INTERNATIONAL DESIGN CONFERENCE - DESIGN 2010
Dubrovnik - Croatia, May 17 - 20, 2010.

DESIGN 2010

THE LONG-RUNNING ISSUE OF REVIEW QUALITY – FINDINGS FROM AN EMPIRICAL STUDY AMONGST INTERNATIONAL REVIEWERS

H. Birkhofer and S. Zhao

Keywords: peer review, review quality, content and formal quality

1. Introduction – background and motivation

Peer review has been conducted in almost all conferences since many years in order to secure the suitability of the contributions to the conference themes and to achieve the required paper quality in form and content. Nevertheless, there are over and over again complaints concerning the intransparency and inconsistency of peer reviews [Anderson 2009]. In this sense, it is not always clear to authors why their contributions have obtained certain evaluations or even have been rejected. Even more confusing is the case when reviews of one paper differ essentially. There are even cases, experienced in several meetings of conference chairs, that one paper gets the evaluation "excellent", whereas another reviewer proposes the paper to be rejected. Not only are those kinds of evaluations confusing for the corresponding authors, but also they question the consensus on what is good and bad and may lead to critizise the objectivity of the peer review system in general.
As in many other organizations, the issue of reviewing is being discussed within the Design Society [Blessing/Chakrabati 2009]. Two years ago, a task force has been founded in order to elaborate on the chances to improve the review quality and to harmonize differing reviews for identical papers. First approaches for dealing with this issue have been presented and discussed at the spring meeting 2009 of the Board of Management (BM) and the Advisory Board (AB) of the Design Society in Boston. During the discussions, again, there were different opinions of what constitutes a good paper as well as diverging conceptions of science and paper quality. All those differing opinions and positions within a group of international scientist even of such a small size constituted the motivation for a survey amongst reviewers and the analysis of the findings which are presented in this paper.

2. Objectives

The study being presented in this paper was conducted in order to elaborate the degree of differing evaluations amongst the reviewers as well as the individual perceptions and opinions. The objective was to obtain as many reviews as possible for one single paper which has been sent out to all potential referees with the same instructions. The results were to be analyzed in terms of frequency distributions, mean average and standard deviations. Based on these results, concurrent and diverging evaluations were to be extracted and, when possible, explained.
Before starting the survey, based on various discussions within the BM and AB of the Design Society as well as with young researches faced with contradicting reviews, the following hypotheses have been proposed:

H1: The exact definition of a paper's quality in terms of content by evaluating it is not/hardly possible.
H2: A paper's formal quality is easier to evaluate than its quality in respect of content.

H3: The reviewer's scientific career (place of higher education, scientific culture, etc.) has an enormous impact on the review result.
H4: The reviewer's review experience affects the review result essentially.
H5: The fact that the reviewer knows or believes to know the author has an essential impact on the review result.

This study does not claim to be statistically firm. Rather, this paper tries to fill the gap between such a scientifically sound study which is only feasible with enormous efforts and all the opinions to arise out of individuals' perceptions. In this sense, despite the quantitative depiction of the results, this study shall be considered in a qualitative way in order to give an approach to improve the review process on a more graspable basis. Being convinced that peer review basically is an adequate and helpful instrument to secure the quality of scientific contributions [Zhao/Birkhofer 2010], still the flaws and the potential for improvements need to be elaborated on, true to the motto: "Better is the enemy of good".

3. Approach and methods

The Design Society can draw upon a large pool of internationally renowned reviewers. All reviewers have received the same paper with the request to evaluate the contribution according to review criteria which were identical for all referees. The collecting and processing of the reviews was conducted with the online tool *Survey Monkey* and the interpretation of the data was completed by the authors of this paper.

3.1 Paper to be reviewed

The basis of the paper to be reviewed was a draft of a contribution for the Design 2008 conference in Dubrovnik with the subject "elementary methods". The topic of the paper was considered to be wide enough, concerning the fundamentals of *Systematic Design* [Hubka 1996, Grabowski 1998], and therefore could be regarded as cross-disciplinary with a certain probability. The draft was shortened to five pages in order to keep the efforts for the reviewer within a feasible limit. Moreover, the title and content was modified and neutralized in respect of authors and references for the purpose of avoiding conclusions to be drawn about the authors. According to the opinions of the author of the reviewed paper, the contribution was adjusted in a way that it corresponds to a moderate submission for a Call for Papers of a Design conference. By doing so, the exploitation of the whole range of the review rating scale was considered possible.

3.2 Review criteria

The review criteria were adopted from those being used in the Design conferences, except for one question which was extracted from the ICED criteria. Additionally, optional questions regarding the individual background of the reviewer were included.

Table 1. Review criteria and questions (partly shortened)

No.	Question/Criteria	Rating scales				
		0	1	2	3	4
1)	Overall quality in respect of content:	Poor				Excellent
2)	Paper's novelty and level of contribution:	General, un-articul. mat.	Repetition known mat.	New applic. known mat.	New contri-but., addition	Innovative contribution
3)	Discourse and conclusions valid?	Not justified, no message	Major omiss. weak justific.	Loose gene-ralizations	Good justi-fication	Strong justi-fication
4)	Industrial or appli-cation perspective?	No comments	Naive arguments	Questionable reflection	Reasonable reflection	Strong reflection

5)	How to improve content of paper:	Comments (open-ended)				
6)	Overall quality regarding formal aspects:	Poor				Excellent
7)	Paper well structured and organized?	Inadequate structure	Irrelevant material	Inadequate length	Reasonable structure	Good structure
8)	Illustrations clear and understandable?	Unacceptable	Major flaws	Partly inadequate	Reasonable, clear concept	Complete, precise
9)	How to improve formal aspects of paper :	Comments (open-ended)				
10)	What takes it to bring paper to accept. form:	Is not acceptable	Acceptable with major revisions	Acceptable with minor revisions	Acceptable as it is	
11)	Importance for Design Society confer.	None	Poor	Questionable	Good	Significant	Exceptional
12)	Your familiarity with the topic:	My area of expert.	Good knowled.	Familiar with	Marginally familiar	Not really familiar	Completely new
13)	Main focus of your scientific career?	North America	South America	Europe	Asia	Australia	Africa
14)	Does it differ from you current place of employment?	No	Yes, I now work in				
			North America	South America	Europe	Asia	Australia	Africa
15)	Years of review exp.?	0 - 1	1 - 5	5 – 10	> 10		
16)	Time of review?	Morning	Afternoon	Evening	Night		
17)	Duration of time to complete this review	< 30 minutes	30 minutes – 1 hour	1 – 2 hours	> 2 hours		
18)	You know/think to know paper's author?	Yes		No			

Question 5 and 9 were to be answered by giving comments which was done by the majority of the participating reviewers. Questions 13-18 helped drawing conclusions about the reviewers' personal experiences and habits. In this paper the metric of question 10 within table 1 was inverted in regard to the questionnaire to get an unified metric with all highest scores representing best values.

3.3 Contacted reviewers and review process

The Board of Management of the Design Society was kind enough to provide the authors of this paper with the complete list of the active reviewers. All reviewers were contacted by e-mail and informed in detail about this undertaking and its motivation. They were asked to review the attached paper and submit their evaluation in the internet on *Survey Monkey* which not only enables the user to create surveys but also helps collecting and analyzing the answers. In order to submit their review, the reviewers had to click on the link which was given in the e-mail and then answer the evaluation questions. Strict neutrality and anonymity were assured. In return for the efforts of completing a review, the access to the analyzed data was promised to those who participated. Table 2 shows an overview of the contacted and participating reviewers.

Table 2. Overview of contacted and participating reviewers

Contacted reviewers	244
Completed reviews	75
Rate of return	30,7%
Additional comments via e-mail	39

Some reviewers declared that they were not able to do the review due to shortage of time. The undertaking was well received by the majority of the additional e-mails the authors have gotten. As to the use of the online tool *Survey Monkey*, there were no comments, so an unproblematic handling can be assumed.

4. Results

Survey Monkey allows a convenient analysis and charting of the answers. The evaluation results were depicted in the form of bar diagrams. Additionally, the mean values (arithmetic mean) as well as the standard deviations were calculated in *Microsoft Excel*.

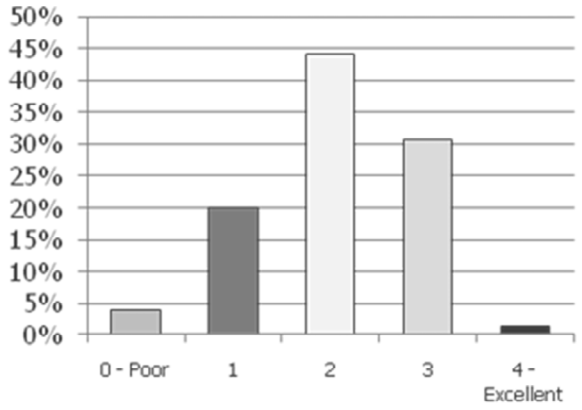
4.1 Primary answers

Primary answers are those answers which were given by the participants to the 18 questions except for questions 5 and 9 which are described in 4.1.2.

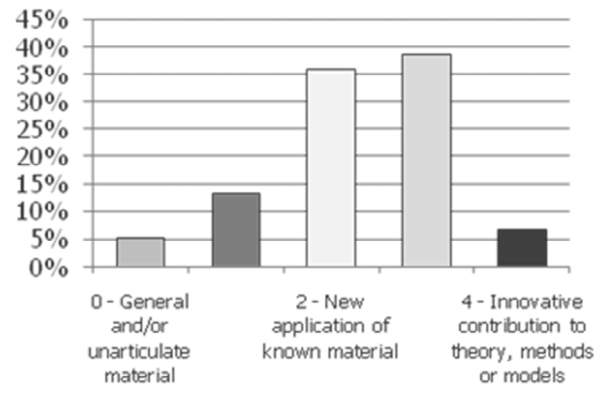
4.1.1 Closed-ended questions

The questions for which there were a definite set of answers the reviewer could choose are called closed-ended questions.

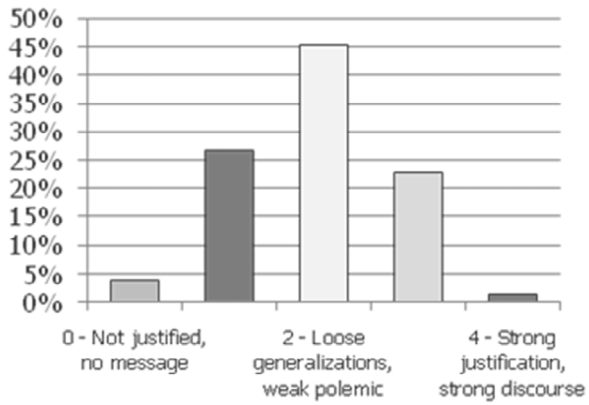
Question 1: Please assess the overall quality of the paper in respect of content:

The „moderateness“ of the paper is emphasized by the mean value of 2,053. The standard deviation of 0,847 is relatively high. The evaluations cover the whole range of rating scales whereas the distribution approximately corresponds to a Gaussian distribution. The heterogeneous evaluation results are in fact critical and confusing for the author of the reviewed paper and seem to underpin hypotheses H1. Another issue to be cleared is the question if the mean value of 2,053 can really indicate the paper's quality.	 $\overline{x} = 2{,}053;\ \sigma = 0{,}847$

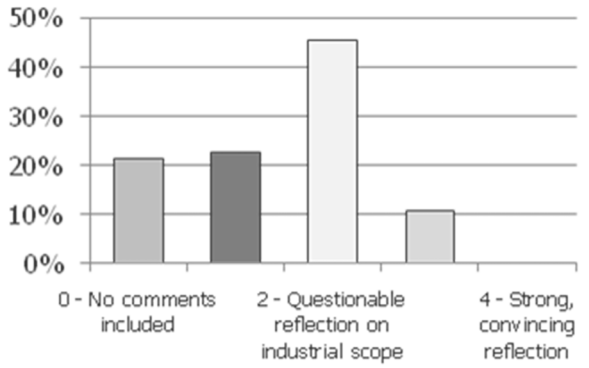
Question 2: Please indicate the paper's novelty and level of contribution:

With a mean value of 2,280, the paper has received a noticeably higher rating in respect of this criterion. However, the standard deviation (0,960) is even higher than in question 1 which indicates a bigger insecurity amongst the reviewers. Assessing the level of a contribution's novelty surely requires profound knowledge about the state of the art in a certain field of research. This can certainly not be the case with reviewers from other research fields.	 $\overline{x} = 2{,}280;\ \sigma = 0{,}960$

Question 3: Are the discourse and conclusions valid?

In question 3 a comparable result as in the first two questions can be found. The mean value of 1,907 shows that the paper is rated lower in this criterion. The standard deviation accounts for 0,835 and is not extremely high. Nevertheless, a lower standard deviation was actually expected as the argument's logic and intelligibility should be easier to assess than the fuzzy concept of scientific quality. This result also underpins hypothesis H1.	 $\overline{x} = 1{,}907; \sigma = 0{,}835$

Question 4: Is an industrial or application perspective reflected in a reasonable way by the author(s)?

The results of this criterion are remarkable. The evaluation of the paper with respect to an industrial context is as expected low as the paper deals with fundamentals of design methods. However, the variation (standard deviation = 0,942) is relatively high. Although, objectively regarded, the paper does barely give any concrete application perspective, the reviewers seem to include own ideas and assumptions in their review. There apparently is a gap between the objective description by the author and the individual assessment by the reviewers.	 $\overline{x} = 1{,}453; \sigma = 0{,}942$

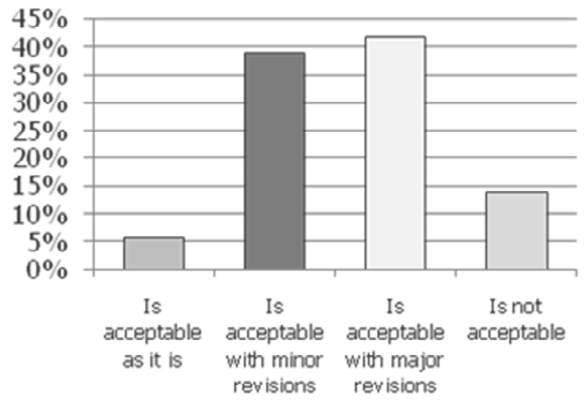
Question 6: Please assess the overall quality of the paper regarding formal aspects:

Assessing the formal aspects of scientific contributions is expected to be easier and more accurate than the assessment regarding the content. But the standard deviation of this criterion (σ=0,842) is almost the same as in question 1 regarding content quality (σ=0,847). If the conclusion, that the assessment of the formal quality also is unclear and fuzzy, can actually be drawn, then this issue needs to be further analyzed. Hypothesis H2, however, is not being underpinned by the results of this criterion.

Question 7: Is the paper well structured and organized and
Question 8: Are the illustrations and tables clear, effective and understandable?

The answers to these two questions which are apparently easy to assess are noticeably heterogeneous. Question 7 even has the highest standard deviation so far (0,963) and also question 8 has a high standard deviation with 0,877. This finding also does not underpin hypothesis H2. Apparently, there is need for action regarding the consolidation of the conceptions of a "good" layout, a convincing structure as well as a clear and understandable graphic presentation of scientific contributions.

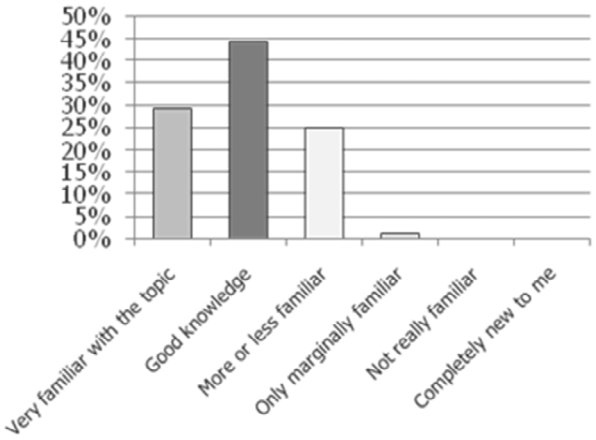
Question 10: Please judge what it takes to bring the paper to an acceptable form:

One out of seven reviewers rejects the paper while more than 40% of the referees would accept the paper without any or only with minor revisions. In other words: Approximately half of the reviewers approves, the other half tends to reject the paper. If reviews are supposed to help the authors, but also the program committee, to assess the quality of a submission and the need for changes, the results of this question can only be disappointing.	 $\overline{x} = 1{,}361; \sigma = 0{,}787$

Question 11: If you consider the whole range of papers recently published at the Design Society conferences, how important would you rank the paper?

Although the mean value of the ratings indicate a certain importance (2,486), the variation of the answers is stunningly high (standard deviation = 0,913). Can this insecurity be ascribed to the reviewers' differing knowledge on the Design Society conferences and the corresponding requirements?

Question 12: Please indicate your familiarity with the topic:

Almost three out of four reviewers consider themselves as familiar with the topic of the reviewed paper. This result can be interpreted in two ways. 1. It is thinkable, that only those reviewers have participated at the review, who are confident that they are able to give a competent evaluation. 2. Assumed the fact that a representative sample of all reviewers of the Design Society has participated, here indeed socially desirable answers can be found as it is unlikely that a reviewer states himself as not competent enough. In this context, it can surely be questioned if every specialist has enough methodic knowledge to give a founded evaluation of the contribution.	

Due to the limitations of pages, questions 13-18 will be discussed only in connection with the results of the answer combinations (see 4.2.).

4.1.2 Open-ended questions

In the open-ended questions the reviewers could comment on the paper and formulate their critique in word. In order to limit the efforts for the reviewers not more than five lines were allowed.

Question 5: Advise the author(s) as how to improve the paper concerning its content:

This question was answered by 92% of the participants . The most named points of criticism with their corresponding frequency are shown in table 3:

Table 3. Most named points of criticism in respect of content

Frequency in %	Criticism
17,0	No perspective of application and benefits
16,0	Paper is too short and lacks details of the analysis
14,2	No real validation or justification
11,3	Lack of references to similar approaches and models
8,5	Objectives of the paper vague or even obscure
7,5	Lack of examples
6,6	Weak argumentation
4,7	Lack of clear and understandable definitions
4,7	Paper is "parochial" and limited on European origin
3,8	Language is weak
2,8	Paper does not lead to scientific novelty
2,8	Paper is too theoretical

In summary, the criticisms are justifiable, even from the reviewed paper's author's point of view, as the fundamentals-oriented content was difficult to understand to begin with, but has even suffered from it being shortened to five pages.

Surprisingly, there were also a noticeable amount of reviewers who provided positive answers:

- Your theory is good and valid.
- Very interesting paper.
- Excellent, interesting and well written paper dealing with a very important subject.
- Paper addresses a key problem in a reasonable and convincing way.
- Section 5.2 is of great importance to industry.

Again, the question has to be put how those evaluations can even arise given all the other (justifiable) negative criticism.

Question 9: Advise the author(s) as how to improve the paper concerning formal aspects:

Only 64% of the reviewers answered this question. The most named points of criticism with their corresponding frequency were "Figures should be more adequately explained" with 48% frequency and "Figures too complex" with 12 % frequency. The rest of criticisms range below 8% frequency.

4.2 Answer combinations

While questions 1-12 correspond to the evaluation of the paper, questions 13-18 had the purpose of characterizing the reviewers. By combining the review results with the personality traits of the referees, the existence of a correlation between the individual, cultural and geographical traits of the reviewer and his review result was to be analyzed. Due to page limitations, again, only the remarkable results are being presented here.

4.2.1 Impact of familiarity with the topic (question 12)

As to the evaluation of the reviewed paper, there is a noticeable impact of the referee's familiarity with the topic (figure 1). While experts ("very familiar", "good knowledge") give lower ratings, reviewers with less knowledge in this field ("more or less" to "marginally") tend to provide benevolent evaluations.

If this impression substantiates and experts are entitled to be more qualified to give objective evaluations, it is indispensable to assign those experts for reviewing the corresponding papers which is

indeed being conducted by responsible program committees since many years. The self-assessment constitutes the basis for the assignment of the submissions by the program committee.

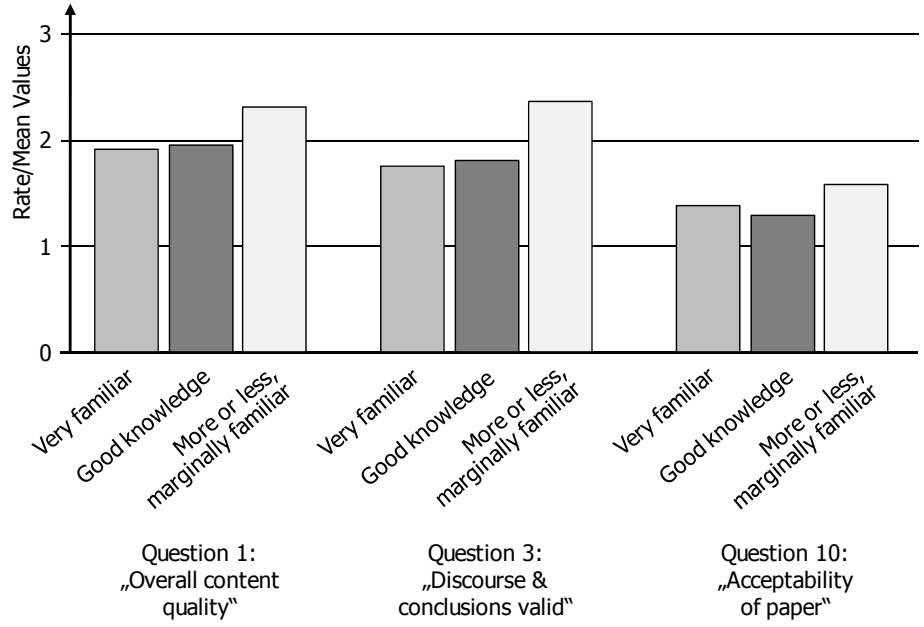


Figure 1. Impact of familiarity with the topic

4.2.2 Impact of place/region of scientific career (question 13)

Various discussions in the recent past have indicated that the place of the scientific/higher education of the referee with its specific understanding of science and cultural impacts influence the evaluation of a paper's quality. This impression is obviously being underpinned by this study (figure 2).

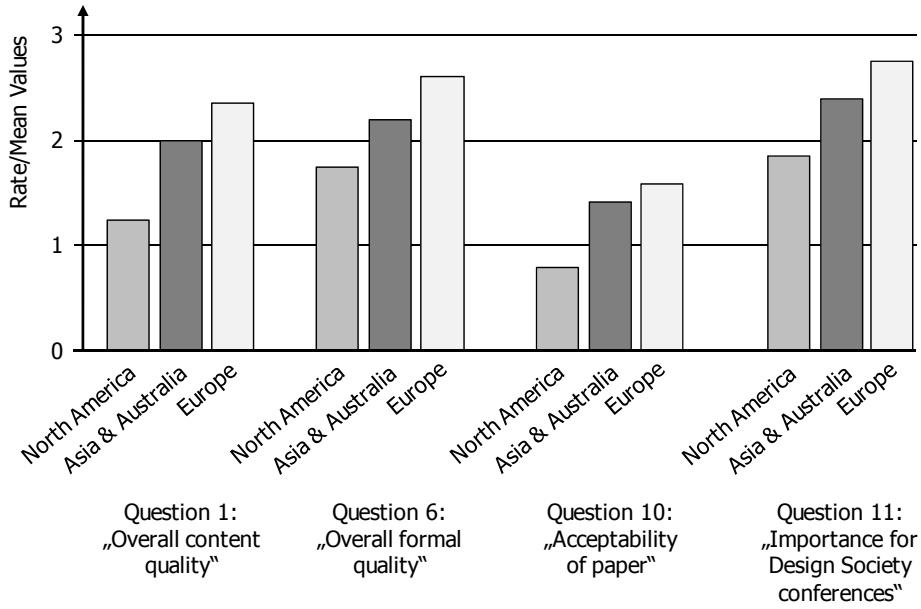


Figure 2. Impact of place/region of scientific career

In the analyzed four criteria the ratings of reviewers from North America were remarkably worse than those from Australia and Asia which were in turn worse than those from Europe. The differences of mean values in all questions between reviewers from North America and those of Europe range within 32% to 85%! These differences are substantial and can influence a review's result in total noticeably which, in turn, underpins hypothesis H3.

4.2.3 Impact of the review experience (question 15)

The years of experience a reviewer has does not show any correlation with the review results at all. Reviewers with 1-5 years, 5-10 years or with more than 10 years of experience give similar evaluations. The assumption that, over the years of review experience, some kind of expertise arises that affects the review result cannot be verified. Thus, hypothesis H4 is not being underpinned.

4.2.4 Impact of the review duration (question 17)

Does short review duration indicate a lower rating of the paper (figure 3)?

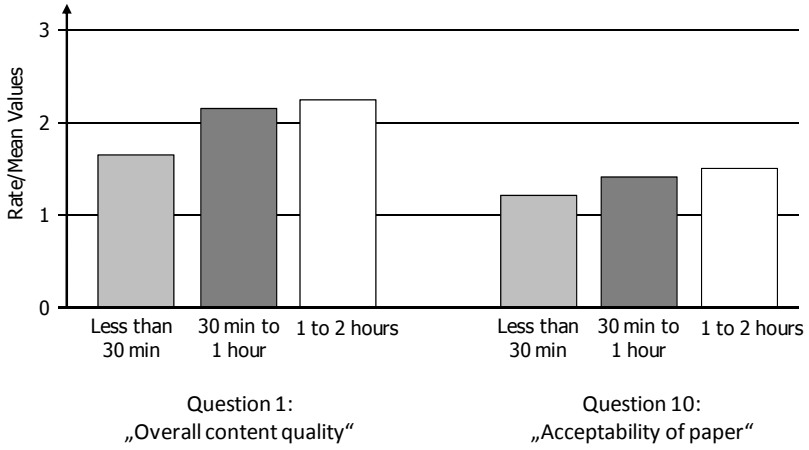


Figure 3. Impact of the review duration

This seems to be the case with the reviewed paper when the review has been completed within 30 minutes. While the mean values of review results in the categories 30 minutes to 1 hour and >1 hour are quite similar, the mean values in the <30 minutes category are significantly lower. A review being completed in a hurry seems to tend to be more critical.

4.2.5 Impact of the author's name recognition (question 18)

If the reviewer knows or thinks to know the paper's author (in this survey this was the case for 25% of the answers), the ratings are dramatically higher (figure 4).

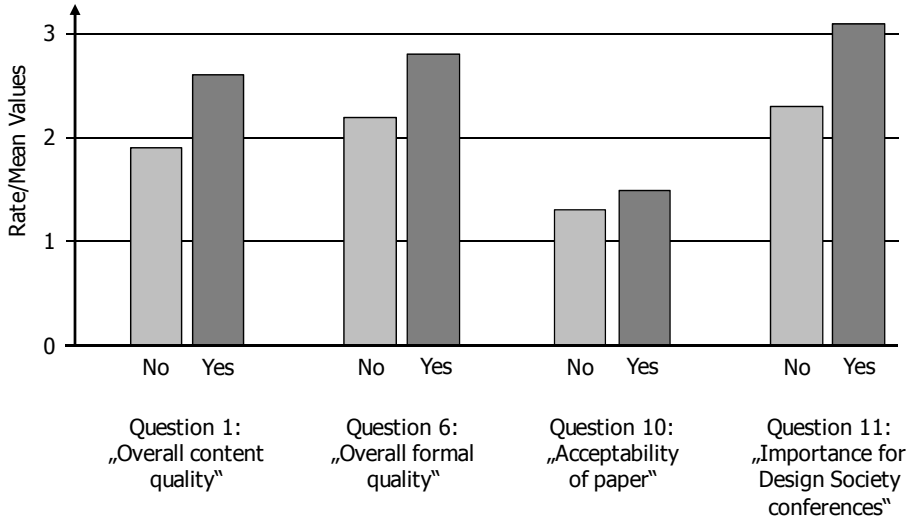


Figure 4. Impact of the author's name recognition (Question 18: Do you know or do you think to know paper's author?)

The differences of the mean values of the questions 1,6 and 11 accounts for about 0,6 which is the highest amongst all correlated mean values. Therefore, hypothesis H5 is being underpinned. As a complete anonymization of a submission is not feasible, this will remain a major flaw of the review

process. In the science community, one knows each other and when reviewing a certain paper, the assessment of the author's character will inevitably influence the evaluation. In this context, newcomers experience more acceptance problems as the community tends to keep to itself.

5. Conclusions

This study shall be considered as a first step in order to make the review process more transparent in terms of elaborating important factors influencing the review result. According to the findings, a strict objective evaluation of a scientific contribution seems to be not feasible. The review process, similar to the design process [Birkhofer 2006], is affected by the character traits of the referee and his scientific imprint. This fact is understood as it is the people who take decisive parts in the processes [Lindemann 2002] with all its likings and dislike [Ioannidis 2005]. The consequences of deficits of peer reviews reach far beyond reviewing papers. It concerns all activities like application for funding or a position in university or industry, where written applications are submitted and evaluated by peers formally or unformally. Several attempts were made in past to overcome such deficits and weaknesses e.g. by Open-Peer-Reviews [Philica], Dynamic-Peer-Reviews [Naboj] or Parallel Open Peer Reviews [Nature] but with varying results solving one problem and creating others. Coming back to the results of this unique survey the authors suggest to carry out further studies to underpin or refute the finding mentioned above. Doing so on a well founded basis of facts this could be a key for designing a better review procedure. In this sense, not the "bargaining" on the review criteria or the rating scales will be rewarding but the adequate consideration of the individuality of the reviewers.

Acknowledgements

The authors would like to thank the Board of Management of the Design Society for providing the reviewers' contact data and especially the 75 reviewers who participated in this survey.

References

Anderson, T.: Conference Reviewing considered harmful. ACM SIGOPS Operating Systems Review, Volume 43 , Issue 2, Pages 108-116.

Birkhofer, H.: ***There is Nothing as Practical as a Good Theory – an Attempt to Deal with the Gap between Design Research and Design Practice.*** *In: Marjaonovic, D. (ed.): Proceedings of the 9th International Design Conference (DESIGN 2006), Dubrovnik, 2006, pp. 7–14.*

Blessing, L.T.M., Chakrabati,A.: DRM, a Design Research Methodology. Springer Dordrecht Heidelberg London New York 2009

Grabowski, H., Rude, S., Grein, G. (Eds.): Universal Design Theory. Aachen, Shaker, 1998.

Hubka, V., Eder, W.E.: Design Science. Springer, Springer London, 1996.

Ioannidis J.P.A., Contradicted and Initially Stronger Effects in Highly Cited Clinical Research, The Journal of the American Medical Association, 2005;294:218–228

Lindemann, U. (Ed.): Human Behaviour in Design, Springer, Berlin Heidelberg New York 2003.

Naboj: Offical website of Naboj

Nature: Overview: Nature's trial of open peer reviews. www.nature.com

Philica: Offical website of Philica

Zhao, S., Birkhofer, H.: Review Quality Management – Applying ISO 9000 standards on the review procedure of the Design Society. Submission for Design Conference, Dubrovnik 2010.

Prof. Dr. h.c. Dr.-Ing. Herbert Birkhofer
Head of Institute
Product Development and Machine Elements (pmd)
Technische Universität Darmstadt
Magdalenenstrasse 4
D-64289 Darmstadt, Germany
Telephone: +49 6151-16-2155
Telefax: +49 6151-16-3355
Email: birkhofer@pmd.tu-darmstadt.de
URL: http://www.pmd.tu-darmstadt.de