Exploring a prototype framework of web-based and peer-reviewed “European Educational Research Quality Indicators” (EERQI)

Exploring a prototype framework of web-based and peer-reviewed "European Educational Research Quality Indicators" (EERQI)

Ton Mooij

Scientometrics (in press) DOI: 10.1007/s11192-014-1429-x
The final publication is available at link.springer.com

Exploring a prototype framework of web-based and peer-reviewed "European Educational Research Quality Indicators" (EERQI)

Abstract Digitization, the Internet, and information or webometric interdisciplinary approaches are affecting the fields of Scientometrics and Library and Information Science (LIS). These new approaches can be used to improve citation-only procedures to estimate the quality and impact of research. A European pilot to explore this potential was called "European Educational Research Quality Indicators" (EERQI, FP7 # 217549). An interdisciplinary consortium was involved from 2008-2011. Different types of indicators were developed to score 171 educational research documents. Extrinsic bibliometric and citation indicators were collected from the Internet for each document; intrinsic indicators reflecting content-based quality were developed and relevant data gathered by peer review. Exploratory and confirmatory factor analysis and structural modeling were used to explore statistical relationships among latent factors or concepts and their indicators. Three intrinsic and two extrinsic latent factors were found to be relevant. Moreover, the more a document was related to a reviewer's own area of research, the higher the score the reviewer gave concerning 1) significance, originality, and consistency, and 2) methodological adequacy. The conclusions are that a prototype EERQI framework has been constructed: intrinsic quality indicators add specific information to extrinsic quality or impact indicators, and vice versa. Also, a problem of "objective" impact scores is that they are based on "subjective" or biased peer-review scores. Peer-review, which is foundational to having a work cited, seems biased and this bias should be controlled or improved by more refined estimates of quality and impact of research. Some suggestions are given and limitations of the pilot are discussed. As the EERQI development approach, instruments, and tools are new, they should be developed further.

Introduction

Innovation in assessment of quality and impact of research

Indicators to characterise the quality or impact of the production of knowledge concerning academic research are important in both Scientometrics and Library and Information Science (LIS) (cf. Barnett, Park, Jiang, Tang, & Aguillo 2014; Cerchiello & Giudici 2014; Hjørland, Nielsen, & Høyrup 2014; Martínez, Herrera, López-Gijón, & Herrera-Viedma 2014). An important development is search engines that automatically use various types of administrative characteristics such as bibliometric characteristics (e.g., author, title of document, title of journal or book, volume number and pages of specific articles). The result may be that articles that use a hyphen or a colon in the title, or with other keywords than the words used in the title, get a higher number of citations (Rostami, Mohammadpoorasl, & Hajizadeh 2014).

Traditionally, the quality of social scientific research has been assessed by "impact indicators" which are based on the Social Science Citation Index (SSCI; cf. Thomson Reuters 2013; see also Campanario 2014). These indicators refer to specific administrative measures of citations in articles in specific scientific journals over a specific period of time (cf. Lee 2011; Selek & Saleh 2014; Su, Deng, & Shen 2014; Williams 2011). Many points of view can then be used to analyse the information available. For example, Fry and Donohue (2014) explore and modify an "Author Affiliation Index" to rank authors, international institutions, and scientific journals in various disciplines. Another example is the use of multilevel network analysis of citations to rank the world's research organisations and universities (Barnett et al. 2014). Fragkiadaki and Evangelidis (2014) distinguish a great many direct and indirect types of citations to characterise papers or documents, authors, and journals. Direct indicators are based on information concerning immediate citations of the unit under scrutiny such as the total number of papers cited or total number of citations. These authors concentrate on indirect citations that consider the direct and indirect impact of a unit's references and citations including both direct and indirect influences over generations of citations.

According to Hjørland (2012), information specialists should be aware that decisions about selection of quality indicators involve consideration of subject-specific theories and evaluation or relevance criteria. Gradmann (2014) further develops this issue by focusing on web-based information and its influence on document creation and publication. His analysis emphasizes the change from a "document" as a container of content, to analysis of its “content”, to identification of its “context” by means of links to information about personal or place names, scientific terms and other characteristics. The result may be a gradual replacement of a document by a dynamic model of richly contextualised aggregations of web resources with new information or semantic characteristics, including new questions and challenges. One of the questions that Gradmann (2014) poses concerns the determination of the quality of these new information developments.

Citation analysis based only on search engines and the use of administrative procedures and operationalisations is also used to indicate the impact or quality of educational research. As stated above, this approach does not imply an interpretation or assessment of the intrinsic quality of the educational research, which can be expressed by content-based criteria such as originality, significance, reliability, validity, or representativeness for a specific population. The reduction of this assessment to extrinsic administrative procedures only is not generally accepted, however (e.g., see Bridges 2009; Centre for Higher Education Research and Information (CHERI) & Hobsons Research 2008; Technopolis 2009a, 2009b). Moreover, such estimation of quality and impact of research runs the risk of increased but unnecessary journal citations and mutual citation within a group of authors publishing in the same journal. Teodorescu and Andrei (2014) investigate “citation circles” in which authors may deliberately cite other researchers for the purpose of inflating bibliometric indicators. This raises questions about the possible manipulation of citations as a means to increase a journal’s impact factor and also the danger of the journal’s isolation from the international scholarly communities (see also Lu, Ajiferuke, & Wolfram in press; Yu, Yu, Li, & Wang in press).

In science in general and in educational research, variation in scientific evaluation criteria and related methodological practices and interpretations has been valued for a long time. Assessing the quality and impact of educational research is not easy, however, because of the variation in scientific evaluation criteria and related methodologies, including differences in data collection practices and interpretation patterns. This diversity is also reflected in differences in priority of specific theoretical or research issues and design procedures across scientific educational journals. Therefore, restricting the focus on administrative operationalisations to assess the quality and impact of educational research has raised questions in the European research community. Initiatives to compensate for this one-sidedness began in the context of the European Educational Research Association (EERA). EERA organises many research networks and functions as a platform to develop or integrate projects (see further http://www.eera-ecer.de/).

The EERQI project

Research initiatives to promote and check the quality and impact of educational research in differentiated ways may produce alternatives to the approach in which this quality is assessed by using administrative or citation operationalisations only. Integration of, or compensation by, other types of indicators are expected to enhance assessment and may also result in improved estimations of the quality and impact of educational research. For these reasons, the European Union (EU) was interested in this type of research. EERA succeeded in getting EU support to conduct a pilot for operationalising and evaluating relevant indicators. This project, 'European Educational Research Quality Indicators' (EERQI), was funded under the Socio-Economic Sciences and Humanities Theme (FP7 # 217549). The project was carried out from 2008 – 2011 and was developed by an interdisciplinary European research consortium of experts from educational science, biblio- and webometrics, information and communication technologies, computational linguistics and publishing houses. The intention was to improve citation-only assessments of the quality and impact of educational and other research (Gogolin, 2008). A research document or paper was chosen as the unit of research (cf. Yu et al. in press).

In EERQI, extrinsic bibliometric and citation indicators were harvested with different search engines such as Google Scholar and Google Web Search (cf. Gradmann, Sieber, & Stoye 2011; Sieber & Stoye 2011). In addition, specific concepts and indicators reflecting the content or intrinsic quality of research documents were distinguished. The first type of intrinsic indicator developed involved estimates of quality that explicated or described aspects of the research such as rigour, originality, significance, integrity, and style (cf. Bridges 2009). This type of information was established by peer review. The second type of intrinsic indicator was to identify internal text signals given within the words, graphs, and metaphors of the text (cf. Sándor & Vorndran 2009). Here assistance was provided by automatic semantic analysis. The EERQI methods and products thus represent various structures, digital facilities, and related practices to connect producers of specific knowledge with users of this knowledge (cf. Dam Christensen 2014). More information about the EERQI project can be found in the final EERQI report, at www.eerqi.eu, and in Gogolin, Åström, and Hansen (2014).

Research question

In line with Alexander (2012, 2014) and Lorentzen (2014), the partners in the EERQI pilot decided that both “subjectivity” and “objectivity” are relevant in the information to be represented by potential EERQI indicators. A main goal of the EERQI project was therefore to collaborate to explore, define, use in practice, and analyse sets of extrinsic and intrinsic quality concepts and their indicators with respect to educational research documents (cf. EERQI project 2010; Mooij 2008; Nolin & Åström 2010).

In addition to the creation of EERQI methods and products, therefore, the statistical relationships between both intrinsic and extrinsic concepts and their indicators had to be explored to construct an initial prototype EERQI framework. The research question to be answered was: How can intrinsic and extrinsic concepts and their indicators be operationalised and what do statistical analyses reveal about the relationships among these concepts and indicators with respect to estimating the quality and impact of educational research? In this article I give an overview of the operationalisation of the two types of concepts, the main development and modeling aspects, and the statistical procedures and their outcomes (cf. also the report of Mooij 2011). Given the data available about intrinsic and extrinsic characteristics of the same documents, however, inclusion of intrinsic concepts had to be restricted to estimates of quality based on peer review.

Operationalisation of concepts and indicators

Intrinsic concepts and indicators

In the EERQI project, intrinsic indicators assumed to indicate the quality of educational research documents were based on the tradition of peer review (see the project documents at www.eerqi.eu; see also Bridges 2009). These concepts and their indicators refer to content-based qualities of educational research documents. They operationalise aspects of documents such as methodology, results, discussion, originality, significance, and validity; see the concepts and their corresponding items in Table 1.

Table 1 about here

The answer alternatives for each item are: ‘not relevant for this text’ (=0), ‘very poor’ (=1), (2), (3), ‘average’ (=4), (5), (6), and ‘excellent’ (=7). A final item, item 21, allows peer reviewers to indicate how closely the document they evaluated relates to their own area of research. Here the answer categories are: ‘Very closely’ (=1), ‘Closely’ (=2), ‘Less closely’ (=3), ‘Not at all’ (=4). The complete set of intrinsic data refers to 177 research documents or articles written by a total of 268 authors.

Peer reviewers scored these documents with respect to all 21 items. These reviewers were partners in the EERQI project or persons who attended the European Conference on Educational Research (ECER) in 2010. Some of the reviewers scored two or more research articles. If available, the scores of various reviewers were aggregated to mean scores per document. However, the data set finally obtained for the present analysis also contained the value 0 (‘not relevant for this text’); this problem could not be avoided because only the aggregated data were available. In univariate analysis using the Statistical Package for the Social Sciences (SPSS, version 17.0), only documents without system-missing values were used. This procedure resulted in item-specific information for 171 documents. Table 1 presents the descriptive statistics of these 21 intrinsic items. The means vary from around 4 (average) to 5; standard deviations vary from 1.05 to 2.03.

Extrinsic concepts and indicators

Extrinsic indicators were web-based and measured aspects of research documents such as number or distribution of citations (per author; across authors; per document; hits resulting from search engines or social network services for a paper or author / combination of authors / title, and so forth; see EERQI project 2011; Gradmann et al. 2011). This webometric information was collected per document, per title, per author, or per title and author (see for detailed information: Gradmann, Haveman, & Oltersdorf 2014). 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. This may imply a bias in favor of multi-author papers, which is comparable to the bias in some other procedures to calculate impact scores. In this respect further improvements are necessary but also rather complicated (cf. Gradmann et al. 2014). The final dataset contained information about 12 extrinsic indicators. Five of these were not included because of too many missing values. Information about the remaining seven extrinsic indicators, their range of scores, means, and standard deviations, is given in Table 2.

Table 2 about here

The variable 'number of citations per paper' [Cit/paper] has a very skewed distribution to the right. This is because the value '0' may reflect 'missing value' or 'no hits'/'no citations'. In this analysis, the latter ('no #') is assumed. The respective scores were 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 then used to explore the relationships between the seven extrinsic variables listed in Table 2. The variables WebMentTitle and MetagerHits are minimally or not related to the other variables. 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 latent factors or concepts: see Table 3.

Table 3 about here

To further explore the relationships between these extrinsic indicators of research documents, the loadings of the five variables on the two factors were rotated (oblique, geomin) within the Exploratory Factor Analysis (EFA) procedure of statistical program

MPlus 6.1 (Muthén and Muthén 1998-2010). The results in Table 4 illustrate that 'Citations per paper (without self-citations)' and 'Web mentions of author in search engine BING' associate to represent a citation concept or latent factor 1, whereas the second concept or latent factor represents numbers of hits by three (other) search engines.

Table 4 about here

Statistical relationships between concepts and indicators

Modeling intrinsic and extrinsic latent factors

Results of a preliminary factor analysis of the intrinsic variables in Table 1 and the information in Table 4 on the extrinsic variables were used to construct a first EERQI measurement model with two intrinsic and two extrinsic concepts or latent factors. A summary of a Confirmatory Factor Analysis (CFA; cf. also Devena, Gay, & Watkins 2013; Mehta & Hull 2013; Mooij & Fettelaar 2013) model to explore the statistical relationships between each concept and the corresponding indicators or observed variables, while taking account of the correlations between the various concepts, shows the following (see also Mooij 2011, 2014).

Latent factor 'Intrinsic1' represents intrinsic indicators reflecting methodological adequacy, completeness and correctness of reporting results, appropriateness of discussion, and originality with respect to methodological procedures; it 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; this factor represents *significance, originality and consistency of the document.* Furthermore, latent factor 'Extrinsic1' refers to the number of citations per document without self-citations and web mentions of the author by search engine BING; it therefore indicates the *number of citations and web mentions by BING.* Latent factor 'Extrinsic2' rather univocally represents the number of hits obtained with search engines Google, CiteULike, and LibraryThing; this factor is associated with the *number of hits in three specific search engines.*

In a next step, the statistical program MPlus (version 6.1) was used to simultaneously check the fit of the measurement model with the four latent factors against the intrinsic scores (Table 1) and the extrinsic scores (Table 2). The outcomes of this model 1 illustrate a rather strong correlation between the two intrinsic factors (0.63) and a weaker correlation between the two extrinsic factors (0.46). The correlation between Intrinsic1 (methodological adequacy of the document) and Extrinsic1 (number of citations and web mentions by BING) is also significant (0.24; $p \leq .05$). This outcome illustrates some overlap between these two concepts and their indicators. Other correlations between intrinsic and extrinsic latent factors are not significant statistically.

Inspection of the modification indices, however, revealed that it was possible to improve this four latent factor model. To explore the statistical consequences, some alternative measurement models were constructed and checked against the model with four latent factors. Alternative model 2a, with the four latent factors, allowed correlation between result indicators Results_1 and Results_2 (items 4 and 5 in Table 1). Alternative model 2b combined intrinsic indicators Results_1, Results_2 and Discussion_2 into a fifth latent factor. This implies the existence of three rather than two intrinsic concepts. An overview of the three alternative models and their statistical Maximum Likelihood (ML) outcomes is given in Table 5.

Table 5 about here

In Table 5, the overall fit of a model is reflected in two statistical indices, the Root Mean Square Error of Approximation (RMSEA) and the Standardized Root Mean Square Residual (SRMR). Both measures are related to the Chi-Square statistic and are influenced by sample size, which implies that a smaller sample results in a less favorable 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 model 2b, with five latent factors, is relatively the best and can be accepted. The measurement model of model 2b is graphically presented in Figure 1.

Figure 1 about here

The relationships between the five latent factors are standardised to facilitate their interpretation as correlations. Correlations between latent factors are free to vary and are represented by two-way arrows between all pairs of latent factors. The regression of each of the indicator variables on their respective latent factor are represented by one-way arrows (per concept only one arrow is given). The total variance of each factor is set to 1. 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. Standardised ML parameter estimates of measurement model 2b in Figure 1 are given in Table 6.

Table 6 about here

The loadings of indicators on specific latent factors in Table 6 are .54 or higher and they are all significant. These explorative results can therefore be accepted for statistical reasons. Taken together, the statistical outcomes in Tables 5 and 6 illustrate that, compared to CFA model 1, CFA model 2b results in a significant improvement in Chi-Square (270.27; df=4; p<.01) and also in acceptable values for RMSEA (0.10) and SRMR (0.08). These results seem plausible and support the notion that both types of latent factors add specific quality information to a prototype EERQI framework. This notion will be checked by a next explorative analysis.

Structural model of intrinsic and extrinsic latent factors

This statistical exploration seeks to assess the effect of peer review on the prototype EERQI framework as given in Figure 1. Explanatory use can be made here of item 21 in Table 1: "'Miscellaneous1': The reviewed article is related to my own area of research". It is hypothesized that the degree to which the reviewed article or document is related to the reviewer's own area of research influences the scores of the three intrinsic concepts or latent factors (cf. also Taylor 2013). Inclusion of this explanatory variable in the CFA model of Figure 1 transforms this model into a causal or structural measurement model as presented in Figure 2. The causal relationships are represented by the three one-sided arrows between item 21 and the three intrinsic latent factors. Furthermore, correlations between the explanatory item and the two extrinsic latent factors are free to vary: see the two-sided arrows in the structural latent factor model in Figure 2.

Figure 2 about here

In Figure 2, the specific indicators for the five latent factors are the same as those in Figure 1. The three intrinsic latent factors are regressed on the explanatory item (Miscellaneous1). The main results of Maximum Likelihood (ML) analysis using MPlus (version 6.1) are given in Table 7.

Table 7 about here

Miscellaneous1 has significant effects on latent factors Intrinsic2 (-0.25; p<.01) and Intrinsic1 (-0.18; p<.01); the effect on Intrinsic3 (-0.13) is non-significant. In a study using a larger sample size the relationship between the third intrinsic factor and Miscellaneous1 could be significant too, however. The present significant results mean that the more closely the reviewed document is related to the reviewer's own area of research, the higher the reviewer's evaluation scores are with respect to significance, originality and consistency (Intrinsic2) and also concerning methodological adequacy (Intrinsic1). These two effects seem to reflect some subjective evaluation bias that may occur in peer review. Table 7 shows no significant statistical relationships between Miscellaneous1 and the extrinsic latent factors.

Discussion

Towards an initial EERQI prototype framework of concepts and indicators

Project EERQI (FP7 # 217549) aimed to operationalise and explore 'European Educational Research Quality Indicators'. In line with major developments in

Scientometrics and LIS as sketched in the introduction section, EERQI combined experts from different fields such as educational science, bibliometrics and webometrics, information and communication technologies, computational linguistics, and publishing houses. The goal was to collaborate and improve citation-only assessments of the quality and impact of educational research documents in particular. The operationalisation used extrinsic bibliometric and citation indicators generated by different search engines. Furthermore, intrinsic or content-based quality of research documents was assessed by evaluation aspects like rigour, originality, significance, integrity, and style; these data were collected by peer review. In addition to the creation and piloting of various instruments, tools, and procedures, one main aim of the EERQI project was to analyse statistical relationships between the extrinsic and intrinsic concepts and their indicators. To this end, the project provided different types of extrinsic and intrinsic data reflecting estimations of the quality and impact of educational research documents.

To try to construct a first prototype of the EERQI framework, statistical relationships between these indicators were explored with respect to 171 educational research documents. The available data set allowed the checking of the relevance of extrinsic concepts and their indicators and intrinsic concepts and their indicators based on peer review; it was not possible to include data of intrinsic indicators based on automatic semantic analysis. Consecutive measurement models and their empirical analysis results confirmed the statistical relevance and functionality of both intrinsic concepts or latent factors and extrinsic concepts or latent factors, and their respective indicators, in a comprehensive EERQI framework. A further statistical check was to determine whether, as expected, the degree to which a reviewed article is related to a reviewer's own area of research influenced the scores of the intrinsic concepts. Empirical testing in a causal structural model (see Figure 2) indeed revealed that the more the reviewed document was related to the reviewer's own area of research, the higher the reviewer's evaluation scores were 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 main relevance of Figures 1 and 2 for determining the quality and impact of educational research documents is that the use of intrinsic indicators seems to add specific quality information to an EERQI framework consisting of extrinsic indicators only, which is the traditional situation. On the other hand, the use of extrinsic indicators may add quality and impact information to a framework consisting of intrinsic indicators only. Moreover, intrinsic indicators as used in peer review should be corrected for reviewer bias (see Figure 2 and Table 7) in order to result in valid information.

Interpretation of EERQI results in Scientometrics and LIS

The EERQI results can be interpreted with respect to actual and potential developments of Scientometrics and LIS. For example, the present results support the relevance of both “objectivity” and “subjectivity” in information taxonomies, as has been claimed by Alexander (2012, 2014). The EERQI findings of Figure 2 and Table 7 furthermore align with those of Taylor (2013), who finds a strong statistical association between work task and criteria used to judge relevance. This researcher concludes that work task criteria may influence the relevance judgment process, which can be used to provide suggestions for the improvement of information retrieval systems and information literacy efforts. Corresponding with Olsen, Lund, Ellingsen, and Hartvigsen (2012), the EERQI pilot combined two traditionally different research approaches in a broad “document conceptual model” and, in addition, made some first empirical steps in this new area. The EERQI pilot result can therefore also be conceived as an example of the new research infrastructures built around different data materials, according to coordinated criteria (see also Finnemann 2014). In doing this, the pilot result may offer some assistance to the solution of some of the problems of data curating from the perspective of research libraries (see further Nielsen & Hjørland 2014).

Furthermore, the contextual web-based strategy of Gradmann (2014) may be promising (cf. also Locoro, David, & Euzenat 2014; Lorentzen 2014), in particular if combined with indicators and data based on automatic semantic analysis as developed in the EERQI project by Sándor and Vorndran (2009). These researchers base intrinsic indicators in internal text signals that are revealed in words, graphs, and metaphors of which text is composed. Sándor and Vorndran (2014a, 2014b) further inform about their natural language processing system highlighting salient sentences or parts thereof that convey the most important threads of an article's content. Their aim is to assist reviewers to focus their activity on the semantic information (e.g., problems, aims, definitions, and results) and the design of the argumentation in an article. Their system is implemented in the "Xerox Incremental Parser" and works with texts in English, French, German, and Swedish. This automatic semantic analysis could be developed further and corresponding data could be included to extend the statistical analyses as conducted in Figures 1 and 2.

Working according to a contextual web-based strategy, Eccles, Thelwall and Meyer (2012) report webometric research as the only source of information to assess the "impact factor" of academic communications and to analyse the impact of online digital libraries and digital scholarly images. These researchers experience some success; however, the impact of digitised resources based on shifting URLs or integrated in large online resources is difficult to measure (see also Gradmann et al. 2014). A webometric-only approach thus needs further investigation and could probably profit from the experiences reported in the EERQI approach by using both webometric and peer review or other indicators. Another suggestion can be based on Lorentzen (2014), who investigated webometrics and web mining. Webometrics explores structures on the web by quantitative analysis of information structures, whereas web mining concentrates on web content and usage by developing methods and algorithms. His conclusion was that research can benefit from collaboration between the webometricians involved in both research fields.

Conclusions

The differentiated conceptual structure of concepts and indicators in Figures 1 and 2, the explorative statistical outcomes in Tables 5 – 7, and the interpretation of those results in terms of Scientometric and LIS developments all support the plausibility, reliability and potential validity of the initial EERQI conceptual framework presented here. A first conclusion then is that a prototype EERQI framework has been constructed. The main goal of this project – to improve citation-only assessments of the quality or impact of educational and other research – thus has been realized, but it is also evident that more research is necessary.

A second conclusion is that the present results are important because the peer review process is the general procedure to judge research documents before these are accepted, or not accepted, by a scientific journal. Peer-review is basic to being cited and EERQI presents evidence that evaluation by peers is biased. A main problem of “objective” impact scores thus is that they are based on “subjective” or biased peer-review scores. From a scientific point of view, this bias should be controlled or be improved by more refined estimates of quality and impact of research, including other “objective” estimates.

Third, the present EERQI framework illustrates various potential improvements that could be realised by further development and practice of EERQI in comparison with the usual SSCI- or Thomson-index in scientific research. For example, as outlined in the above section, it seems worthwhile to introduce intrinsic indicators based on automatic semantic analysis into this prototype framework, either by extending the present dataset and analysis results or by designing new research. Moreover, in the introduction section many other developments were shown to take place concerning extrinsic indicators in particular. It seems that much research will be needed to come to a more or less satisfying approach in combining intrinsic and extrinsic concepts and their indicators. Statistical analyses and checks could be performed in a way similar to those described in this paper, to continue enhancing the process of measuring the quality and impact of research documents.

Limitations and improvements

In addition to the creation of different methods, tools, and products, EERQI also taught some lessons and provided information on limitations which follow-up development or analysis should seek to eliminate. These lessons include the following experiences.

1. The present pilot covered a fairly small yet exemplary number of research documents and reviewers. Future research should increase the numbers of documents and reviewers included and also make use of concepts, indicators, and data based on automatic semantic analysis of research documents (cf. Locoro et al. 2014; Sándor & Vorndran 2014a, 2014b).

2. Future research should also pay close attention to details of the assessment of indicators, improve the assessment of interobserver reliability of reviewed information, and take the possible effects of language differences into account.
3. Furthermore, the relationship between the prototype EERQI framework and its concrete use to improve the scoring of both the quality and impact of research documents should be given more attention.
4. Finally, the EERQI pilot put forth a number of tools that can be either taken up and further developed by (other) research teams or applied to next assessment processes in Scientometrics and LIS.

Acknowledgements

The pilot 'European Educational Research Quality Indicators' (EERQI) was EU funded under the Socio-Economic Sciences and Humanities Theme (FP7 # 217549). The project was developed and conducted by an interdisciplinary European research consortium. I want to thank all partners for their collaboration. Moreover, I am grateful to Daan Fettelaar MSc for his assistance in performing the necessary analyses and to Dr. Maureen Snow Andrade for her check of the appropriate use of English language.

References

Alexander, F. (2012). Assessing information taxonomies using epistemology and the sociology of science. *Journal of Documentation, 68*(5), 725-743.

Alexander, F. (2014). Devising a framework for assessing the subjectivity and objectivity of information taxonomy projects. *Journal of Documentation, 70*(1), 4-24.

Barnett, G.A., Park, H.W., Jiang, K., Tang, C., & Aguillo, I.F. (2014). A multi-level network analysis of web-citations among the world's universities. *Scientometrics, 99, 5*–26.

Bridges, D. (2009). Research Quality Assessment: Impossible science, possible art? *British Educational Research Journal, 35*(4), 497-517.

Campanario, J.M. (2014). The effect of citations on the significance of decimal places in the computation of journal impact factors. *Scientometrics, 99,* 289–298.

Centre for Higher Education Research and Information (CHERI) and Hobsons Research (2008). *Counting What is Measured or Measuring What Counts? League tables and their impact on higher education institutions in England. Report to HEFCE.* http://www.hefce.ac.uk/data/year/2008/countingwhatismeasuredormeasuringwhatcounts/ Accessed 25 March 2014.

Cerchiello, P., & Giudici, P. (2014). On a statistical *h* index. *Scientometrics, 99,* 299–312.

Dam Christensen, H. (2014). The framing of scientific domains: About UNISIST, domain analysis and art history. *Journal of Documentation, 70*(2), 261-281.

Devena, S.E., Gay, C.E., & Watkins, M.W. (2013). Confirmatory Factor Analysis of the WISC-IV in a Hospital Referral Sample. *Journal of Psychoeducational Assessment, 31* (6), 591-599.

EERQI project (2010). *State of the Art Reports on EERQI Project Parts. Preparatory meeting for the second EERQI workshop, 18-19 March.* http://eerqi.eu/de/page/final-report Accessed 11 January 2011.

EERQI project (2011). *European Educational Research Quality Indicators. Project final report*. http://eerqi.eu/de/page/final-report Accessed 23 June 2012.

Eccles, K.E., Thelwall, M., & Meyer, E.T. (2012). Measuring the web impact of digitised scholarly resources. *Journal of Documentation, 68*(4), 512-526.

Finnemann, N.O. (2014). Research libraries and the internet: On the transformative dynamic between institutions and digital media. *Journal of Documentation, 70*(2), 202-220.

Fragkiadaki, E., & Evangelidis. G. (2014). Review of the indirect citations paradigm: Theory and practice of the assessment of papers, authors and journals. *Scientometrics, 99,* 261–288.

Fry, T.D., & Donohue, J.M. (2014). Exploring the Author Affiliation Index. *Scientometrics, 98,* 1647–1667.

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

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

Gradmann, S. (2014). From containers to content to context: The changing role of libraries in eScience and eScholarship. *Journal of Documentation, 70*(2), 241-260.

Gradmann, S., Haveman, F., & Oltersdorf, J. (2014). Studies in correlative assessing of intrinsic and extrinsic indicators of quality. In I. Gogolin, F. Åström, & A. Hansen, *Assessing quality in European Educational Research. Indicators and approaches* (pp. 60-84). Wiesbaden, Germany: Springer VS.

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

Hjørland, B. (2012). Is classification necessary after Google? *Journal of Documentation, 68*(3), 299-317.

Hjørland, B., Nielsen, H.J., & Høyrup, H. (2014). Introduction to the special issue: Perspectives on research libraries. *Journal of Documentation, 70*(2), 198-201.

Lee, C.A. (2011). A framework for contextual information in digital collections. *Journal of Documentation, 67*(1), 95-143.

Locoro, A., David, J., & Euzenat, J. (2014). Context-based matching: Design of a flexible framework and experiment. *Journal on Data Semantics, 3*(1), 25–46.

Lorentzen, D.G. (2014). Webometrics benefitting from web mining? An investigation of methods and applications of two research fields. *Scientometrics, 99,* 409–445.

Lu, K., Ajiferuke, I., & Wolfram, D. (in press). Extending citer analysis to journal impact evaluation. *Scientometrics.*

Martínez, M.A., Herrera, M., López-Gijón, J., & Herrera-Viedma, E. (2014). H-Classics: Characterizing the concept of citation classics through H-index. *Scientometrics, 98,* 1971–1983.

Mehta, V., & Hull, D.M. (2013). Structural validity of the professional development profile of the LoTi digital-age survey. *Journal of Psychoeducational Assessment, 31*(1), 61-71.

Mooij, T. (2008). *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 presented at the final conference of the international collaborative project ‘European Educational Research Quality Indicators’ (EU Project 217549; FP7) in Brussels, University Foundation, 15-16 March 2011. Nijmegen, The Netherlands: Radboud University, ITS.

Mooij, T. (2014). A prototype empirical framework of intrinsic and extrinsic EERQI indicators. In I. Gogolin, F. Åström, & A. Hansen, *Assessing quality in European Educational Research. Indicators and approaches* (pp. 121-138). Wiesbaden, Germany: Springer VS.

Mooij, T., & Fettelaar, D. (2013). School and pupil effects on secondary pupils’ feelings of safety in school, around school, and at home. *Journal of Interpersonal Violence, 28*(6), 1240-1265.

Muthén, L.K., & Muthén, B.O. (1998-2010). *Mplus User’s Guide. Sixth Edition.* Los Angeles, CA: Muthén & Muthén.

Nielsen, H.J., & Hjørland, B. (2014). Curating research data: The potential roles of libraries and information professionals. *Journal of Documentation, 70*(2), 221-240.

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

Olsen, B.I., Lund, N.W., Ellingsen, G., & Hartvigsen, G. (2012). Document theory for the design of socio-technical systems: A document model as ontology of human expression. *Journal of Documentation, 68*(1), 100-126.

Rostami , F., Mohammadpoorasl , S., & Hajizadeh , M. (2014). The effect of characteristics of title on citation rates of articles. *Scientometrics, 98,* 2007–2010.

Sándor, Á., & Vorndran, A. (2009). *Detecting key sentences for automatic assistance in peer reviewing research articles in educational sciences.* http://dl.acm.org/citation.cfm?id=1699757 Accessed 10 March 2011.

Sándor, Á., & Vorndran, A. (2014a). Highlighting salient sentences for reading assistance. In I. Gogolin, F. Åström, & A. Hansen, *Assessing quality in European Educational Research. Indicators and approaches* (pp. 43-55). Wiesbaden, Germany: Springer VS.

Sándor, Á., & Vorndran, A. (2014b). Enhancing relevance ranking of the EERQI search engine. In I. Gogolin, F. Åström, & A. Hansen, *Assessing quality in European Educational Research. Indicators and approaches* (pp. 56-59). Wiesbaden, Germany: Springer VS.

Selek, S., & Saleh, A. (2014). Use of *h* index and *g* index for American academic psychiatry. *Scientometrics, 99,* 541–548.

Sieber, J., & Stoye, D. (2011). *Description of aMeasure.* Berlin: Humboldt-Universität, Institut für Bibliotheks- und Informationswissenschaft. http://edoc.hu-berlin.de/oa/reports/reJ3Xv4PJ82ZM/PDF/29B6vgnyGba6.pdf Accessed 27 June 2012.

Su, X., Deng, S., & Shen, S. (2014). The design and application value of the Chinese Social Science Citation Index. *Scientometrics, 98,* 1567–1582.

Taylor, A. (2013). Examination of work task and criteria choices for the relevance judgment process. *Journal of Documentation, 69*(4), 523-544.

Technopolis (2009a). *Identification and dissemination of lessons learned by institutions participating in the Research Excellence Framework (REF) bibliometrics pilot. Results of the round one consultation – report to HEFCE.* http://www.hefce.ac.uk/data/year/2009/identificationanddisseminationoflessonslearnedbyinstitutionsparticipatingintheresearchexcellenceframeworkrefbibliometricspilotresultsoftheroundoneconsultation/ Accessed 21 February 2012.

Technopolis (2009b). *Identification and dissemination of lessons learned by institutions participating in the Research Excellence Framework (REF) bibliometrics pilot. Results of the round two consultation – report to HEFCE.* http://www.hefce.ac.uk/data/year/2009/identificationanddisseminationoflessonslearnedbyinstitutionsparticipatingintheresearchexcellenceframeworkrefbibliometricspilotresultsoftheroundtwoconsultation/ Accessed 21 February 2012.

Teodorescu, D., & Andrei, T. (2014). An examination of "citation circles" for social sciences journals in Eastern European countries. *Scientometrics, 99,* 209–231.

Thomson Reuters (2013). *Journal Citation Reports®.* http://thomsonreuters.com/products_services/science/science_products/a-z/journal_citation_reports/ Accessed 20 June 2013.

Williams, R.B. (2011). Citation systems in the biosciences: A history, classification and descriptive terminology. *Journal of Documentation, 67*(6), 995-1014.

Yu, T., Yu, G., Li, P.-Y., & Wang, L. (in press). Citation impact prediction for scientific papers using stepwise regression analysis. *Scientometrics.*

Table 1. Concepts and items assessing intrinsic quality (n research documents=171)*

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

* Items 1 – 20 range from 0 – 7; item 21 has range 1 – 4 (see text).

Table 2. Variables assessing extrinsic quality (n research 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 WebMentAuth	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 hits matching the author's name	.00	3265.00	219.91	448.85
5 MetagerHits	Metager hits matching the author's name	.00	133.00	4.74	16.16
6 CiteULikeHits	CiteULike hits matching the author's name and title of article	.00	486.00	21.32	60.55
7 LibraryThingHits	LibraryThing hits matching the author's name and title of article	.00	651.00	29.34	89.95

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

Factor	Eigenvalue	% of Variance	Cumulative %
1	2.61	52.24	52.24
2	1.15	23.04	75.28
3	.61	12.13	87.40
4	.44	8.88	96.28
5	.19	3.72	100.00

* These variables are: Cit/paper, WebMentAuth, GoogleHits, CiteULikeHits, LibraryThingHits.

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

Variable name	Description	Factor 1	Factor 2
Cit/paper (sqrt)	Citations per paper without self-citations using the full title of the article	0.92	-0.00
WebMentAuth	Web mentions of author in search engine BING; number of URL's of pages matching the query submitted	0.41	0.10
GoogleHits	Google Web Search hits matching the author's name	0.02	0.95
CiteULikeHits	CiteULike hits matching the author's name and title of article	0.00	0.69
LibraryThingHits	LibraryThing hits matching the author's name and title of article	-0.11	0.87

Table 5. Comparison of different Confirmatory Factor Analysis models

Alternative measurement models	X^2	**df**	**RMSEA**	**SRMR**
1. Model with 4 latent factors (2 intrinsic, 2 extrinsic)	1028.66	269	0.13	0.07
2a. As Model 1, but with error covariat. Result_1 - Result_2	785.23	268	0.11	0.07
2b. Model with 5 latent factors (3 intrinsic, 2 extrinsic; Fig. 1)	758.39	265	0.10	0.08

Table 6. Maximum Likelihood parameter estimates (standardised) of the measurement model in Figure 1

	Factor loadings					
	INTRINS1: Methodol. adequacy	**INTRINS2: Sign./orig./ consist.**	**INTRINS3: Results**	**EXTRINS1: # cit./Web BING**	**EXTRINS2: Hits 3 search**	**R2**
Methods_1	0.91**					0.82**
Methods_2	0.86**					0.74**
Methods_3	0.91**					0.84**
Results_1			0.97**			0.94**
Results_2			0.98**			0.95**
Discussion_1	0.88**					0.78**
Discussion_2			0.79**			0.62**
Discussion_3		0.66**				0.43**
Discussion_4		0.61**				0.37**
Originality_1	0.79**					0.62**
Originality_2		0.80**				0.63**
Originality_3		0.87**				0.76**
Significance_1		0.90**				0.81**
Significance_2		0.91**				0.83**
Significance_3		0.81**				0.66**
Significance_4		0.72**				0.52**
Validity_1		0.54**				0.29**
Validity_2		0.79**				0.62**
Validity_3		0.84**				0.71**
Miscellaneous2		0.84**				0.71**
Cit/paper (sqrt)				0.59**		0.35**
WebMentAuth				0.69**		0.47**
GoogleHits					0.980**	0.96**
CiteULikeHits					0.674**	0.46**
LibraryThingHits					0.803**	0.65**

	Factor covariances (correlations)			
	INTRINSIC1	INTRINSIC2	INTRINSIC3	EXTRINSIC1
INTRINSIC2	0.62**			
INTRINSIC3	0.74**	0.48**		
EXTRINSIC1	0.24	0.15	0.19	
EXTRINSIC2	0.15	0.09	0.11	0.46**

Fit indices: X^2 (265)=758.385 (p=0.000); RMSEA=0.104; SRMR= 0.077.

* 0.01≤p≤0.05; ** p<0.01.

Table 7. Maximum Likelihood factor parameter estimates (standardised) of the structural model in Figure 2

	Factor covariances (correlations)				
	INTRINSIC1	INTRINSIC2	INTRINSIC3	EXTRINSIC	
INTRINSIC2	0.60**				
INTRINSIC3	0.74**	0.46**			
EXTRINSIC1	0.25	0.16	0.20		
EXTRINSIC2	0.15	0.09	0.11	0.46**	
	Direct effects				
	INTRINSIC1	INTRINSIC2	INTRINSIC3	EXTRINSIC	EXTRINSIC2
Miscellaneous1	-0.18*	-0.25**	-0.13		
	Correlations				
Miscellaneous1				0.03	0.02

Fit indices: X^2 (284)=779.56 (p=0.00); RMSEA=0.10; SRMR= 0.08.

* 0.01≤p≤0.05; ** p<0.01.

Fig. 1. Graphic presentation of CFA measurement model (five concepts or latent factors)

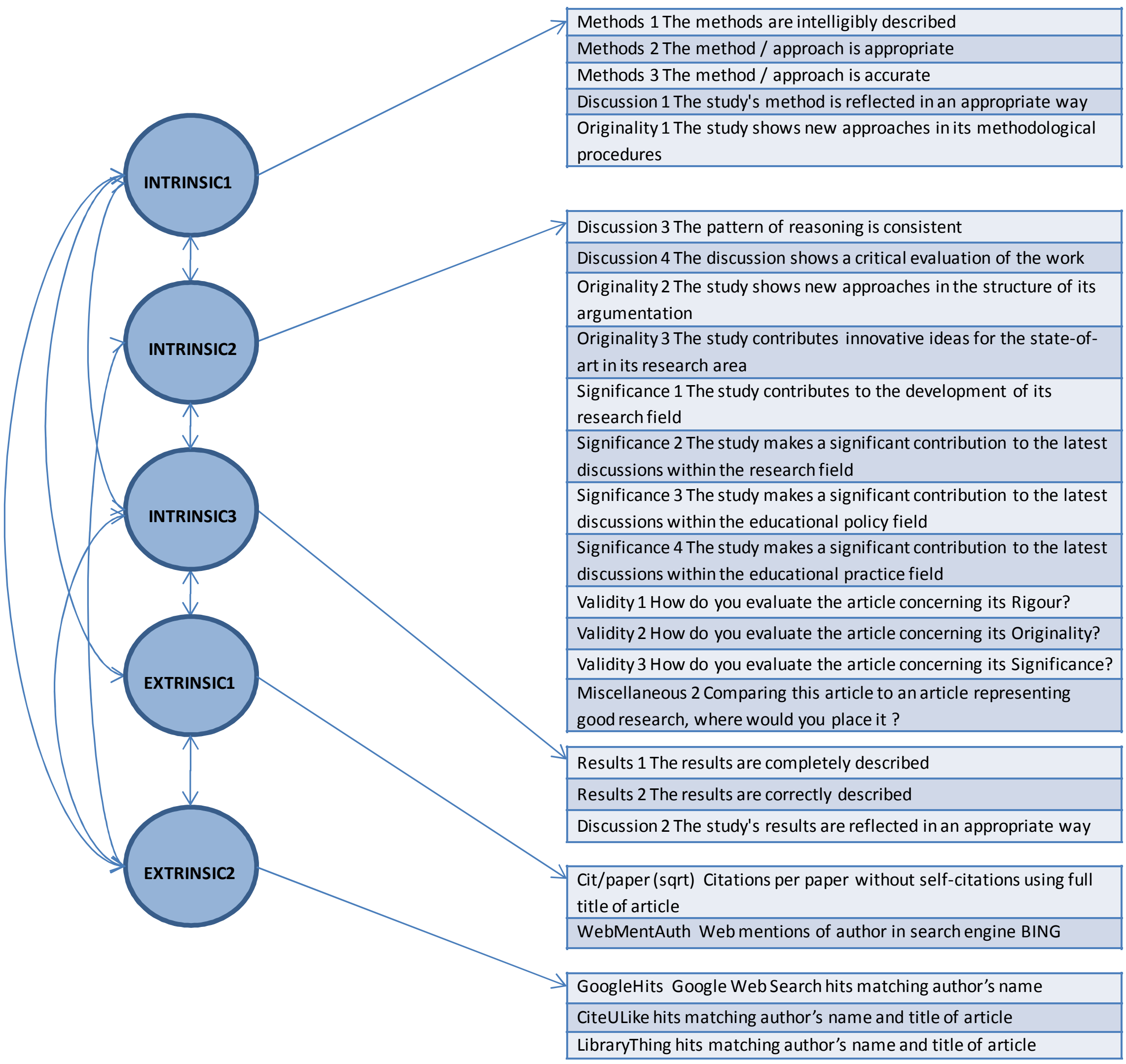

Fig. 2. Structural model of peer review effect on three intrinsic and two extrinsic concepts or latent factors (indicators: see Figure 1)