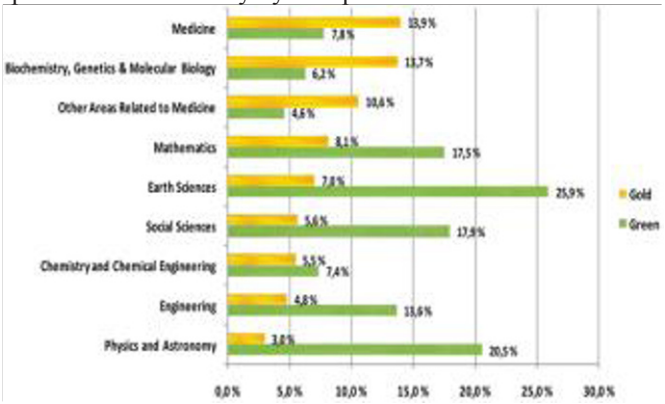
55 Especially, many field specific investigations have been carried out on the coverage of Google Scholar, which show that important national publications and journals are not represented. For the field of education a recent German study is available with further references: Leinenkugel/Dees/Rittberger, 2011: 160-170

56 GS does not publish information about its selection procedure, which is often criticized as "the secrecy of coverage" (English Wikipedia). The above-mentioned German study reveals the interdependence between absence in GS and insufficient presence in the internet.

57 Work Programme 2012. Capacities 2011: 5

journals were already available in open access format, but with big differences among disciplines. Surprisingly, the Social Sciences were in the upper middle field with 23.5%.

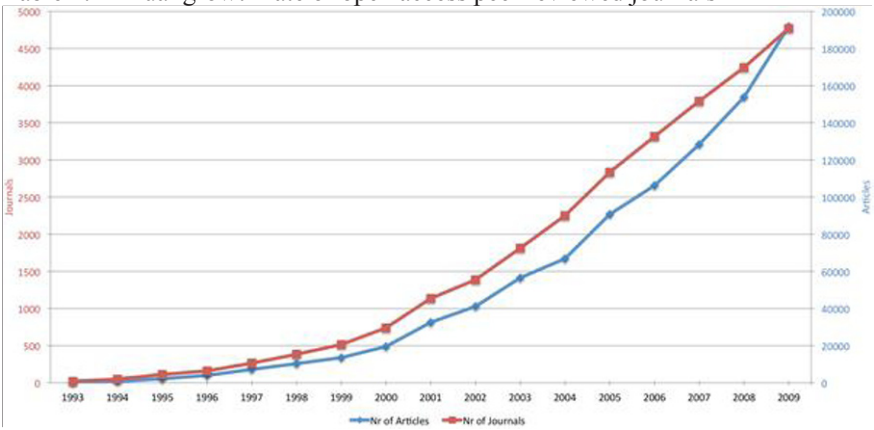
Table 1: Open access availability by discipline 200958



The situation in 2009 goes back to a remarkable 10-year period of development. "Since the year 2000, the average annual growth rate has been 18% for the number of journals and 30% for the number of articles."59 In the 1990s, open access was pioneering work.

[58] Cf all data: Björk B-C. et al., 2010
[59] Laakso M. et al., 2011

Table 2: Annual growth rate of open access peer reviewed journals



This growing availability of valuable open access data will generally enhance scholarly web communication and also better the conditions for semantic applications which are connected with the terms "Web3.0" or "Semantic Web"60. If the necessary standardization (basically the broader implementation of RDF-formatting) succeeds, the structured analyzis of web content will be much more powerful.

(3) Scholarly work in European countries is affected by the Europeanization of nearly all public aspects of life. Europeanization goes with a growing contingent of international cooperation and collaboration. This increases the need for supra-national communication regarding scientific developments, planning and results. Even though Europe supports the challenge of multi-lingual communication, the internationalization of scholarly exchange seems to coincide with an immanent drive towards English as the single Lingua Franca. Generally speaking, small language areas move faster in the direction of English publications than the French, German or Spanish speaking communities (remarkable exception of a large language community moving towards English is China). In spite of these strong traditions of national language publishing there is and will be a growing number of publications (mostly in English, but not only) indexed by traditional bibliometric instruments. This means the *Web of Science* will remain a first stage assessment tool for international reputation.

On the other hand, as long as some countries continue to publish outstanding educational research in national outlets and languages, there will be a need for

[60] Cf. Fensel, Dieter et al., 2011

translations, multilingual instruments to support comparative observations and international awareness of national specifications (topics, approaches).

4 The future of science metrics

The above outlined trends of the scholarly publication and communication culture lead to new approaches of measurement, nearly all of which are based on internet services. A popular new term comprising these approaches is "webometrics". The web offers different sources and indicators to assess scholarly activity; the problem is that most of these sources (web services) were not primarily meant to assess scholarly activity, but to communicate observations, opinions or connections. Others were installed to count very specific usage. As indicators of assessment of scientific productivity, they often seem to lack coverage, soundness, standardization.

Nevertheless, having in mind the limitations of traditional bibliometric instruments, proponents of webometrics vote for a new definition of bibliometrics, that includes metrics of web citations, bookmarking and linking, and usage. Two of the most promising approaches in webometrics will be inspected more closely here.

(1) A quite prominent line of webometrics makes use of the collections of major internet players, or rather Google, in order to apply classical citation analysis. Especially *Google Scholar*, but also *Google Books* are used to build new bibliometric tools.61. A service like *Publish or Perish*62 by Harzing is frequently used, even though its limitations are well known. Recently, Google Scholar published a similar service called *Google Scholar Citations* which might improve standardization (name identification is still difficult), as user feedback can be the basis of an optimization process.

Comparisons of the results of web-based bibliometrics with traditional tools like WoS or Scopus show that the overall effort to clean web-based data is still very high and can only be applied in a scientific project scenario. Nevertheless, the results are promising and underline the suggestion that traditional bibliometrics and webometrics are complementary and should be used in combination.

> *"Our study indicates that there are substantial numbers of citations to academic books from Google Books and Google Scholar and hence it may be possible to use these potential sources to help evaluate research in book-oriented disciplines. Most notably, the possibility to locate cited references*

61 Kousha/Thelwall/Rezaie, 2011
62 http://www.harzing.com/pop.htm

in many academic books through Google Books provides new opportunities to assess citations from books to books (but see the limitations below) that were not traceable before through traditional article-based citation indexes (e.g., WoS and Scopus). Due to the relatively moderate overlap between Google Scholar and Scopus citations, a combination of the two is recommended rather than just one of them."[63]

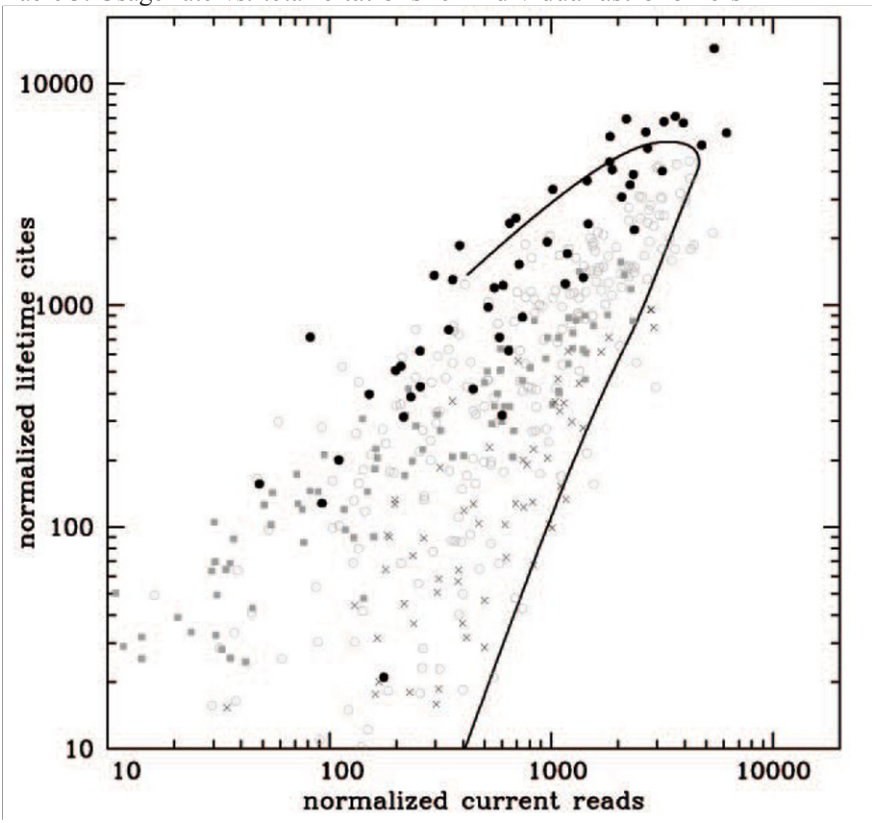
(2) Kuntz/Bollen64 (2010) review the vast amount of research on the very recent phenomenon of usage metrics. Institutional and disciplinary repositories, publishers, database services and library catalogues provide a huge amount of data on the usage (download, lending, acquisition) of scientific publications. Usage data extend the coverage of traditional bibliometrics not only with respect to other publication types, but also a lot more types of reference (downloads, links, web mentions, citations) can be measured in order to develop indicators. Usage metrics have been applied and evaluated on nearly all targets of bibliometric measurement: Papers, authors, institutions, countries, scientific networks…

"Citation rates and usage rates: The two measurements, although both related to the usefulness of an article, have very different properties. Usage rates decrease monotonically with time following publication, even as citation counts increase monotonically. Usage rates are a measure of the current use; citation counts are a measure of all past use. By taking the combined citation counts and usage rates for an aggregation of papers by a single author one obtains a two dimensional measure of that author's productivity or usefulness, which, in addition to the author's age, gives substantially more information than citation counts alone when evaluating performance."[65].

Citations and usages are different aspects of publications. They might need to be weighted differently, but they both provide data on forms of scholarly awareness and reception.

[63] Kousha/Thelwall/Rezaie, 2011: 15
[64] Kurtz/Bollen, 2010: 3-64
[65] Kurtz/Bollen, 2010: 28

Table 3: Usage rate vs. total citations for individual astronomers[66]



The solid line is a model for the most productive scientists at different ages.

"Considerable challenges still exist with regard to the standardization of recording and aggregation of usage data. In the present situation usage data are recorded in a plethora of different formats, each representing a different permutation of recording interfaces, data fields, data semantics, and data normalization."[67]

[66] Kurtz/Bollen, 2010: 29
[67] Kurtz/Bollen, 2010: 15

Project-based efforts to define processes (MESUR[68]) and to enhance standardization and compliance (COUNTER[69],SUSHI[70]) are still work in progress, but with sustainable structures after project funding. As with the EERQI instruments the aims of these projects are not to substitute traditional bibliometrics, but to extend coverage and to integrate new aspects of assessment.

5 Wrap-Up and Visions

Looking back at EERQI in the light of these parallel and future perspectives, some windows into future developments are open. If a lesson learned from the developments in the last 20 years is that a single tool of observation and assessment will not reflect the essentials and changes of scholarly output at once, we will have to rely on parallel and complementary developments for the time being. Instruments which integrate a combination of indicators and tools (like aMeasure) will be needed to enable the necessary effect of combination. On all levels of development, the standardization and calibration of indicators and tools will represent an enormous task. The complexity of scholarly publication and communication culture - in terms of publication types, formats and languages - determines the complexity of services necessary to capture the multi-faceted scholarly reality.

The idea of establishing a non-commercial service which is independently run by the scientific community and resistant against manipulation is understandable but misleading. The challenges of representative coverage, continuous maintenance and – again – standardization will necessitate building on commercial services from Google, Elsevier or Mendeley. Control of web-based services – be they commercial or not yet commercial – depends on the participation of the web community – in our case the scholars. Only scholars who adjust their mode of publication to an international publication culture will be visible. In the future, publishing in open access format will be one of the critical gates to awareness.

Webometrics and their diverse specifications will play a central role, even though they are also still biased - towards the English language and of course towards online publications. This bias will probably be reduced as more and more relevant publications will be available online, be they in English or in other languages which retain significance in scholarly communication. While unmistakable signs insure that the English language will sustain and extend its position

[68] http://mesur.informatics.indiana.edu/
[69] http://www.projectcounter.org/
[70] http://www.niso.org/workrooms/sushi

as international scientific communication language, the multilingual approach adapted by EERQI is still backed by European reality.

A-step-by-step approach to the Semantic Web, as the vision of an ontology-based structuring of web content, will continuously improve the conditions for research communication, but also research collaboration and easy re-use of research results and data, as a major part of research will take place in the web. In the field of computer linguistics, which was not specifically addressed in this article, many researchers are beginning to use the large body of resources available on the web to enhance scientific work and cooperation, especially in the SSH disciplines. Virtual research environments in these fields are designed around core facilities supporting semantic and lingual processes and collaboration[71].

There are at least three different and not yet integrated approaches to make the web better visible, the optimization of retrieval, webometrics, and the semantic and linguistic analyses. They all have been picked up by EERQI, as their combination is a very promising way to enlighten the world of scholarly productivity (not only, but also in social sciences).

6 References

Björk B-C, et al. (2010) Open Access to the Scientific Journal Literature: Situation 2009. PLoS ONE 5(6): e11273.doi:10.1371/journal.pone.0011273 http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0011273

Chotkowski La Follette, Marcel (Ed.) (1982): Quality in Science. http://mitpress.mit.edu/catalog/item/default.asp?ttype=2&tid=9519&mode=toc

Fensel, Dieter et al. (2011): Semantic Web Services. Springer Heidelberg. Haslam, N. & Laham, S. M. (2010): Quality, quantity, and impact in academic publication. European Journal of Social Psychology, 40/2, p. 216–220.

Ilg, Stefan. & Boothe, Brigitte (2010): Qualitative Forschung im psychologischen Feld: Was ist eine gute Publikation? FQS Forum: Qualitative Sozialforschung, 11/2, Art. 25. http://www.qualitative-research.net/index.php/fqs/article/view/1371/2976

Kousha, Kayvan, Thelwall, Mike & Rezaie, Somayeh (2011): Assessing the citation impact of books: the role of Google Books, Google Scholar, and Scopus. Journal of the American Society for Information Science and Technology. Published online in Wiley Online Library. DOI: 10.1002/asi.21608

[71] A major and exemplary European project is DARIAH: http://www.dariah.eu/

Kurtz, Michael J. & Bollen, Johan (2010): Usage Bibliometrics. In: Annual Review of Information Science and Technology, Volume 44, Issue 1, 3-64. http://onlinelibrary.wiley.com/doi/10.1002/aris.144.v44:1/issuetoc

Laakso M, et al. (2011): The Development of Open Access Journal Publishing from 1993 to 2009. In: PLoS ONE 6(6): e20961. doi:10.1371/journal.pone.0020961. http://www.plosone.org/article/info%3Adoi%2F10.1371%2Fjournal.pone.0020961#s3

Leinenkugel, Phillipp, Dees, Werner & Rittberger, Marc (2011): Abdeckung erziehungswissenschaftlicher Zeitschriften in Google Scholar In: Information und Wissen: global, sozial und frei? Griesbaum, Joachim et al. (Eds.) Boizenburg: Hülsbusch, 160-170.

Weingart, Peter & Winterhager, Matthias (1984): Die Vermessung der Forschung. Theorie und Praxis der Wissenschaftsindikatoren. Frankfurt/M/New York: Campus.

Work Programme 2012 Capacities (2011). European Commission C(2011)5023 of 19 July2011. ftp://ftp.cordis.europa.eu/pub/fp7/docs/wp/capacities/capacities-intro-wp-201201_en.pdf

Quality and Quality testing in the Humanities Perspectives from Research Funding

Axel Horstmann
Translation: Judith Keinath and Sarah McMonagle

Short Summary
The discourse on quality presents some difficulties for the humanities. At the same time, an increasing number of voices emphatically assert that robust appraisals of quality are in fact possible in the humanities and call urgently on that community to take an active role in the development and application of quality standards and corresponding procedures. Against this background, this chapter considers the problems and possibilities of quality and quality testing from the perspective of research funding, focusing on literary studies in the German-speaking context. Particular attention is given to the peer review and how it may be configured as a communicative process of quality assessment, as well as to 'professionalism' as a cornerstone of quality assurance within literary studies.

1 Quality testing - the only option?

"What the hell is quality"

This title question from Elisabeth Lack and Christoph Markschies's edited volume on "Quality Standards in the Humanities" (2008), sounds somewhat crude at first.[72] But after 295 pages, this question manages to lose all trace of crudeness and sounds more like a deep sigh; a mixed sigh of helplessness, impatience and perhaps even displeasure at the obstinate way in which this question seems to elude any conclusive, widely acceptable and above all practicable answer.

"What the hell is quality"

[72] The term 'humanities' here refers to German-speaking countries and mainly to linguistics, literary studies, cultural studies, including the philological disciplines, as well as to historical studies, philosophy, archaeology, theology, the fine arts and musicology. The subject of 'quality' raises very similar questions and problems for the social sciences, however this article does not deal with the social sciences. For a definition of the humanities see Donovan 2008, 76.

I. Gogolin et al. (eds.), *Assessing Quality in European Educational Research,*
DOI 10.1007/978-3-658-05969-9_13, © Springer Fachmedien Wiesbaden 2014

It is a comparatively new question – at least when specifically focusing on the humanities. Søren Kjørup, for instance, did not take note of it in his introduction to the humanities – published in Danish in 1996 and in German in 2001 – or at least did not consider it important enough to address it either explicitly or even to devote a section to it. Even the term 'quality' and its related terms – 'quality measurement', 'quality assurance', 'quality improvement', 'quality management', to name but a few – do not appear in the index (Kjørup 2001). The same goes for Achim Geisenhanslüke's second and unrevised edition of *Introduction to the Theory of Literature*, published in 2004. In contrast, however, Hans Bergemann's 'Select Bibliography on Quality Assurance in Journals for the Humanities and other Scholarly Disciplines' (2010) contains some 1,227 titles. Included in Helwig Schmidt-Glintzer's conference proceedings *On Ascertained Knowledge and New Insights* (2010), this is just a select bibliography! And those who carry out a Google search on "quality assurance" in relation to the humanities, will retrieve around 22,200 hits after 0.21 seconds.[73]

"What the hell is quality"

Yet what has thus become a central question in humanities discourses in Germany within a few years is not entirely promoted by the relevant disciplinary representatives. It seems, rather, that although the question is frequently raised externally, it is only ever lead reluctantly.[74] The reasons for the boom in quality discourse are manifold and are addressed so often in the relevant publications that it is not necessary to detail them here. In any event, this discussion impinges on the humanities as a discipline that, to a large degree, has become unsure of itself or does not really seem to have acquired self-confidence – despite all encouraging reports and funding programmes.[75] The 'crisis symptoms' of the humanities are well known and frequently discussed:

- Dwindling consensus within and between subjects on targets, subject matters and methods

[73] Accessed 15 January, 2011.
[74] The German Federation of Historians, for instance, is known to have refused to take part in a review conducted by the German Science Council; see Krull 2010, manuscript 2ff. Many thanks to the author for allowing me access to this paper.
[75] Examples include the German Science Council's recommendations regarding the development of the humanities from 2006, relevant funding offers from the German Research Foundation (DFG), as well as private funding sources such as the Fritz Thyssen Foundation, the German Stiftverband, the Volkswagen Foundation and the ZEIT-Foundation Ebelin and Gerd Bucerius ('Focus on the Humanities').

- The dissolution of subject boundaries and increasingly diverse discourses
- Lack of peer-reviewed journals that set standards
- Low expectations and chances of success, as well as (real or perceived) disadvantages when competing for funding, in particular third-party funding
- Lack of self-confidence and widespread sentimentality
- Loss of legitimacy and a defensiv approach to legitimisation
- No professional portrayal of capabilities and achievements to the public.[76]

It is little wonder, then, that the humanities find it difficult to engage in a debate on quality. To a certain extent, they appear to be downright traumatised in their reactions to a field in which they have become outsiders - a position they have not quite chosen voluntarily, but to which they have at least contributed. For, despite widespread criticism of excessive 'evaluitis' and questionable rating methods elsewhere, there is a broad consensus in the scientific community outside of the humanities, as well as at the policy and adminstrative levels, that the modern scientific system can no longer do without quality measurement, assessment and assurance anymore. Moreover, it is agreed that fundamental criticism of quality measurement does not make sense unless it aims at the specific improvement of both criteria and techniques, in particular as an increasing number of researchers depend on third-party funding for their work. And lest the floodgates be thrown open to arbitrariness, third-party funding ought only be assigned in good conscience following a thorough expert evaluation of the 'quality' of the respective candidates, their past performance and their newly planned projects.

It would be a distortion of the facts, though, to deny that this perception has by now spread among humanities scholars who are now deeply convinced that disclosure on quality and performance has to be part of their accountability. However, one of the few issues on which the humanities community is largely in agreement is that the mere adoption of criteria, indicators and procedures from other disciplines, and in particular the natural sciences, be consistently rejected. Rather, when it comes to quality, quality measurement and quality assurance, there is an almost universal insistence on special status for the humanities.

[76] See, for example, Herbert 2009, 31ff., as well as Krämer's 'Reply' 2009, 43ff.; also, Schmidt-Glintzer 2010, 29; Braungart 2008, 99ff., esp. 110f.

2 Special case humanities?

As already stated, there is widespread consensus that the quantitative methods of quality measurement deployed largely in the natural sciences are not useful to the humanities as they put them structurally at a disadvantage.[77] However, that is where the common ground ends, since there is anything but agreement on what ensues. On the part of the humanities, reactions range from a general rejection of any measurement ("the mind [Geist] is immeasurable") to ambivalent statements ("yes, but"), differentiated considerations ("partly-partly") and the more or less resigned admission that under the given circumstances no convincing alternative to quality measurement might be available, to the more offensive counter-argument that on the contrary it is possible to develop appropriate criteria and practical methods for the humanities. Achieving this, however, would require the relevant disciplines to take the initiative to devise and apply them themselves.[78]

The cited volume *What the hell is quality?* details impressively this wide-ranging discussion and the broad spectrum of positions therein. It also shows, however, that even those who flatly contest the possibility of serious quality *measurement* for the humanities, by no means refrain from making quality *judgments* in their daily business. Indeed, they make such assessments all the time: whether grading academic qualifications, reviewing manuscripts for publication in journals or book series, writing reviews of publications or appraising funding proposals, or, in the case of appointment procedures, when inspecting research centres and scientific institutes. In such instances there certainly seem to be agreed criteria. One also finds in many fields within the humanities a remarkably broad consensus on the academic standing of individual researchers, of publications and publication series, of institutes and institutions – but this is usually spoken of amongst one another and only ever off the record.[79] So it seems that what is missing is not so much the possibility of consensus, but the actual willingness to explicitly argue one's own valuation and, if need be, to have to face public debate. This applies as much to research as it does to teaching, where Ulrich Herbert has described both dramatically and tellingly "a cartel of silence" and a "firm code of silence", any breach of which is considered an unpardonable violation of unwritten guild laws (Herbert 2009, 36; Herbert/Kaube 2008, 41).

[77] In particular, the level of third-party funding as a preferred performance indicator is rightly criticised; see, for example, Krull 2010, 6f. For an empirically oriented approach to developing quality criteria for research in the humanities, see Hug/Ochsner/Daniel 2010, 91ff.
[78] See Felt 2008, 289f.; Schmidt-Glintzer 2010, 49; Krull 2010, 19f.
[79] See Herbert/Kaube 2008, 40; and Herbert 2009, 34.

Indeed, even among the sceptical one finds, for all intents and purposes, references to criteria and indicators that reliably determine quality in the humanities. Thus, according to Ulrich Herbert and Jürgen Kaube:

> It is wrong to say that there are no widely accepted quality standards within the humanities. They are, however, phrased informally, are usually confined to one subject area, and for a number of subjects are not readily transferable to other subdisciplines. Of course, the breadth of knowledge of the relevant material, the degree of reading, the analytical sharpness, the ingenuity and originality of the research, the plausibility of findings, and lastly the aesthetics of the langauge used, apply everywhere. (Herbert/Kaube 2008, 40)[80]

Helwig Schmidt-Glintzer substantiates this in terms of academic publications and, aside from general scholarly criteria, names a number of "quality attributes":

- Originality
- Professionalism
- Intelligibility
- Reflexivity
- Referentiality
- Focus on the reader
- Mono- or multilingualism (Schmidt-Glintzer 2010, 81)

Sybille Krämer goes one step further and differentiates between first-, second- and third-degree standards (Krämer 2009, 45ff). Furthermore, Wilhelm Krull names the four "Is" as "quality indicators for the humanities": infrastructure, innovation, interdisciplinarity and internationality (Krull 2010, 14ff).

Little more can be added at this point, except perhaps the reservation expressed by Ulrich Herbert and Jürgen Kaube: "... there is considerable flexibility as to the application of these criteria – anyone who quietly and contentedly acknowledges just how excellently each of these categories applies to his or her own work will admit this" (Herbert/Kaube 2008, 40f.).[81]

---

[80] Similar listings can be found in the Volume *What the hell is quality?* under Suder 2008, esp. 255; also Nießen 2008, esp. 261f.; see also Krämer 2009, 45ff.

[81] Similar to Schmidt-Glintzer 2010, 16, Herbert and Kaube use this opportunity to point out that the "small subjects" apparently find it easier to hold a discourse on quality, compared with the less manageable "mass subjects" (Herbert/Kaube 2008, 45); see also Herbert 2009, 31f.

Yet is it really just the considerable scope for flexibility in the *application* of these criteria that allows verdicts on quality to remain so divergent? That presents little opportunity for mutual consent and even gives rise to doubts on the feasibility of reaching any consensus on benchmarks?

I believe it is not only the leeway of these criteria that renders the feasibility and development of serious quality benchmarks in the humanities so difficult; a further difficulty is heterogeneity and, to a certain extent, the tendency for criteria to contradict themselves. This is where the issue becomes tricky, since the matter of quality standards in the humanities that achieve consensus at the same time raises the crucial question as to the humanities' academic status. This status essentially depends on whether 'good scholarship' can be plausibly demonstrated in the humanities and what differentiates that from less good scholarship or non-scholarship. Georg Braungart has thus remarked that the problem with quality measurement not only affects the epistemological foundations of the humanities, but ultimately their identity and security of survival (Braungart 2008, 103f.).[82]

3 Case study: literary studies

Not being a literary scholar myself, I don't promise the experts anything that is essentially new on the topic of 'quality and quality testing' within their own discipline. I focus on this discipline nonetheless, as it plainly depicts the problems, possibilities and perspectives within quality discourse. That is to say, if the humanities in general have difficulty in reaching consensus on subject matter and methods, then this is particularly the case for literary studies. *How* or *as what* 'literature' is viewed – or should be viewed – is disputed. For instance, 'hermeneutics', 'structuralism', 'deconstruction' and 'discourse analysis' – to name just some concepts of literary theory (see Geisenhanslüke 2004, 42ff.) – are not only completely different ways of *how* 'literature' is approached, but also different conceptualisations of *what* 'literature' itself is as a subject-matter. Achim Geisenhanslüke's reduced definition of 'literary studies' as a 'scholarly discipline' therefore comes as no surprise: "of course it concerns a certain knowledge about literature" which is fundamentally different to "knowledge of

[82] See also Braungart 2008, 104: "The status and future of the humanities are being […] decided on […] during academic strategic and policy disputes on evaluation methods, on procedures of quality assurance, and on criteria for the allocation of resources which depend on interests. This is conveyed internally and externally as a negotiation of disciplinary standards and cultural-economic conditions on the one hand, and externally-imposed (and therefore from those essentially outside of the subject area) criteria on the other hand".

literature" (2004, 8). The status of this knowledge, and with it the status of literary studies as a scholarly discipline, thus remains open. Geisenhanslüke, in turn, hopes for decisive assistance from within literary theory, whose specific task it is to substantiate "what literary studies entail and how they can legitimise themselves in contrast with other types of knowledge" (ibid.). Herein lies the problem, since literary theory itself does not provide any unambiguous answer to this basic question, according to Geisenhanslüke. He admittedly considers this to be rather an advantage:

> On the one hand, the fact that there is no longer the one theory of literature, but a multitude of rivalling approaches, may be deplored as a loss of clarity. On the other hand, this might be hailed as a sign of increasing complexity that has vastly expanded the possibilities of literary studies in the last few decades. (ibid., 15)[83]

Leaving aside the question of whether this is anything more than sheer optimism, even Geisenhanslüke cannot deny that such a "principle of plurality" – as conveyed through the "multiplicity of methods in literary studies" – gives rise to "differences that ultimately lead to conflict the authority of the various methods themselves" (ibid., 11). For the time being there seems to be no clear winner in this conflict and therefore no generally accepted answer to the question of how to determine quality, at least where methodologies are concerned. The issue is not made any simpler by the fact that many experts claim that literary studies have lost their subject matter, i.e. literature, in the face of excessive theoretical discussions (see ibid., 12). To what purpose and for whom are literary studies pursued if there is no longer a common object for research and teaching to reliably draw upon?

Literary studies are thus a prime example of the difficulties that the humanities face in the disussion on quality, quality assessment and quality assurance. For the time being no panacea can be hoped for – certainly not from me and at this point. What I can and hope to contribute in the following section are experiences from, insights into and reflections on how quality testing, assessment and assurance are depicted from the point of view of research funding. Of course, I hope that this will also serve to draw some conclusions for literary studies. I would like to state in advance that I will focus exclusively on *research* in the humanities, although well aware that *teaching* in the humanities poses problems that are no less serious in the same regard.[84]

[83] Examples might include the lengthy and ultimately fruitless disputes between advocates of descriptive and prescriptive, quantitative and qualitative, empirical and theoretical approaches.
[84] See Herbert 2009, 40; Herbert/Kaube 2008, 41f.

4 Quality testing within the framework of research funding

Manfred Nießen and Frank Suder have already dealt with this subject in detail in the repeatedly cited volume *What the hell is quality?* They offer revealing insights and appraisals gained during their many years of experience with the German Research Foundation (DFG) and the Fritz Thyssen Foundation, respectively (Suder 2008, 251ff.; Nießen 2008, 259ff.).

From my own experiences with the Volkswagen Foundation, I can only underscore most of what they say. First of all, Nießen's reference to the specific situation that all research funding must face in terms of quality assessment is important. Research funding is not - at least not primarily - about appraising previous academic accomplishments. Rather, it is about assessing the academic validity and significance of undertakings for the future (Nießen 2008, 260), i.e. a *planned* event, research project or publication. Of course – and Nießen rightly points this out – the past performance of the candidates plays an important role here. Although no funding decision can be based on this alone, it allows for a degree of confidence to be placed in the candidates and their ability to reach their proposed objectives. Incidentally, those essential quality indicators and criteria are the same as being applied in the assessment of demonstrated research achievements – but just prospectively and expanded to include the vital question on the 'risk of failure' that the potential funding body is willing to incur.[85]

It goes without saying that quantifying methods, such as counting the principal investigator's publications and measuring their impact, are therefore only auxiliary in such assessments.[86] The decisive method in research funding remains the peer review, to which, despite its undeniable weaknesses, no convincing alternative has been found.[87] To counter weaknesses such as subjectivity, mainstream bias, risk aversion, and susceptibility to extra-academic interests and influences, funding institutions administer "references for peer reviewing" to the expert reviewers. They make clear what they deem to be the most relevant assessment criteria – contribution to the scientific discussion, plausibility, personal qualifications, commensurable costs, etc. – and commit the reviewers to "general rules of good practice".[88] Reviewers for the Volkswagen

[85] See Suder 2008, 256.
[86] See Braungart 2008, 103ff.; Krull 2010, 5ff.
[87] See Schmidt-Glintzer 2010, 37; against this background, Donovan 2008, 76ff., enquires after the most suitable procedure for quality assessment in the humanities and likewise sticks with the peer review; see also Hornbostel 2008, 68. Nießen 2008, 265, defines the peer review as "organised and thus well-regulated reasoning about criteria concerning applications, carried out by those who belong to the same argumentative framework as the applicants".
[88] When sending the respective application documents to reviewers, the Volkswagen Foundation also usually encloses its "Guide to Peer Review" or refers to the relevant pages on its website.

Foundation, for instance, are assured of unconditional confidentiality through anonymised critical assessments and suggestions. This enables frank and earnest assessments to be reached, even in delicate situations. The foundation expressly notes this in its relevant memoranda so that applicants know in advance what they have to be prepared for. While this does not provide the frequently demanded transparency around the naming of reviewers, it does provide transparency regarding the reviewing procedure itself.

Of course, neither review guidelines nor appeals to fairness can entirely prevent the influences of personal sympathies or antipathies in the review process. It is then the task of those who appraise the reviewers' opinions to identify any such buried 'arguments' and to qualify them when weighing up.

That does not always prove easy and such biases that prejudice assessment often remains undetected. This is especially the case for the humanities where consensual decision-making is rather exceptional, regardless of whether it concerns a small symposium or a large project, an individual person or an entire research institute. Most likely, personal biases that cloud reviewers' perceptions are in many cases (also) responsible for the differences in reviews. This makes it all the more important to configure instruments and techniques for assessment in a way that allows such influences to be detected and neutralised. Against this background, the peer review procedure for research funding has been methodologically refined and improved. In many funding initiatives of the Volkswagen Foundation, for example, individual written peer reviews are no longer the only basis of decision-making. In particular, large funding applications are put through a multi-stage procedure. Here, a predominantly subject-related individual review process is followed by a comparative examination of shortlisted applicants by a multidisciplinary and international expert panel through joint reviewer meetings. Then, funding offers that are made to individuals always include a personal presentation by the respective applicant. This yields a remarkable finding: no matter how much the appraisals of the panel members differ at the outset, unanimity is almost always reached by the end of the process – even for applications from the humanities! This does not just show that, as previously discussed, there are indeed criteria and standards that find general consensus in the humanities; it also underscores the significance of personal discussion and the face-to-face exchange of arguments, points of view and evaluations for an assessment in the humanities.[89]

[89] See Nießen 2008, 271; for Nießen "the obligation to argue when talking to colleagues" is the decisive factor. Beiner 2009, 43ff., also considers the connection between "dialogism" and "intersubjectivity" to be the distinguishing feature in humanities research as "discursive practice"; see also Schmidt-Glintzer 2010, 74, who recommends a "network-based quality assurance" in order to neutralise any possible prejudices held by individual peers.

This is not to say that communication is always easy and that the exchange of arguments, points of view and evaluations leads to unanimous decisions without any problems. On the contrary! Ultimately, it is not only necessary to assess the quality and significance of any one application, but also to decide on the criteria by which such quality and significance can be assessed and - especially in the case of multidisciplinary projects – on the value to be attributed to each vote in the process.

What, then, counts as an argument in the context of a reviewing procedure? What does power of judgment mean and who can claim it?

5 Levels of assessment, arguments, expertise

This much should be certain: In matters of academic quality and significance, only someone with sufficient academic expertise is entitled to deliver a judgment! The various levels of assessment also appear to be somewhat beyond dispute. They are:

- Subject and research question
- Methodology and research design
- Qualifications of the relevant parties
- Anticipated results and impact
- Cost-income analysis
- Chances of success
- Position vis-a-vis competing projects

But this is where things become complicated, most notably in the humanities. What some may consider to be an exciting topic or groundbreaking research question, others may take to be rather absurd or academically questionable; where some evaluate methodology and research design as rock solid, others criticise its lack of innovation; where one opinion deems the professional qualifications of the relevant party to be insufficient for the planned project, another review praises the courage of venturing into unknown disciplinary territory; where some regard methodological risks to be acceptable in view of anticipated academic breakthroughs, others demand more certainty around results; where one reviewer lauds intended findings to be highly significant to a respective discipline, another criticises their lack of interdisciplinary relevance and sees a glaring disproportion between costs and benefits.

The daily grind of reviewing offers plenty of examples to illustrate this, admittedly rather simplistic, typology of controversial arguments. The

humanities consistently provide striking examples of how difficult it is to to achieve consensual decisions under such conditions, since it is not just about reaching agreement on every level of assessment, but also on how each of these levels ought to be weighted in principle among themselves. For instance, can a topic and research question of a given project be so innovative, exceptional and exciting that a certain lack of clarity regarding method and design be deemed unavoidable and therefore acceptable to a certain extent?

Of course, such questions emerge in other disciplines too. But the additional difficulty in the humanities is - as already mentioned - that disagreement frequently prevails not only about research design and methodologies, but also whether a certain project can be considered a relevant research subject at all. Common 'frontiers of research' that seem to offer certain guidance to the scientific community of the natural and engineering sciences, are hardly to be found within the humanities. Although that is not without reason: In the humanities, addressing the ethics of Aristotle can be no less groundbreaking for philosophical research than contributions to the moral aspects of modern prenatal diagnosis.[90]

Furthermore, humanities scholars tend to stress their disagreement with items under discussion, allowing disproportionate time for criticism, even if it concerns only marginal points that in principle do not affect the final acceptance.[91] This may be due to the fact that research subjects, questions and concepts are more strongly linked to the individual personalities who engage with them than in other academic areas. Manfred Nießen has pointed out this peculiarity, identifying it as the main reason quality judgments in this field are taken much more "personally" and why rejections particulary "offend" (Nießen 2008, 264f.). There is no doubt that strategic thinking, which places a joint interest in strengthening (third-party funding for) the humanities above and

[90] Schmidt-Glintzer 2010, 42, points to the fact that although the humanities also strive for "innovation", "grappling with tradition" remains part of them; see Nießen 2008, 263: "Debates in the humanities […] are not about 'the one' new space and its 'mappability' at a given time, but rather about competing or complementing ways to position"; against this background, Felt 2008, 289, calls for a closer look to be taken at the hitherto relatively unknown contexts of production within the humanities; see also Beiner 2009, 59ff. Of course it is also possible to name areas in the humanities where joint efforts of the scientific community are made over longer periods of time; this is the case for extensive volumes of works, dictionary projects, reference material or other types of documentation and development projects, up to archaeological excavations that are often conducted or supervised by scientific academies as 'long-term projects'. Such projects provide the humanities with indispensable *foundations* for their work, yet as such only account for one part of the entire research spectrum. Certain eras can also become such *frontiers of research* – take, for example, research into the Baroque period.

[91] See, for example, Krull 2010, 8ff. In contrast, Schmidt-Glintzer 2010, 11, specifically underlines the indispensability of "difference" in humanities discourse; see also Krämer 2009, 43ff.

beyond personal needs for distinction and differentiation, can be very difficult to develop – a hindrance that should not be underestimated in competition with other disciplines.

Finally, it should not be overlooked that the transfer of findings and results to the non-academic public is becoming increasingly important as an assessment criterion (as well as academic quality criteria) in the humanities.[92] Sybille Krämer, for instance, expressly welcomes the fact that the humanities not only publish in the academically exalted formats of journal articles and monographs, but also in feature pages, exhibition catalogues, the telecommunicative media and through various lectures that the public are more likely to notice (Krämer 2009, 50).

And Georg Braungart appears to be convinced that the difficult question surrounding "quality in the humanities" will increasingly concern "their ability to impart results" (Braungart 2008, 110f.).[93] How then, in case of doubt, should academic value be balanced against non-academic impact?

Should this be the case – and it seems likely that it is – then not only will the spectrum of knowledge, abilities and experiences necessary to pass a proper judgment on quality expand; the question of who possesses those manifold competencies and should therefore be called upon to review a complex project intensifies: experts from the directly relevant discipline, professionals from alternative disciplines to allow comparative examination, or even adequately skilled lay people to represent public interest? And what should the composition of such a panel look like in this case? This much seems to be clear: the judgment of disciplinary peers alone no longer suffices in funding decisions on sophisticated projects in the humanities, particularly for those projects that aim to cross traditional disciplinary borders. Fortunately, such projects are no longer one-off cases.

In case of doubt, whose vote should be decisive?

Of course, this question cannot be answered in such an abstract and general form. Manfred Nießen alluded to it in his analysis of "reviewing as opinion formation in a social context". According to him, quality assessment in the humanities is always, and probably to a higher degree than in other disciplines, a

[92] The Volkswagen Foundation was clearly aware of the vital importance of conveying research results to the public with the establishment of its funding initiative 'Key Topics in the Humanities' in 1998. Among other things, assessment depended on the ability of applicants to make a convincing case in this respect. This criterion still applies.

[93] Braungart invites the humanities to make use of their real relation to society and to bring their "cultural capital" to bear.

"matter of interpretation and negotiation" (Nießen 2008, 262). This comes as no surprise since the humanities thrive especially on discourse, also where reviews are concerned.[94] Whoever expects quick opinions and judgments without the 'ifs and buts' hasn't quite grasped this discursive character. Reliable assessments of academic activities, either accomplished or intended, are conclusively reached by way of a communicative process. And this process, Nießen rightly remarks, must come to an end for pragmatic reasons: applicants eventually need to know whether their projects are going to be funded or not. Yet the nature of this process remains open-ended which means, in principle, that its results are always revisable.[95] This may sound unsatisfactory and disappoint those who hope for a conclusive formula for 'right and wrong' in this context, but it does not change the facts.

"Opinion formation in a social context" allows everything in principle to be questioned and discussed: the various levels of assessment, the weighting of each level, the plausibility and significance of the arguments presented and, last but not least, the competence of those involved in the reviewing process. From the outset, it is anybody's guess which arguments and assessments will win out in the end. That does not mean that funding decisions are incidental to the review process. Quite the contrary is the case – at least where the panel manages to reveal *everything* that might argue on behalf of or against a certain project, i.e. *all arguments and counter-arguments* (including the 'buried' ones), and *to scrutinise jointly their content and significance*. The persons involved need to be able to listen carefully, to deal fairly with opposing opinions, not to cling to their own position at any cost or impose it on others, and allow themselves to be won over by the better argument, i.e. purely and simple to muster the 'communicative reason' necessary for the success of any serious dispute – both within and outside of academia.

From my own experience in funding I can confirm that, despite the challenging conditions, this can be achieved in the humanities. However, it must also be noted that quality assessment here requires higher-than-average amounts of time and personnel – at least if it is to be carried out seriously, yielding reliable results.[96]

And what are the implications for literary studies?

---

[94] See Beiner 2009, 50ff.

[95] Felt 2008, 284f., takes a different view. Disputing Nießen's conversational character thesis, she challenges the eventuality of continuing communication once a funding decision has been made.

[96] See Donovan 2008, 95. Braungart 2008, 109, perceives an (imminent) absurd disproportion between the costs of eleborate evaluation procedures and the funds available for research itself.

That literary studies serve as a prime example of the difficulties faced by the humanities in quality discourses can only be underscored from the perspective of research funding. But here the problem is further exacerbated by the fact that not only are the understanding of and approach to literature disputed, but so too are the goals and potential recipients that literary studies should focus on. In the background of quality discourse, there therefore looms a question of wider impact:

6 Why literary studies and for whom?

Undoubtedly literary studies serve to secure, expand, deepen and augment academic knowledge of literature. Equally obvious are those to whom this acquired knowledge is aimed: disciplinary colleagues and contemporaries – just the same as in other subjects.

But does this suffice as an answer to the question above? One would have to reckon with the biting criticism of Tristram Shandy who has accused the "friends of scholarship" of writing "new books over and over … the way apothecaries make new mixtures" by "pouring water from one jar into another" (Sterne 1966, 357).

It is all the more pressing, then, to impart literary research findings beyond the academic community and to deliver the benefits to those who share in the subject matter of literature as readers and who want to be attracted, informed, entertained, inspired and enthused by it. If the humanities are generally expected to present their compiled knowledge to a broader public, then this is particulary obvious for literary studies. They would be well advised to take this seriously and not to disappoint unnecessarily. For if they manage to convince a broader public that literary research is not just the preserve of a self-absorbed scientific community, but that it offers insights from which society can largely benefit – e.g. by keeping society from intellectual impoverishment, by immunizing against attempts of stultification, by unlocking intellectual potential and offering cultural guidance –, only then, I believe, will literary studies remain viable in the long-term. This isn't just a matter of preserving their personnel and financial resources in the face of competition from other disciplines; it also concerns the lifeblood that connects literary studies with society and ensures the necessary 'external' influx of cultural and social vigour.

As well as academic performance, literary studies must therefore also take transferability to a non-academic public into account when designing quality standards. First and foremost, the literary studies community itself needs to de-

velop pertinent criteria and define corresponding specifications – provided it wants to retain jurisdiction in this crucial field.

However, a further problem emerges. Helwig Schmidt-Glintzer has rightly pointed out that the humanities cannot do without value judgments (Schmidt-Glintzer 2010, 35; see also Krämer 2009, 44). And literary studies are no exception here: when decisions must eventually be reached on, say, whether a literary text ought to be included in the academic canon, value judgments simply cannot be avoided. The same applies to readers who expect advice from literary experts on whether it is 'worth' reading a certain book - either for education or edification, for entertainment or to pass the time. A consequential development within the humanities in the nineteenth century was the relinquishment of literary criticism from academic studies to become the sole responsibility of the feature pages. The reasons behind this development cannot be expanded upon here.[97] But it remains to be noted that by dispensing *a priori* with real avenues for literary criticism, literary studies passed up the chance to awaken public interest and thus to improve their public image.

That is not to say that literary studies should lower or forego professional aspirations in favour of a consumer-friendly shallowness. Rather, meticulously executed research of the highest possible expertise that ensures the reliability of their results, coupled with a no-less professional and comprehensible transfer of these results beyond the narrow circle of experts, is the desired balance. The latter, however, by no means concerns a transfer only to the 'general' public. In a time of highly specialised research, the notion of the 'public' begins in neighbouring academic subjects. For if inter- and transdisciplinary research is not just some hollow phrase, then it presupposes comprehensibility surrounding experts' activities and results. What has therefore been termed the "rhetorical quality" of presentation (Schmidt-Glintzer 2010, 21), is not just some decorative term in academia that can be dispensed with if need be; rather, it helps to ensure proper reception to academic works.

This poses an additional challenge to quality discourse in literary studies: Not only is it necessary to develop and apply a sufficiently complex concept of 'quality', but the notion of 'professionalism' and how it can be shaped in teaching and through studies must also be reconsidered. Quality cannot be attained without professionalism, and not just in the present context. But since literary studies are so closely associated with cultural life in general, they need to be especially interested in maintaining professionalism (this applies to other humanities subjects that are just as close to society). Although it is not always possible to draw a distinct line between the more popular scientific publications and re-

[97] See Horstmann 1992, 186ff.

search literature that is aimed at an academic audience (cf. Hornbostel 2008, 60), and although the humanities themselves can be interpreted as part of a comprehensive history of tradition and reception,[98] it must be borne in mind that museum tours are not the same as studying history, visits to the opera are not musicology, and reading books is not the same as studying literature.

The difference is based on professionalism; professionalism "when dealing with languages, texts and images", as Sybille Krämer has put it (Krämer 2009, 45); professionalism, not just in the sense of scholarliness and mastery of professional techniques, but primarily as the ability to make oneself *understood*. "To do everything in a way that will be comprehensible to others" – Sybille Krämer is right in pointing out that, by adopting this maxim, academic pursuits directly connect with everyday activities. But this essentially depends on adopting "a rational and reasonable application of these symbolic instruments vis-à-vis our daily habits as *standards of the first order*". According to her catalogue of requirements

> This includes clarity of speech, the ability to define terms, clarity of theses, the plausibility of explanatory statements, transparency and soundness of argument, responsibility for texts and quotations by referring back to 'sources' and, last but not least, a certain economy in use of terms, thoughts and text length in general (Krämer 2009, 45)

Admittedly, professionalism in the humanities goes beyond the mere standardisation of everyday practices. Here too, I concur with Sybille Krämer that, due to its commitment to the "values of the Enlightenment", professionalism is particularly characterised by its ability to keep a distance from certain issues, by reflexivity and the ability to respond constructively to criticism (ibid., 46f) – including intellectual honesty that is equally prepared to oppose both the academic mainstream as well as extra-academic interference. Here we have come full circle in relation to the public: history studies are necessary to protect the interested layperson from popular misguidance; musicology is required to sensitise music lovers against savvy deception; and literary studies are necessary to cultivate a grasp for the differences between high-circulation trash and literary art.

To both develop and apply standards of quality and professionalism in this challenging context is, as already discussed, first and foremost up to the scientific community itself – and this must be done increasingly through international co-operation.[99] This applies as much to literary studies as it does to the humanities in general. Here, however, is not the place to offer recommendations and

[98] See Gadamer 1975.
[99] See Nießen 2008, 268ff.

suggestions to this end. All the more, it must be emphatically underlined that quality discourse as basic professional self-reflection will only succeed by way of communication – and even then it would be best not to count on long-lasting results.

Friedrich Schleiermacher pointed out that when no one can any longer claim to possess the one, true knowledge, nothing but dialogue will help; a dialogue as an exchange of different opinions, perceptions and assessments in the joint *search for knowledge* (Schleiermacher 1976). If I am not mistaken, this applies in an exemplary way to quality discourse in literary studies. After all, it is hardly likely that anybody there will want to claim – much less be able to claim – monopolies on definition. The quality of the results of this discourse will ultimately depend on *how* this discourse is conducted; it can only be hoped that the necessary professionalism will be allowed to prevail – for the sake of literary studies, for the sake of the interested public and, last but not least, for the sake of literature.

6 References

Beiner, Marcus, Humanities. Was Geisteswissenschaft macht. Und was sie aus macht, Berlin 2009.

Bergemann, Hans, Auswahlbibliographie zur Qualitätssicherung in (geisteswissenschaftlichen Zeitschriften, in: Helwig Schmidt-Glintzer, Von gesichertem Wissen und neuen Einsichten. Dokumentation einer Expertentagung zum Thema „Geisteswissenschaftliche Zeitschriften – Referenzsysteme und Qualitätsstandards“, Wiesbaden 2010.

Braungart, Georg, Qualität und Qualitäten: Forschungsmessung in den Geisteswissenschaften? In: Elisabeth Lack / Christoph Markschies (Hg.), What the hell is quality? Qualitätsstandards in den Geisteswissenschaften, Frankfurt am Main / New York 2008, 99-111.

Donovan, Claire, Das zweiköpfige Lama zähmen: Die australische Suche nach den besten Evaluierungsmethoden für die Geisteswissenschaften, in: Elisabeth Lack / Christoph Markschies (Hg.), What the hell is quality? Qualitätsstandards in den Geisteswissenschaften, Frankfurt am Main / New York 2008, 74-98.

Felt, Ulrike, Angemessen messen? Die Qualität von Forschungsprojekten in den Geisteswissenschaften, in: Elisabeth Lack / Christoph Markschies (Hg.), What the hell is quality? Qualitätsstandards in den Geisteswissenschaften, Frankfurt am Main / New York 2008, 273-291.

Gadamer, Hans-Georg, Wahrheit und Methode. Grundzüge einer philosophischen Hermeneutik, 4. Aufl. Tübingen 1975.Geisenhanslüke, Achim, Ein führung in die Literaturtheorie. Von der Hermeneutik zur Medienwissenschaft, 2. Aufl. Darmstadt 2004.

Hempfer, Klaus W. / Philipp Antony (Hg.), Zur Situation der Geisteswissenschaften in Forschung und Lehre. Eine Bestandsaufnahme aus der universitären Praxis, Stuttgart 2009.

Herbert, Ulrich, Geisteswissenschaftliche Standards in Forschung und Lehre, in: Klaus W. Hempfer / Philipp Antony (Hg.), Zur Situation der Geisteswissenschaften in Forschung und Lehre. Eine Bestandsaufnahme aus der universitären Praxis, Stuttgart 2009, 31-42.

Herbert, Ulrich / Jürgen Kaube, Die Mühen der Ebene: Über Standards, Leistung und Hochschulreform, in: Elisabeth Lack / Christoph Markschies (Hg.), What the hell is quality? Qualitätsstandards in den Geisteswissenschaften, Frankfurt am Main / New York 2008, 37-51.

Hornbostel, Stefan, Gesucht: Aussagekräftige Indikatoren und belastbare Datenkollektionen. Desiderate geisteswissenschaftlicher Evaluierung, in: Elisabeth Lack / Christoph Markschies (Hg.), What the hell is quality? Qualitätsstandards in den Geisteswissenschaften, Frankfurt am Main / New York 2008, 55-73.

Horstmann, Axel, Antike Theoria und moderne Wissenschaft. August Boeckhs Konzeption der Philologie, Frankfurt am Main / Berlin / Bern / New York / Paris / Wien 1992.

Hug, Sven E. / Michael Ochsner / Hans-Dieter Daniel, Entwicklung von Qualitätskriterien für die Forschung in den Geisteswissenschaften – Eine Explorationsstudie in den Literaturwissenschaften und der Kunstgeschichte, in: Qualität in der Wissenschaft. Zeitschrift für Qualitätsentwicklung in Forschung, Studium und Administration 4 (2010), 91-97.

Kjørup, Søren, Humanities. Geisteswissenschaften. Sciences humaines. Eine Einführung. Aus dem Dänischen von Elisabeth Bense, Stuttgart / Weimar 2001

Krämer, Sybille, Replik, in: Klaus W. Hempfer / Philipp Antony (Hg.), Zur Situation der Geisteswissenschaften in Forschung und Lehre. Eine Bestandsaufnahme aus der universitären Praxis, Stuttgart 2009, 43-51.

Krull, Wilhelm, Vom Nutzen und Nachteil der Qualitätsbewertung für die Geisteswissenschaften. Vortrag an der Universität Zürich am 2. März 2010 (Manuskript).

Lack, Elisabeth / Christoph Markschies (Hg.), What the hell is quality? Qualitätsstandards in den Geisteswissenschaften, Frankfurt am Main / New York 2008.

Nießen, Manfred, Begutachtung als Urteilsbildung im sozialen Kontext, in: Elisabeth Lack / Christoph Markschies (Hg.), What the hell is quality? Qualitätsstandards in den Geisteswissenschaften, Frankfurt am Main / New York 2008, 259-272.

Schleiermacher, Friedrich, Dialektik, hg. v. R. Odebrecht, Darmstadt 1976; Nachdruck der Ausgabe Leipzig 1942.Schmidt-Glintzer, Helwig, Von gesi chertem Wissen und neuen Einsichten. Dokumentation einer Expertentagung zum Thema „Geisteswissenschaftliche Zeitschriften – Referenzsysteme und Qualitätsstandards“, Wiesbaden 2010.

Sterne, Laurence, Das Leben und die Ansichten Tristram Shandys. Deutsch von R. Kassner, Berlin / Darmstadt / Wien 1966.

Suder, Frank, Lohnt der Aufwand? Zum Thema Drittmittel von Stiftungen, in: Elisabeth Lack / Christoph Markschies (Hg.), What the hell is quality? Qualitätsstandards in den Geisteswissenschaften, Frankfurt am Main / New York 2008, 251-258.