

THE EUROPEAN EDUCATIONAL RESEARCH QUALITY INDICATORS (EERQI) PROJECT¹

Ingrid Gogolin
Antje Hansen

The EERQI project was motivated by the fact that the international notion of scientific quality as being the main determinant on which research is funded and supported may cause undesired side effects, if the questions of how quality is interpreted and how it is measured are not adequately answered. Current instruments for ‘measuring’ quality via citation counting and similar methods do cause such side effects, as they are strongly biased and largely inadequate for research in the Social Sciences and Humanities. The EERQI project developed an approach to detect the quality of research texts – with educational research serving as model case for Social Sciences and Humanities (SSH) – by applying an intelligent combination of different approaches that complement each other. This is what we call the EERQI Prototype Framework. It consists of products and methods that can serve as alternatives in processes of assessment of quality in SSH research. The possibility of multilingual assistance of assessment processed by EERQI’s multilingual search engine and automatic semantic analysis are tailor-made for strengthening the European research space, but may also be inspiring for other areas of the world that are not adequately recognised in the current instruments of quality assessment. The EERQI products and methods consist of a content base with educational research texts in four European languages that were exemplary included in the EERQI project: English, German, French and Swedish. Furthermore, a multilingual search engine was developed that is capable of finding educational research texts in the Web in the four EERQI languages. An approach to automatic semantic analysis and first tests of a citation analysis method that has the potential to be further developed for the application to educational research (and other SSH) texts do also belong to the EERQI products. Last, but not least, the project developed and tested a set of text-immanent (intrinsic) indicators for the detection of quality in educational research publications, together with an accompanying peer review questionnaire that operationalised those indicators. The instrument was positively tested for reliability and practicality in the European educational research community. The EERQI team would welcome any further development and adaptation of its approaches that take the specific situation of research in other areas of the world into account.

CONTEXT AND OBJECTIVES OF THE PROJECT

All across the world, the structures and control mechanisms of publicly funded research have changed dramatically in the last decade. There are many widely discussed causes of these developments. The set of causes on which the EERQI-project concentrated is based on the evocation of the ‘ability to compete internationally’ – a request that is expressed vis-à-vis national research landscapes in Europe as well as the whole European research area, and possibly also research areas in other continents.

A metaphor that is either explicitly used or implicitly resonates in the existing discourses is *quality*. The discovery, improvement and promotion of research quality and the quality of research outcomes – such as publications – are the driving motives for the tendency to re-evaluate and redevelop structures for the research area, for redesigning the funding of research institutions and projects, and for instituting control and legitimisation systems that are (or intend or pretend to be) helpful for decision-makers.

In the framework of these developments, the questions of *how quality is interpreted and how it is measured* are of fundamental importance. Analyses dealing with this question supplied the starting point for the development of the research project ‘European Educational Research Quality Indicators (EERQI)’ focusing on questions such as: What constitutes and marks the current quality control systems that are applied in contexts of governance and funding, irrespective of the genre and the type of research that is at stake? And what are possible (desired and undesired) effects of these systems on research that is conducted in the European research area, especially in the domains of the social sciences and the humanities?

The project was developed by a truly interdisciplinary European research consortium, a unique composition of experts from educational science, biblio- and webometrics, information and communication technologies, computational linguistics and publishing houses.

Educational research is particularly suitable for considerations and research on such questions, because it can be considered to be prototypical for vast areas of the whole field of the social sciences and humanities. Educational science and research combine a wide spectrum of theoretical and methodological approaches – from primarily philosophical-historical methodologies as used in the humanities to psychologically or sociologically based empirical observations of individual development, education, training or Bildung (formation); from hermeneutical interpretation, single case studies to the generation and statistical analysis of great amounts of survey data. This manifests relevant characteristics of knowledge production, which are also found in other disciplines in the social sciences and humanities.

Another reason why it was appropriate to choose educational research as a model is that the visibility of education and learning as a policy space and its emergence as a significant area of policy are not matched by useful analyses of its operation. Policy in education and educational research is no longer the sole domain of the nation-state, but has become a key feature of a ‘Europeanizing’ process – or even: of global developments. ‘New Learning’ through social innovation is central to the knowledge economy, allowing education to be compared, promoted, researched and improved in its role as a key part of the knowledge economy. However, the contribution of European education research to broader global debates is hampered by the way it is organised. Distinctive and fruitful traditions of work are locked into national intellectual resources and it is a slow process of enabling them to move across borders. Thus, there is a need to intensify networks and agree on common standards paving the way to a shared discourse space for European educational researchers – or even beyond this.

Before the project started, a firm review state of the art research on quality assessment was carried out. The review resulted in a generic judgment that can briefly be articulated as follows:

The existing instruments do not lead to valid results, because they do not measure what they claim to measure. An example to illustrate this statement is quality assessment based on citation indices and journal rankings. As yet, this has been the most common approach in vast areas of quality assessment in higher education.

The central and joint quality criterion that is used in these instruments is ‘international visibility’ of research findings. This is expressed by the placement of the publication, namely in journals with a good reputation, and by the number of citations of a publication. This approach is – for example – characteristic of the Social Science Citation Index, a commercial instrument, owned by the US American publishing group Thomson Reuter. Its results often play an important role in reporting systems on research achievement. A closer look at the documentation of the journals represented by this index reveals (for the field of educational science according to the “Journal Citation Report”) the following:

In total, 201 educational research journals were incorporated in the rankings of 2009. More than half of these journals are published by US American publishers. An additional 25% derive from British publishing houses. The next ‘largest’ nations in this ranking are the Netherlands (with 4% of cited journals) and Germany (with 3% of publications). Altogether, 15 nations across the world are represented in the ranking. India is not among them.

Another perspective on the Journal Citation Report reveals that 89% of the publications that were ranked in 2009 are in English. The next ‘largest’ languages with 2.5% and 2% respectively are German, Spanish and Turkish. In total, eleven languages are represented by the index. A language such as French is not included.

Although the owners of these instruments are constantly striving to improve their methodologies and although these methodologies differ more or less extensively between the instruments, the general problem of their validity is illustrated by these findings of the preparatory EERQI survey. These kinds of approaches do not produce valid information in the sense that they pretend to do. The illustration shows that the intended international relevancy of the included publications cannot be proven. The rankings are heavily biased. They essentially refer to US American or UK publications and to publications written in English. International visibility as a quality criterion must be translated here to. Visibility of products from a narrow selection of national research spaces to the rest of the world. The provided information is perfectly suitable to substantiate the dominance of a 'minority' of regional and linguistic research areas.

This means in fact, that these methodologies do not reflect an adequate coverage of (not only) European scientific publications, in particular in the social sciences and humanities. Not only are most other languages than English and publications produced outside the US and the United Kingdom ignored but also are other types of publication, for example books excluded. This means a bias to actual publication practice in educational research as well as the whole field of SSH. Another criticism is that mere citation counting may, if at all, indicate research quality in natural sciences, where a 'cumulative model' of citing is broadly practised. A citation indicates that the citing author positively refers to and builds on the work of a cited author. In the social sciences and humanities however, this is not the main tradition and function of citations; instead, citations are often used for contradiction or rejection of another position, or for negotiation. Thus, if citations are simply counted rather than analysed in their intention, the significance of the result is weak.

Based on such analyses, the motivation for the development of the EERQI project was, in a nutshell, the observation that the strategies of assessment that were developed in 'hard science'-contexts have to be heavily criticised for their methodological weakness and lack of validity. At the same time, there was a genuine desire to develop new approaches that can serve better for the aim of detecting research quality in our domains. This desire unites the SSH-research community as well as relevant stakeholders from other spheres, such as publishing houses, research funding and political decision making.

The EERQI team's general intention was to develop useful tools that support the process of quality detection, generally aiming to

- a) raise the transparency and value of the process of quality detection itself; and
- b) make the task better manageable and less time consuming.

It was not EERQI's objective to develop one single method, such as another index that can compete with the existing ones. The aim was the development and testing of a set of tools with different functions that can support and accompany the process of detecting research quality in texts - from the moment of identification of a text to the moment of the conclusive determination of quality. The new tools to be developed were compiled in a broader prototype framework, each tool addressing a specific part of the assessment process:

Before being able to assess educational research documents for their quality they have to be identified and gathered. Therefore, when searching for a specific term, it is the task to identify relevant educational research documents and make them available to the user. The EERQI-project developed a specific search engine that is able to identify educational research texts (see <http://makalu.xrce.xerox.com/eerqi/>). The stock that the search engine comprises of contains a wide range of documents, those being freely available in the World Wide Web and those being in the possession of publishing houses or research institutions – which are normally not freely accessible. Since the relevance of the harvested documents plays a crucial role, it was the task of the developers to refine the search engine in order to gather only educational research documents relevant to a respective search term. Taking into consideration the

European context in which educational research documents are published in different languages, the search engine needed additional multilingual functions so that it was able to deliver results to the search term in several languages. As the EERQI project could only develop a prototype method, four languages as examples: English, German, French and Swedish were included.

After retrieving the documents, their quality could then be assessed with different approaches. This involved the improvement of 'classic' indicators (e.g. amount of citations; classification of a journal). But additionally, the project developed other approaches. Very soon in the research process, the EERQI team decided that quality detection should differentiate between two different types of indicators:

- one type that was external to the text, such as bibliometric and webometric features was called 'extrinsic indicators'; and
- another type that was internal in the text - namely the signals that were given within the words, graphs, metaphors of which the text was composed was called 'intrinsic indicators'.

In order to assist an evaluator (a reader) with the detection of quality, the extrinsic and intrinsic indicators have both to be applied to the text. Thus, tools had to be developed to assist with their easy detection. Measuring extrinsic characteristics of research publications involves the harvesting of relevant pieces of information from different search engines such as Google Scholar, Google Web Search etc. The detection of the intrinsic indicators is a much more complex process. Here, assistance can be provided by automatic semantic analysis, a tool developed by the EERQI team members from Xerox Laboratories in Grenoble, France. Another relevant tool is the EERQI peer review questionnaire. This instrument comprises of operationalised items of the intrinsic indicators that indicate internal features of the quality of a text. The reliability and acceptance of this questionnaire was tested with a positive result within the educational research community; responsible were the team members from University of Hamburg, Germany and the European Educational Research Association. The combination of methods and approaches to assist a reader in the process of quality detection is what we called the EERQI Prototype Framework.

The EERQI Prototype Framework consists of:

- a content base and the search and query engine that support the detection of potential quality via identification of relevant educational research texts in different (electronically available) sources (developed by EERQI partners RRZN, Hannover, and ISN, Oldenburg).
- a tool called 'aMeasure' that identifies extrinsic characteristics of research publications by using Google Scholar, Google Web Search, MetaGer, LibraryThing, Connotea, Mendeley, and citeulike (developed by EERQI partner Humboldt University, Berlin, Germany).
- a linguistic technology that allows for the automatic identification of key sentences to indicate parts of documents to which reviewers should pay particular attention (Automated semantic analysis, developed by EERQI partners Xerox and DIPF).
- a Peer Review Questionnaire that contains a tested operationalisation of the intrinsic indicators that were developed by the EERQI project (developed by University of Hamburg and EERA).

Following section explains in detail examples of the EERQI products.

2. Example No. 1: 'Intrinsic Indicators' and the Peer Review Questionnaire

As already mentioned, the EERQI consortium decided very early in the project's process to distinguish between intrinsic and extrinsic indicators. The hypothesis was that both types of indicators may be relevant for evaluation and assessment processes, but the kind of information wanted on the performance of a text can differ and may or may not be cor- or interrelated. The identification of cor- or interrelations between the two types of indicators belonged to the project's aims. Although the funding period is finished, the data analysis for this purpose is not yet terminated. First attempts show that on a holistic

basis there are only weak ties between both types of indicators. Multilevel analysis however, is still ongoing.

The development of the intrinsic indicators was an iterative process which included several hundred experts that were nominated by European Educational Research Associations. On this basis, a comprehensive set of five generic quality indicators emerged:

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

These indicators were operationalised, again in a consultation process, and transferred into the ‘EERQI Peer Review-Questionnaire’ that was tested in three waves. In each wave, statistical analysis was carried out in order to test the reliability of the items.

The final version of the questionnaire contains three scales with respect to the indicators rigour, originality and significance with all in all 16 items. All test and item characteristics show good to very good values, which could be approved for subsamples of different areas of educational research with different cultural and linguistic backgrounds. The following table illustrates the values for the scales:

Table 1
Overview Final Scale Values

Scale	Subscale	Number of Items	Reliability	Mean Value of Item validity
Rigour		9	.92	.76
	Methods & Approaches	3	.83	.72
	Results	2	.94	.64
	Discussion	4	.90	.82
Originality		3	.91	.78
Significance		4	.91	.78

Especially helpful for the reviewing process were the reviewers’ comments on the relevance of the indicators and the practical use of the questionnaire. The acceptance and indication of usefulness of the questionnaire was generally very high; especially in reference to educational research texts that derive from empirical studies. Moreover, the statistical results show that the questionnaire can well be applied across different areas of educational research. The analysis of the qualitative responses substantiates that the questionnaire includes the most important indicators in the field of quality assessment in educational research publications.

Different suggestions for an extended use of the questionnaire, for instance, a further development for the purpose of training (especially new) researchers in assessment tasks were made. These possibilities should be further explored within the educational research communities in the future.

To sum up, EERQI's intrinsic indicators and their operationalisation as shown in the questionnaire were successfully tested. The acceptance of the instrument in the European educational research community appears to be high according to our investigations. The instrument is now available on the EERQI website for implementation and further development. This will hopefully support the intentions to facilitate and to raise the transparency of assessment processes in educational research, and thus, enhance the quality of quality assessment procedures as such.

3. Example No. 2: Automatic Semantic Analysis

The aim of the EERQI approaches 'automatic semantic analysis' is to assist human evaluators in the time-consuming process of quality detection. A method to do this is 'key sentence extraction'. EERQI-partner XEROX developed a tool that automatically detects and highlights key sentences in educational research articles in English, French, German and Swedish, and we have tested its performance. The aim of key sentence detection is twofold in EERQI: a) providing reading assistance for peer-reviewers, and b) improving the relevance ranking of the EERQI search engine.

(a) Providing Reading Assistance to Peer-reviewers

The approach consists in highlighting salient sentences that provide textual evidence for the peer reviewers in order to back up their evaluation. This means that the tool does not 'evaluate' automatically, but suggests to a reader which are the parts of the texts that are relevant for founding the judgment on. The approach is based on a consensus in the EERQI team according to which an evaluation by peers is supposed to judge 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 inherent in the evaluation that can be carried out with the EERQI Peer Review Questionnaire. Following these evaluation criteria, the tool that was developed highlights key sentences that describe research problems, purposes and conclusions related to the topic of the article as indicated by keywords.

The underlying supposition driving this approach is that by highlighting information, a relevant and coherent representation of the flow of the article is added to it. This complements and completes the representation that is provided by the structure of the article, i.e. title, summary, outline. Such features are sometimes, but not always made explicit in a text. A summary, when present, gives concise information of the overall issues in the article, but is not sufficient for evaluation. Section headings and article structure play an important role for synthesizing the development of the arguments, but in the domain of educational sciences - as in social sciences and humanities in general - they do not follow general patterns and thus are in many cases not indicative of the underlying argumentation.

In order to evaluate the appropriateness and usefulness of the results of this approach, several tests were carried out. One of these tests focused on the time that was needed for evaluation procedures, using exemplary texts with and without highlighted sentences. Our conclusions are the following. Highlighting allows the support of peer reviewers' evaluation process according to the scales of significance, originality and the items for style. Evaluation with respect to the items referring to 'integrity' and the scale for rigour however, can as yet not be supported by the highlighted sentences. Here, further development of the method would be necessary. Another relevant result is that the highlighting makes it possible to rapidly filter out bad quality. In this respect, processing of the highlighted texts took *four times shorter* than the text without highlights.

Another testing (peer review exercise) concerned the question if the highlighted sentences do in fact cover the most relevant contents of the articles. Peer-reviewers who participated in the testing of the questionnaire were asked to briefly summarize the reviewed article in their own words. They were asked

to state the main subject of the reviewed article, the research problem, the main conclusions or results and open questions.

The summaries were written in the same language as the reviewed article. The sentences of these summaries were compared to the highlighted sentences. The following results have been obtained:

- 84% of the human summary sentences correspond to a sentence from the article.
- 56% of the corresponding sentences fulfil the criteria of salient sentences.
- 68% of the salient sentences were detected by XIP, the parser developed by Xerox.
- On average, the automatically highlighted sentences showed four times more of the nouns from the text than occurred in the human summary sentences.

These results show that the automatically detected sentences cover a considerable proportion of human summary sentences. This suggests that automatic highlighting does indeed have the potential of providing key sentences for peer reviewers, and thus allow for less time consuming processes of quality detection.

(b)Enhancing the Ranking of the Search Engine

The assumption here is that if the query term inserted in the EERQI search engine occurs in the highlighted key sentences, the article is more relevant and thus the article gets ranked higher in the search engine's results. In this way, the key sentences are used as snippets in the mechanism that ranks the results of the search engine.

CONCLUSION

In the EERQI project, we also tested other methods in order to unveil their appropriateness for educational research - for example, the method of citation counts that is used in most current approaches for quality assessment. In former evaluations, manifold shortcomings have been uncovered in scientific analyses of the processes and outcomes of this method. EERQI aimed at providing evidence for the invalidity of this underlying assumptions, namely the idea that research communication is merely cumulative in the sense that one piece of research is cited by another in order to build on it for accumulating knowledge. It was one of our approaches to find out whether an alternative use of citations could be a means to overcome the restrictions that are bound to citation counting. The main result of this EERQI approach was that citations in educational research (and other SSH) articles function differently than those in some areas in natural science. Citations may not be motivated by a reference to 'good quality' of the cited text; they may be motivated by the wish to contradict, to criticise - in other words, by the intention to indicate the opposite of good quality. Functions like these could be important factors to be taken into account for the further development of citation indexes. In present methodologies however, one citation counts as one unit independently of its motivation or function; this is obviously an insufficient, if not a misleading information in assessment processes.

The EERQI results as well as all products are explained in more detail on the website www.eerqi.eu and will be published in a volume on "Assessing Quality in European Educational Research – Indicators and Approaches" (working title that will presumably appear in summer 2012. The results were presented at several occasions to the scientific community for verification and acceptance, not only in Europe. These presentations addressed educational research societies, experts in the field, representatives of research funding agencies, and promotion and evaluation bodies at the national and European level. All in all it can be concluded that the aim of the project – the development of a prototype framework that consists of tools which give support to 'human' assessment procedures, but do not replace the human judgment – was met. A special value of the project's processes and results is that we actuated a disciplinary discourse on shared notions of quality as they occur in educational research publications. Moreover, the project could show ways to integrate multilingualism in attempts to identify texts that might be relevant to a researcher who does not master the respective languages. This points to possibilities of strengthening the ties among European educational researchers and allow for more international cooperation and

collaboration, even if the different partner languages are not known to the others. Models of cooperation with one working language, but inclusion of research that is carried out in several different languages, should be better feasible in future. The EERQI project also shows that there is still a lot of work to be done. In the life span of the project, we were able to work out prototypes of the tools and methodologies that we considered to be promising. The research community is now invited to take up the results – the EERQI team welcomes all interested researchers to make use of the data and products that the project developed and advance the approaches.

*1. This paper is based on the EERQI Project's Final report to the European Commission. The project was funded under the EU's 7th Framework, Social Sciences and Humanities Scheme, from 2008 to 2011. See for further information: www.eerqi.eu. A slightly different version of this text appeared in the *Revisita de Investigacion Educativa* 30. 1.2012, pp.13-27.*

Authors

Prof. (Ms.) Ingrid Gogolin, EERQI-Coordinator, Universität Hamburg, Von-Melle-Park 8, 20146 Hamburg, Germany.

Ms. Antje Hansen, Universität Hamburg, Von-Melle-Park 8, 20146 Hamburg, Germany.