A project funded under the Socio-Economic Sciences and Humanities Theme (SSH)

European Educational Research Quality Indicators (EERQI): A first prototype framework of intrinsic and extrinsic indicators

Paper for the final EERQI conference, 15-16 March 2011

University Foundation
Rue d'Egmont 11
1000 Brussels

Prof. Dr. Ton Mooij

10th March 2011

Radboud University, ITS (t.mooij@its.ru.nl)
Open University of the Netherlands, Celstec (ton.mooij@ou.nl)

Abstract

The goal of the international project 'European Educational Research Quality Indicators' (EERQI) is to improve upon citation-only based assessments of the quality or impact of educational and other research. One of the project activities is to distinguish 'objective' bibliometric and citation indicators, or 'extrinsic indicators' of research documents, from 'subjective' indicators reflecting 'intrinsic' qualities like rigour, originality, significance, integrity, and style. Different pilots were conducted to collect data with respect to these two types of indicators. In addition to other presentations at the final conference of the project, the goal of this paper is to analyse and report about some of the data gathered. Despite the lack of time and methodological problems of the data collected, two research questions will be answered:

1) What are the results of applying the 'intrinsic criteria' rigour, originality, significance, integrity, and style, in a peer review evaluation of educational research articles?
2) What are the results of analysing the relationships between both intrinsic and extrinsic indicators, to construct an example of a prototype EERQI framework?

In a first pilot, 78 peer reviewers assessed 117 research documents according to rigour, originality, significance, integrity, style, and miscellaneous. Results of principal components analysis first of all reveal an evaluation dimension reflecting overall significance, originality, and scientific relevance in theoretical, policy and practical points of view. A second component points at the consistency and logically communicative structure of the document. A third evaluation dimension mainly accentuates ethical aspects. These outcomes encourage the approach to assess intrinsic quality indicators by peer reviewing.

In a second pilot, peer reviewers evaluated 20 intrinsic indicators that operationalise the concepts methodology, results, discussion, originality, significance, validity, and miscellaneous. In item 21 a peer reviewer completes information about the degree to which the document evaluated was related to the own area of research. The dataset of 2nd March 2011 consists of 177 research documents or articles written by a total of 268 authors. Extrinsic indicators of the same research documents contain various types of web-based information.

A first measurement model was constructed with two intrinsic latent factors and two extrinsic latent factors. The intrinsic factors and extrinsic factors are correlated significantly. The correlation between latent factor Intrinsic1 ('methodological adequacy of the document') and latent factor Extrinsic1 ('number of citations and Web mentions by BING') is also significant. This outcome illustrates some overlap between intrinsic and extrinsic indicators which deserves more attention for reasons of both interpretation and modeling of EERQI.

Exploration of possible improvements of the model resulted in a second measurement model with three intrinsic and two extrinsic latent factors. Significant correlations exist between the intrinsic and between the extrinsic factors; no significant correlations are found between intrinsic and extrinsic factors. This empirical outcome supports the notion that use of intrinsic indicators may add specific quality information to EERQI consisting of extrinsic indicators only; on the other hand, introduction of extrinsic indicators may add specific quality information to EERQI containing intrinsic indicators only.

A final exploration concerns the hypothesis that the degree to which the reviewed document is related to the reviewer's own area of research, will influence the score of intrinsic latent factors. Empirical testing of a structural model reveals that, the more the reviewed document is related to the reviewer's own area of research, the higher the document is evaluated by the reviewer with respect to 1) significance, originality and consistency and 2) methodological adequacy. No relationships exist between the reviewer's own area of research and the two extrinsic factors. The effects on two intrinsic factors seem to reflect some subjective evaluation bias that may take place in peer reviewing. These different outcomes seem to support the validity of both the conceptual framework and the empirical research.

It is concluded that an example of a prototype EERQI framework can be constructed. This is in line with the main goal of the EERQI project: To improve upon citation-only based assessments of the quality or impact of educational and other research. Moreover, present learning experiences can support future EERQI methodological, developmental and research activities that can also integrate semantic latent factors and indicators.

1. Introduction

For many years, impact indicators in educational research like the Social Science Citation Index (SSCI) were based on measurement of citations in publications or documents in specific scientific journals. Nowadays, search engines automatically use various types of 'objective' or 'extrinsic characteristics' e.g., bibliometric or semantic characteristics of publications or documents that may be located in many different web-based sources. The goal of the international project 'European Educational Research Quality Indicators' (EERQI; FP7 # 217549) is to improve upon citation-only based assessments of the quality or impact of educational and other research (cf. Gogolin, 2008). The ambition of the project is to (1) develop and test new methods and sets of indicators for quality assessment of publications in the field of educational and social science, and (2) extend the content base of publications open for quality assessment in this field, with regard to type of texts and publication languages (EERQI project, 2010).

To realise these ambitions different activities were undertaken (see for an overview EERQI project, 2010). Attempts were made to develop and conduct automatic semantic analyses and specific bibliometric and citation analyses on research papers (cf. Sándor & Vorndran, 2009). These activities occurred in relationship to the development of a search engine (Sieber & Stoye, 2011) to harvest relevant information in the Internet (cf. also Nolin & Åström, 2010). Also, in addition to bibliometric and citation indicators (Gradmann, Sieber, & Stoye, 2011), specific indicators reflecting the more 'subjective' or 'intrinsic' quality of research documents were distinguished. The intrinsic indicators refer to the content of a publication or research document and are supposed to explicate or describe aspects like rigour, originality, significance, integrity, and style (cf. Bridges, 2009).

In former EERQI papers (Mooij, 2008a, 2008b) I sketched a possible approach to empirically explore and statistically analyse the relationships between sets of intrinsic and extrinsic indicators of the quality of research documents. In this paper I will analyse both intrinsic and extrinsic indicators and their relationships, to construct a prototype framework of EERQI. In the project, different pilots were conducted to collect data with respect to these two types of indicators. My goal is to use some of the data gathered to answer twoe research questions:

1) Pilot 1: What are the results of applying the criteria rigour, originality, significance, integrity, and style, in a peer review evaluation of educational research articles?
2) Pilot 2: What are the results of analysing the relationships between both intrinsic and extrinsic indicators, to construct an example of a prototype EERQI framework?

Data analyses to answer these research questions were limited in scope and time available. It was for example not possible to rely on a complete account of data collection and preparation steps that were carried out in executing the pilots. Notwithstanding this situation, the measurement models and their statistical outcomes illustrate in exemplary ways that the goal to improve upon citation-only based assessments of the quality or impact of educational and other research seems realistic and can be achieved. In addition, the present statistical approaches and outcomes can be compared with methodological procedures and results of other approaches used in the EERQI project. A thorough comparison of different methodological and analysis approaches, and an adequate evaluation of their potentially different outcomes, will pave the optimal road to future developments of EERQI.

2. Pilot 1: Intrinsic quality indicators

2.1 Specification

In 2010, a first pilot concerned the assessment of intrinsic indicators by peer reviewers. The items used to concretise the concepts were formulated by the EERQI project partners. A total of 78 peer reviewers evaluated and scored 117 research documents according to rigour,

originality, significance, integrity, style, and miscellaneous.[1] Table 1 presents the concepts, items used with respect to each concept, and the results of univariate analysis of the items. The Social Package for the Social Sciences (SPSS; version 17.0) was used in the empirical analysis of the data. The results show that the item means usually vary around 2 ('agree'), whereas the standard deviation (SD) ranges from .49 to 1.07.

Table 1 – Concepts and items to assess intrinsic quality (Pilot 1; n documents=117)

	N	Mean	SD
Rigour*			
The study uses an appropriate methodology and method in a careful and thorough manner.	113	2.08	.75
The conclusion is justified by the evidence in the study.	113	1.99	.69
The argumentation in the study is clear, coherent and internally consistent.	112	1.92	.81
The study is reflexive about its own limitations.	108	2.47	.81
Originality*			
The study shows awareness of previous work in the field and makes its own contribution clear (in relation to content, methodology, and findings).	112	2.12	.80
The study is original, creative or innovative in a significant (i.e. non-trivial) way.	109	2.28	.69
Significance*			
The theoretical and/or practical context of the study is made clear.	115	2.04	.81
The work makes a *significant* contribution to educational research, theory, policy or practice.	113	2.36	.86
Integrity*			
It appears that the study is genuinely the work of the named author (i.e. evidence of plagiarism or unacknowledged derivativeness would be a counter indicator of quality).	111	1.53	.52
The study has respected (where applicable) the ethical principles which normally operate in this field.	98	1.67	.49
Style*			
The title of the article fits with the contents.	114	1.73	.77
The study communicates in an appropriate way.	113	1.82	.68
The study is well (appropriately) organised.	110	1.98	.79
It is a pleasure to read the study.	111	2.09	.86
Miscellaneous			
The reviewed article is related to my own area of research.**	114	2.80	.96
In case this peer review of the study would be used to decide upon publication/non-publication in a scientific journal, would the reviewed study be accepted?***	112	2.16	1.07

* Scored as: I completely agree=1; I agree=2; I disagree=3; I completely disagree=4.
** Scored as: very closely=1; closely=2; less closely=3; not=4.
*** Scored as: yes=1; yes, with *minor* amendments=2; yes, with *major* amendments=3; no=4.

2.2 Principal components analysis

Principal components analysis was used to explore the covariation in the scoring of all 16 intrinsic quality items. Some analysis results are given in Table 2.

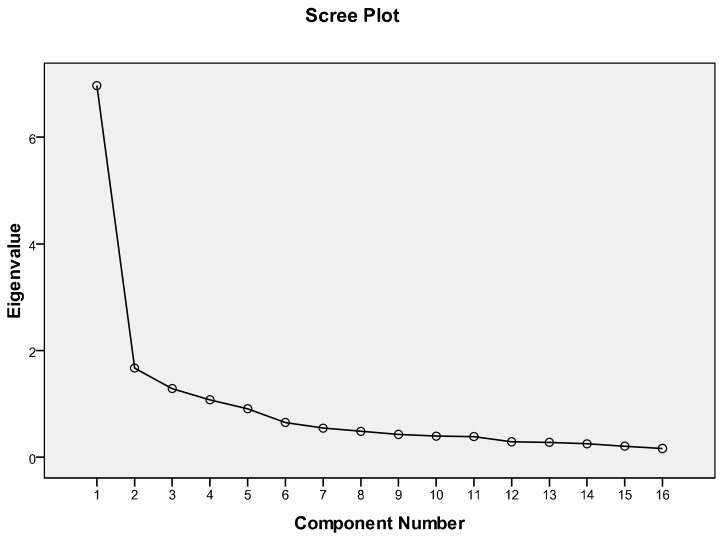
Table 2 – First principal components of judgements of intrinsic quality (Pilot 1)

Principal component	Initial Eigenvalues	% of variance	Cumulative %
1	6.97	43.54	43.54
2	1.67	10.45	53.99
3	1.29	8.04	62.03
4	1.08	6.73	68.76

The results of Table 2 emphasize the existence of one main principal component reflecting one important dimension of evaluation (see the respective percentages of variance). This result is shown also in the graph of the Scree test: see Figure 1.

[1] The research documents represent three different European languages. The data set contains scores of peer reviewers who are partner in the EERQI project. Some of the reviewers scored two or more research documents. In combination with the low number of reviewers, this differentiated data structure does not facilitate the assessment of interobserver reliability or multilevel analyses between and within languages and / or reviewers, respectively.

Figure 1 –Scree test results with respect to judgements of intrinsic quality (Pilot 1)



In addition, the relative importance of the intrinsic items with respect to the first four components is shown in their loadings on these components: see Table 3. These results indicate the existence of one huge overall evaluation criterion regarding 15 items (excluding the item about the relatedness of the reviewed article to the reviewer's area of research). This one-dimensional evaluation domain was confirmed by Alpha scale analysis with respect to the 15 items (using listwise deletion of missing values; n documents reviewed=81). The results reflect a good overall scale (Cronbach's Alpha=.91; scale mean=30.19; scale SD=7.64).

Table 3 – Unrotated principal components and item loadings of judgements of intrinsic quality (Pilot 1)

	Component			
	1	2	3	4
In case this peer review of the study would be used to decide upon publication/non-publication in a scientific journal, would the reviewed study be accepted?	.844		-.238	
The study is well (appropriately) organised.	.803	.261	.259	
The study uses an appropriate methodology and method in a careful and thorough manner.	.797			
The argumentation in the study is clear, coherent and internally consistent.	.773	.274	.245	
The conclusion is justified by the evidence in the study.	.770			
The study communicates in an appropriate way.	.750		.321	
It is a pleasure to read the study.	.749	.259		
The work makes a significant contribution to educational research, theory, policy or practice.	.748		-.329	
The theoretical and/or practical embeddedness of the study is made clear.	.694			.256
The study shows awareness of previous work in the field and makes its own contribution clear (in relation to content, methodology, and findings).	.661			
The study is original, creative or innovative in a significant (i.e. non-trivial) way.	.615	-.318	-.305	.395
The study is reflexive about its own limitations.	.575		-.511	-.321
The study has respected (where applicable) the ethical principles which normally operate in this field.	.384	-.643	.383	
The title of the article fits to the article.	.405	.554		-.494
It appears that the study is genuinely the work of the named author (i.e. evidence of plagiarism or unacknowledged derivativeness would be a counter indicator of quality).	.385	-.477	.482	
The reviewed article is related to my own area of research.		.532		.617

Extraction Method: Principal Component Analysis.

Further exploration of the data was carried out by varimax rotation of the four components. The rotated results in Table 4 are interpreted as follows. Component 1 refers to an evaluation dimension reflecting overall significance, originality, and scientific relevance in theoretical, policy and practical points of view. The second component points at the consistency and logically communicative structure of the document reviewed in particular. The third evaluation

dimension mainly accentuates ethical aspects. The item about the relationship of the reviewed article to the field of expertise of the reviewer is hardly or not related to the first three evaluation dimensions.

Table 4 – Varimax rotated principal components and item loadings of judgements of intrinsic quality (Pilot 1)

	Component			
	1	2	3	4
The work makes a significant contribution to educational research, theory, policy or practice.	.831			
The study is original, creative or innovative in a significant (i.e. non-trivial) way.	.794			
In case this peer review of the study would be used to decide upon publication/non-publication in a scientific journal, would the reviewed study be accepted?	.791	.350		
The conclusion is justified by the evidence in the study.	.668	.414		
The study uses an appropriate methodology and method in a careful and thorough manner.	.657	.385	.223	
The study is reflexive about its own limitations.	.640	.382	-.243	-.314
The study shows awareness of previous work in the field and makes its own contribution clear (in relation to content, methodology, and findings).	.610	.304		
The theoretical and/or practical embeddedness of the study is made clear.	.560		.412	.280
The title of the article fits to the contents.		.842		
The study is well (appropriately) organised.	.409	.678	.256	.296
The argumentation in the study is clear, coherent and internally consistent.	.388	.670	.229	.282
The study communicates in an appropriate way.	.354	.603	.418	
It is a pleasure to read the study.	.460	.535		.389
The study has respected (where applicable) the ethical principles which normally operate in this field.			.812	
It appears that the study is genuinely the work of the named author (i.e. evidence of plagiarism or unacknowledged derivativeness would be a counter indicator of quality).			.767	
The reviewed article is related to my own area of research.			-.213	.802

Extraction Method: Principal Components Analysis.
Rotation Method: Varimax with Kaiser Normalization.

2.3 Conclusion

Although the sample size is relatively small, the results of pilot 1 encourage the approach used to assess intrinsic quality indicators by peer reviewing. In a second pilot, both intrinsic and extrinsic indicators will then be scored with respect to a new set of documents.

3. Pilot 2: Intrinsic and extrinsic quality indicators

3.1 Intrinsic indicators

The experiences with intrinsic indicators in pilot 1 resulted – among other things – in a reformulation of their conceptual structure and the corresponding items. In pilot 2 the intrinsic indicators operationalise the concepts: methodology, results, discussion, originality, significance, validity, and miscellaneous. A total of 20 items was formulated to assess these concepts: see Table 5. The answer alternatives for each item are respectively: 'not relevant for this text' (=0), very poor' (=1), (2), (3), 'average' (=4), (5), (6), 'excellent' (=7). By responding to item 21, a peer reviewer could complete information about the degree to which the document evaluated was related to the own area of research. Here the answer categories are: ' Very close' (=1), 'closely' (=2), 'less closely' (=3), 'not at all' (=4).

Table 5 – Concepts and items to assess intrinsic quality (Pilot 2; n documents=171)

Variable name	Description	Min.	Max.	Mean	SD
Methods_1	The methods are intelligibly described	.00	7.00	4.02	2.03
Methods_2	The method / approach is appropriate	.00	7.00	4.70	1.63
Methods_3	The method / approach is accurate	.00	7.00	4.34	1.78
Results_1	The results are completely described	.00	7.00	4.51	1.66
Results_2	The results are correctly described	.00	7.00	4.53	1.67
Discussion_1	The study's method is reflected in an appropriate way	.00	7.00	3.94	1.82
Discussion_2	The study's results are reflected in an appropriate way	.00	7.00	4.51	1.69
Discussion_3	The pattern of reasoning is consistent	1.00	7.00	5.48	1.10
Discussion_4	The discussion shows a critical evaluation of the work	.00	7.00	4.67	1.47
Originality_1	The study shows new approaches in its methodological procedures	.00	7.00	3.39	1.63
Originality_2	The study shows new approaches in the structure of its argumentation	.00	7.00	4.16	1.35
Originality_3	The study contributes innovative ideas for the state-of-art in its research area	.50	7.00	4.52	1.33
Significance_1	The study contributes to the development of its research field.]	1.00	7.00	5.02	1.28
Significance_2	The study makes a significant contribution to the latest discussions within the research field	1.00	7.00	4.82	1.30
Significance_3	The study makes a significant contribution to the latest discussions within the educational policy field	.00	7.00	4.62	1.52
Significance_4	The study makes a significant contribution to the latest discussions within the educational practice field	.00	7.00	4.51	1.57
Validity_1	How do you evaluate the article concerning its Rigour?	.00	7.00	4.72	1.38
Validity_2	How do you evaluate the article concerning its Originality?	1.00	7.00	4.82	1.05
Validity_3	How do you evaluate the article concerning its Significance?	1.00	7.00	5.03	1.22
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
Miscellaneous1	The reviewed article is related to my own area of research…	1.00	4.00	2.40	0.54

The dataset resulting from pilot 2 consists of 177 research documents or articles written by a total of 268 authors. Peer reviewers scored these documents with respect to 20 intrinsic indicators (cf. Table 5) plus item 21 about the relationship of the document to the own research area. For each document, peer review evaluation scores were aggregated by calculating their mean across reviewers.[2] In the univariate analysis using SPSS (version 17.0) only documents without system missings were used, which results in item specific information of 171 documents. Table 5 presents the descriptive statistics of these intrinsic items. Their means vary around 4 (average) or 5; standard deviations vary from 1.05 to 2.03.

3.2 Extrinsic indicators

Extrinsic indicators usually measure aspects of research documents like for example number or distribution of citations (per author; across authors; per document; hits resulting from web search engines for a paper or (combination of) author(s); and so forth). The information about extrinsic indicators used in pilot 2 was provided per author. Because research documents constitute the unit of analysis, this extrinsic information was aggregated per document. In case of more than one author per document, the available information per indicator has been aggregated across authors per document.[3]

[2] The dataset containing both intrinsic and extrinsic scores of 177 research documents became available at 2nd March 2011. The dataset contains scores of peer reviewers who are partner in the EERQI project or attended the European Conference on Educational Research 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 low number of reviewers, the actual data structure does not facilitate the assessment of interobserver reliability or multilevel analyses between and within languages and / or reviewers, respectively.

[3] Identification of documents and authors is based on variable 'revID' (named 'CODE' in earlier datasets). Each record starts with the character 'd' or 'e', followed by a number and sometimes another character has been added. Each character added seems to represent another record in the database, which may identify authors in case a document has more than one author.

The dataset of 2nd March 2011 contains information about 12 extrinsic indicators. Five of these were neglected.[4] Information about the remaining seven extrinsic indicators, their range of scores, means and standard deviations is given in Table 6.

Table 6 – Variables and descriptions to assess extrinsic quality (Pilot 2; n documents=171)

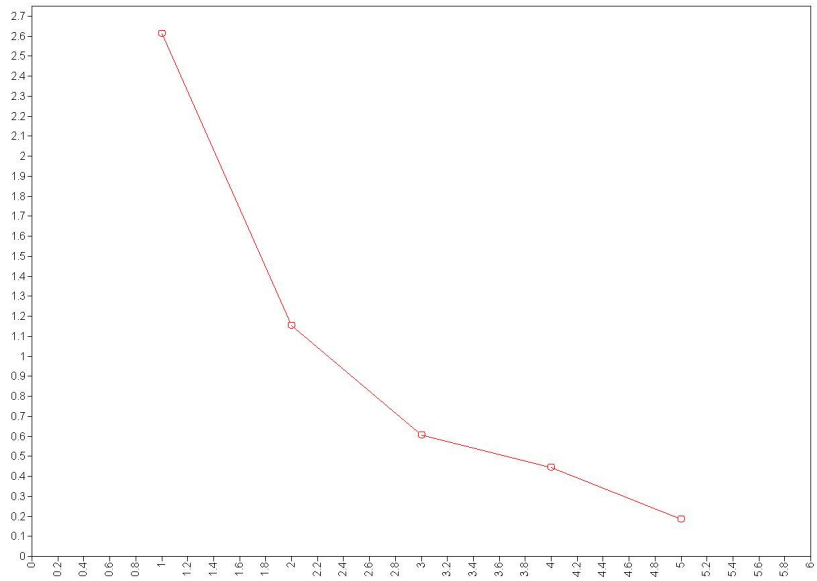
Variable name	Description	Min.	Max.	Mean	SD
Cit/paper	Citations per paper without self citations using the full title of the article	.00	804.81	18.48	64.36
WebMennAuth	Web mentions of author in search engine BING; number of URL's of pages matching the query submitted	2.00	1791.00	352.23	280.33
WebMentTitle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

The variable 'number of citations per paper' [Cit/paper] has a very skew distribution to the right.[5] Therefore, 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 used to explore the relationships between the seven extrinsic variables of Table 6. The variables WebMentTitle and MetagerHits are hardly 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 at the existence of two underlying factors: see Table 7 and the results of the Scree test in Figure 2.

Table 7 – Eigenvalues and explained variances for extracted factors of five extrinsic variables (Pilot 2)

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

Figure 2 - *Scree test results with respect to extrinsic indicators*



[4] These are 'ConnoteaHits', 'MendReader', 'Downloads08, 'Downloads09, and Downloads10'. With the first two variables, scores on all documents were 0; the Downloads variables had a lot of missing values.
[5] The value '0' may reflect 'missing value' or 'no hits'/'no citations'. In this paper the latter ('no #') is assumed.

The loadings of the five variables on the two factors were rotated (oblique, geomin): see Table 8. The results illustrate that 'Citations per paper (without self citations)' en 'Web mentions of author in search engine BING' characterise factor 1, whereas the second factor represents numbers of hits by three other search engines.

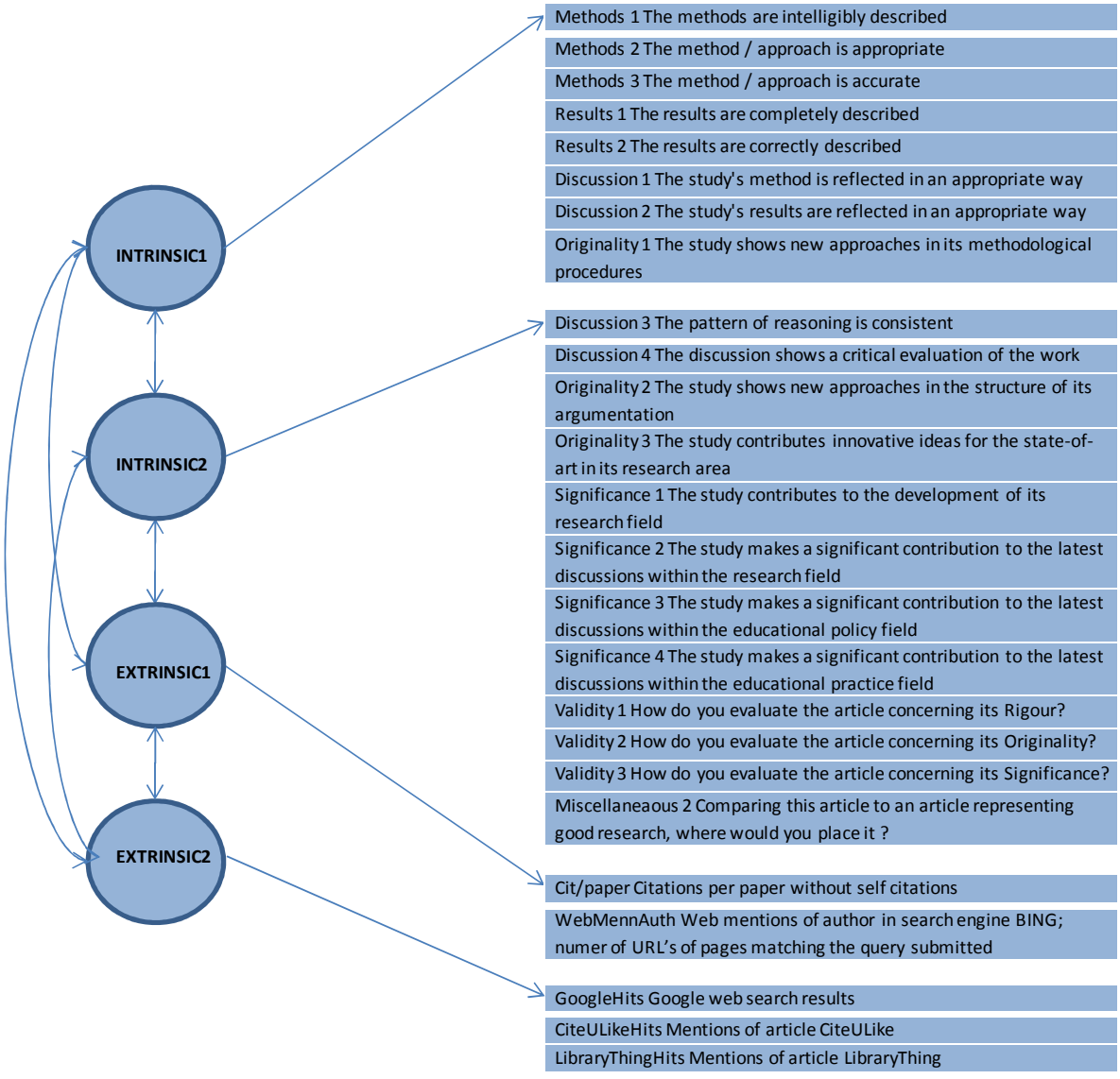
Table 8 - Factor loadings of extrinsic variables after oblique (geomin) rotation (Pilot 2)

Variable name	Description	Factor 1	Factor 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

3.3 Modeling intrinsic and extrinsic latent factors

Using results of a former factor analysis with the intrinsic variables of Table 5, a measurement model was constructed with two intrinsic factors and two extrinsic factors (cf. Table 8). The model is given in Figure 3.

Figure 3 – Graphic presentation of a CFA measurement model with four latent factors



In Figure 3, latent factor 'Intrinsic1' represents intrinsic indicators characterising methodological adequacy, completeness and correctness of reporting results, appropriateness of discussion, and originality with respect to methodological procedures. *Intrinsic1* then 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* therefore 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* then 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 then associated with *number of hits in three specific search engines.*

Figure 3 specifies a 'Confirmatory Factor Analysis' (CFA) to check the relationships between each latent factor and the own specific indicators or observed variables, while taking account of the correlations between various varlatent factors.[6] The variance of each observed indicator variable is explained by both the regression on the specific latent factor and the own specific error variance; error variances between observed indicators may be correlated. The statistical program MPlus (version 6.1) is used to simultaneously check the fit of the measurement model of Figure 3 against the intrinsic scores (Table 5) and the extrinsic scores (Table 6). The outcomes of Maximum Likelihood analysis are given in Table 9.

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 below Table 9. 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 9 it can be seen that RMSEA=0.129 and SRMR=0.072.

Table 9 furthermore demonstrates that the two intrinsic factors correlate rather strongly (0.631) and the two extrinsic factors are correlated to a lower degree (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 which deserves more attention for reasons of both interpretation and modeling of EERQI. The other correlations between intrinsic and extrinsic latent factors are not significant statistically. This implies that use of intrinsic indicators may add quality information to EERQI consisting of extrinsic indicators only, or that introduction of extrinsic indicators may add quality information to EERQI containing intrinsic indicators only.

Given the data available and the small sample in pilot 2, the overall results in Table 9 confirm the first empirical check of the validity of the measurement model of Figure 3. Moreover, the confirmatory factor loadings and the variances explained per indicator (R^2) are relatively high. However, inspection of the modification indices reveals that improvement of Figure 3 seems possible. To explore the statistical consequences, some alternative models were constructed and checked against the model presented in Figure 3 and Table 9. An overview of the alternative models and their statistical outcomes is given in Table 10.

[6] In the measurement model of Figure 3, the relationships between 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 double-sided arrows between all pairs of latent factors. The regressions of each of the indicator variables on their respective latent factor are represented by one-sided arrows. The total variance of each indicator is set to 1.

Table 9 – ML parameter estimates (all standardised) of the measurement model in Figure 3

	Factor loadings				
Latent factors / **Indicators**	INTRINSIC1: Method. adequacy	INTRINSIC 2: Sign./orig./consist.	EXTRINSIC1: # citat./web BING	EXTRINSIC2: Hits 3 search engines	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
Miscellaneous2		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
LibraryThingHits				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: $\chi^2(269)=1028.656$ (p=0.000); RMSEA=0.129; SRMR=0.072; ~ =non-significant (p>0.05).

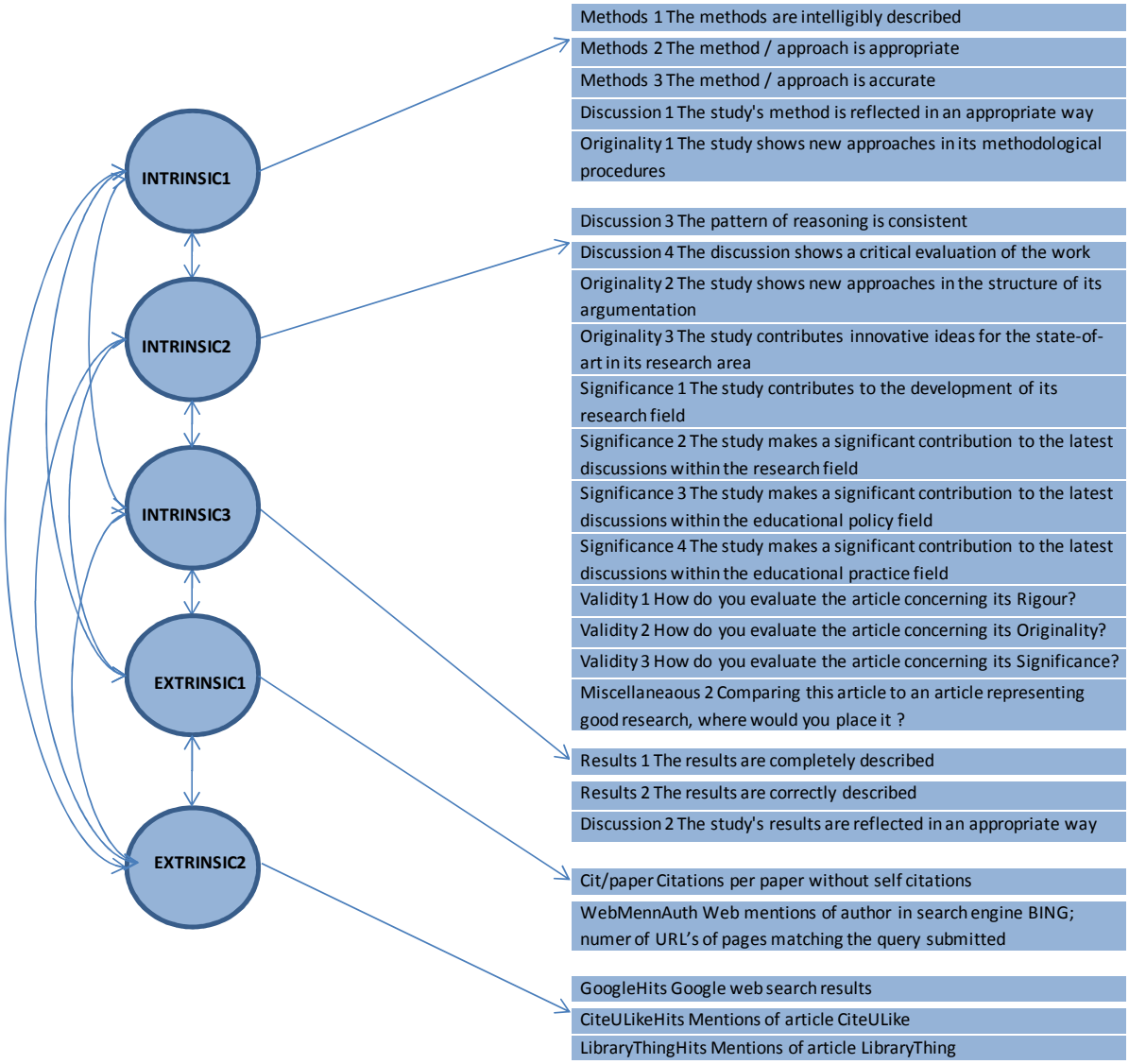
In Table 10, model 1 is the model given in Figure 3 and Table 9. Model 2a of Table 10 allows correlation between result indicators Results_1 and Results_2. Compared to model 1, model 2a demonstrates a decrease in Chi-Square of 243.424 with a difference of only one degree of freedom (df). This difference between model 1 and model 2a is highly significant: model 2a results in a significant improvement of model 1. This is shown also in the values of RMSEA (0.106) and SRMR (0.070).

Table 10 – Comparison of different CFA models

Alternative measurement models	χ^2	df	RMSEA	SRMR
1. Model with 4 latent factors (2 intrinsic, 2 extrinsic; Figure 3)	1028.656	269	0.129	0.072
2a. As Model 1, but with error covariation Result_1 - Result_2	785.232	268	0.106	0.070
2b. Model with 5 latent factors (3 intrinsic, 2 extrinsic; Figure 4)	758.385	265	0.104	0.077

Additional explorative analysis of various parameters suggests to combine intrinsic indicators Results_1, Results_2 and Discussion_2. This implies the existence of three instead of two intrinsic latent factors. This suggestion changes the CFA model of Figure 3 into the CFA model of Figure 4.

Figure 4 – Graphic presentation of a CFA measurement model with five latent factors



The statistical outcomes in Table 10 illustrate that, compared to CFA model 1, CFA model 2b results in a significant improvement in Chi-Square (270.271; df=4; p.<.01) and acceptable values of both RMSEA (0.104) and SRMR (0.077). To facilitate more detailed comparison between model 1 (Figure 3, Table 9) and model 2b (Figure 4), some statistical information of model 2b is presented in Table 11.

Table 11 demonstrates that Results_1, Results_2 en Discussion_2 are good estimates of Intrinsic3 (0.968, 0.975 and 0.787, respectively). Compared to the range of the variances (R^2) of these indicators in Table 9 (0.609 to 0.626), the introduction of latent factor Intrinsic3 or 'Results quality' increases R^2 of these indicators to a range between 0.620 en 0.951. Furthermore, Table 11 shows significant correlations between the three intrinsic factors and between the two extrinsic factors; however, contrary to the result in Table 9, intrinsic and extrinsic latent factors in Table 11 are not correlated significantly. Like the outcome of Table 9, this result concerning the relationships between intrinsic and extrinsic latent factors in Table 11 needs more attention for reasons of both interpretation and modeling issues in the EERQI conceptual framework. Moreover, this empirical outcome again supports the notion that the use of intrinsic indicators may add specific quality information to EERQI consisting of extrinsic indicators only and that introduction of extrinsic indicators may add specific quality information to an EERQI framework containing intrinsic indicators only.

Table 11 – ML parameter estimates (all standardised) of the measurement model in Figure 4

	Factor loadings					
	INTRINSIC1: Method.adequacy	INTRINSIC2: Sign./orig./cons.	INTRINSIC3 Results	EXTRINSIC1 # citat./web BING	EXTRINSIC2 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
LibraryThingHits					0.803	0.645

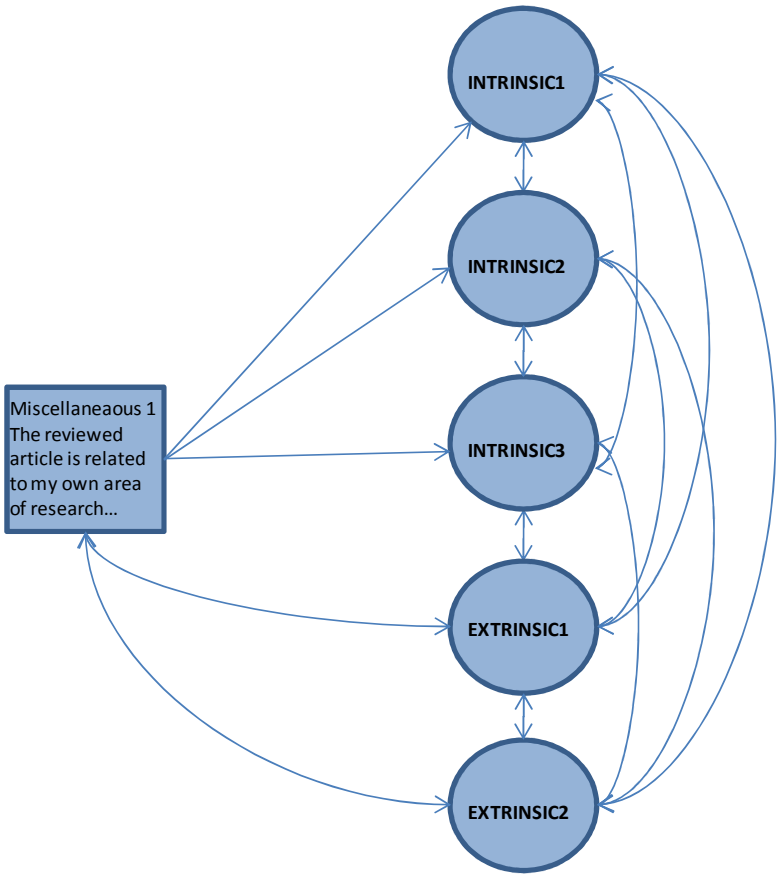
Factor covariances (correlations)				
	INTRINSIC1	INTRINSIC2	INTRINSIC3	EXTRINSIC1
INTRINSIC2	0.620			
INTRINSIC3	0.740	0.476		
EXTRINSIC1	0.236~	0.148~	0.188~	
EXTRINSIC2	0.146~	0.085~	0.113~	0.460

Fit indices: $\chi^2(265)=758.385$ (p=0.000); RMSEA=0.104; SRMR= 0.077; ~ =non-significant (p>0.05).

3.4 Structural model of intrinsic and extrinsic latent factors

A final exploration is directed at the possible explanation of the CFA model in Figure 4. 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 Miscellaneaous1 in Table 5) influences the scores of the intrinsic latent factors. Inclusion of this explanatory variable in the CFA model of Figure 4 transforms this model into a causal or structural model including intrinsic and extrinsic latent factors and their indicators. The structural latent factor model is illustrated in Figure 5. In this Figure, the specific indicators of the latent factors are the same as those in Figure 4. Moreover, Figure 5 illustrates that the three intrinsic latent factors are regressed on the explanatory item Miscellaneaous1 ('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.

Figure 5 – Structural model with intrinsic and extrinsic latent factors (indicators not included)



The main results of Maximum Likelihood (ML) analysis using MPlus (version 6.1) are given in Table 12. Miscellaneaous1 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 the reviewed document is related to the reviewer‘s own area of research, the higher the document is evaluated by the reviewer with respect to significance, originality and consistency (Intrinsic2) and the higher the document is evaluated concerning methodological adequacy (Intrinsic1). The two effects seem to reflect some subjective evaluation bias that may take place in peer reviewing. No relationships exist between Miscellaneaous1 and the two extrinsic factors. The fact that degree to which the reviewed document is related to the reviewer‘s own area of research is related to intrinsic factors, and not to extrinsic factors, can be interpreted to support the validity of the conceptual framework.

Table 12 - ML factor parameter estimates (all 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
Miscellaneaous1	-0.176	-0.247	-0.128~		
	Correlations				
Miscellaneaous1				0.029~	0.020~

Fit indices: $\chi^2(284)=779.559$ (p=0.000); RMSEA=0.101; SRMR= 0.075; ~ =non-significant (p>0.05).

4. Conclusions

4.1 A first prototype framework

Different pilots were conducted to collect data with respect to 'intrinsic' and 'extrinsic' indicators of research documents. The research questions to be answered are:

1) Pilot 1: What are the results of applying the criteria rigour, originality, significance, integrity, and style, in a peer review evaluation of educational research articles?
2) Pilot 2: What are the results of analysing the relationships between both intrinsic and extrinsic indicators, to construct an example of a prototype EERQI framework?

In the first pilot, peer review scores of indicators of the criteria rigour, originality, significance, integrity, and style were subjected to principal components analysis. The results reveal an evaluation dimension reflecting overall significance, originality, and scientific relevance in theoretical, policy and practical points of view. A second component points at the consistency and logically communicative structure of the document. A third evaluation dimension mainly accentuates ethical aspects. These outcomes answer the first research question.

In the second pilot, peer review scores or intrinsic indicators operationalise aspects of the methodology, results, discussion, originality, significance, validity, and miscellaneous of research articles or documents. Extrinsic indicators of the same documents are 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 focused on the hypothesis that the degree to which a reviewed article is related to the reviewer's own area of research, will influence the scores of the intrinsic latent factors. Empirical testing in a causal structural model reveals that the more the reviewed document is related to the reviewer's own area of research, the higher the document is evaluated by the reviewer with respect to 1) significance, originality and consistency, and 2) methodological adequacy. No significant correlations exist between the reviewer's own area of research and the extrinsic factors.

The differentiated relationships and outcomes of pilot 2 support the validity of both the conceptual framework and the empirical research. It can be concluded that an example of a prototype EERQI framework has been constructed. The relevant conceptual framework is presented in the combination of Figure 5 and Figure 4. Other types of information, for example semantic indicators and factors, can be integrated in these Figures and relevant research. Given the statistical outcomes related to Figures 5 and 4 in Tables 9 – 12, it is furthermore concluded that the conceptual EERQI framework was checked successfully in a first empirial test. The main goal of the EERQI project - to improve upon citation-only based assessments of the quality or impact of educational and other research - has been supported. However, some limitations of the study have to be presented also.

4.2 Methodological aspects

From a methodological point of view the present study has a number of limits, for example:

1) the exact sampling procedures of documents and reviewers need to be spelled out;
2) numbers of documents and numbers of reviewers in the pilots are rather low;
3) procedures to operationalise / interpret variables or outcomes need to be detailed;
4) ratio of number of documents / number of indicators (pilot 2: 171/25) is relatively low;
5) more attention should be given to distribution characteristics of the variables;
6) attention should be given to interobserver reliability in the reviewing information;
7) possible effects of language differences are not taken into account;
8) multilevel analysis was not applied because of low numbers of documents/reviewers;
9) empirical testing may be hampered by low numbers of documents/reviewers;
10) lack of time to carefully check, analyse and interpret data and report about results.

4.3 Future steps

The EERQI project has many sides and strong potentials. Partners can be convinced of its significance, originality and consistency (latent factor Intrinsic 2). Continuation of these strong potentials, in combination with an improved focus on both methodological adequacy (latent factor Intrinsic 1) and semantic indicators and latent factors, will build optimal steps to the future of EERQI.

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.

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 aMeasure.* Berlin: Humboldt-Universität, Institut für Bibliotheks- und Informationswissenschaft.