REVIEWING PEER REVIEW, AN EYE TRACKING EXPERIMENT OF REVIEW BEHAVIOUR

Boa, Duncan R; Hicks, Ben
University of Bristol, United Kingdom

Abstract

The quality of peer reviews is a longstanding issue within the Design Society with concerns over the consistency and transparency of reviews raised frequently. Previous research has sought to qualify these concerns by describing the variability of review scores and correlating them with academic's backgrounds. This paper aims to update and advance the current understanding of peer review within the Design Society by characterising review behaviour through the addition of eye tracking. Seventeen academics attending Design 2014 took part in an experiment. The results of the experiment are discussed in this paper with the aim of answering two research questions: do different review strategies exist and what are they? And, do character traits of reviewers affect reviewer strategy? Results confirm findings from previous research, suggesting little has changed since the topic was last reported and that inconsistency remains a problem. However, some of the cause of review inconsistency is potentially explainable through identified review strategies evident from eye tracking data.

Keywords: Peer review, Human behaviour in design, Eye tracking

Contact:
Duncan R Boa
University of Bristol
Department of Mechanical Engineering
United Kingdom
dr.boa@bristol.ac.uk

Please cite this paper as:
Surnames, Initials: *Title of paper*. In: Proceedings of the 20th International Conference on Engineering Design (ICED15), Vol. nn: Title of Volume, Milan, Italy, 27.-30.07.2015

1 INTRODUCTION

Publication at international conferences is a key aspect of academic research allowing for the dissemination and discussion of ideas amongst a community. Peers, other academics within the field, assess individual papers to determine their suitability and quality for a conference. The system is reliant on good will and has some notable long-standing issues regarding the quality of reviews.
Birkhofer and Zhao (2010) highlight the main problems reported with peer-review within the Design Society as being a lack of transparency and inconsistency. In the same paper the authors conduct an experiment in which a single paper is peer-reviewed by 75 different academics from The Design Society. Birkhofer and Zhao found a substantial degree of variation in the results of the reviews, supporting the concerns over a lack of consistency. The authors made various conclusions regarding the 75 reviews and the academics that conducted them. They found that an objective appraisal of a paper's scientific contribution is not feasible, and that the character traits and scientific background of the reviewer affect review process, and that reviews conducted in haste were more critical. However, based on their study method little can be determined about the individual's review process. It is not known whether there exists a correlation between experience and review strategy, or if there are different review strategies amongst academics or a common model. This paper aims to investigate review strategy by repeating the study of Birkhofer and Zhao with the addition of eye-tracking to monitor and measure the process of peer-reviewing a conference paper. For the purpose of this study review strategy encompasses two interrelated dimensions. The first is the sequence in which sections of a paper are viewed. The second concerns the distribution of engagement across the sections.
During Design 2014, seventeen academics attending the conference participated in an experiment in which they were asked to review a conference paper from a Design Society proceeding. While reviewing the paper the individual's had their gaze recorded. The findings from that experiment are discussed in this paper and two research questions are addressed:

1. Do different review strategies exist and what are they?
2. How do character traits of reviewers, such as topic familiarity and review experience, affect review strategy?

2 METHOD

In this section we describe the characteristics of the population, the metrics used, the stimuli, the experimental procedure and the eye tracking equipment.

2.1 Population

Seventeen academics attending the Design 2014 conference were recruited to take part in the experiment. Academics ranged in position from professors to doctoral students. Six nationalities were represented with all participants possessing fluent English. The average age was 37.5 (SD 10.9) with an average number of years of review experience of 6.8 years (SD 6.0). One outlier had a review experience of approximately 40 years. Of the participants 14 were male and 3 were female. The characteristics are given in Table 1 and grouped into four classes as described by Birkhofer and Zhao.

Table 1. Population descriptive statistics

Review group	Expert reviewers	Advanced reviewers	Intermediate reviewers	Novice reviewers
Experience level	> 10 years	5 > 10 years	1 > 5 years	0 > 1 years
Average experience (SD)	13.5* (2.1)	7.5 (2.1)	3.3 (0.5)	0.0
Average age (SD)	45.9 (11.4)	33.0 (2.8)	36.0 (5.6)	26.5 (1.7)
Language **	4 Eng 1st, 3 Eng 2nd	1 Eng 1st, 1 Eng 2nd	4 Eng 2nd	2 Eng 1st, 2 Eng 2nd
Sex	5 M, 2 F	2 M	4 M	3 M, 1 F

*One outlier with 40 years experience was excluded from the average
** English as a first language, English as a second language

2.2 Eye tracking metrics and AOI generation

Eye tracking allows the recording of an individual's gaze with reference to a given scene. In this instance a remote eye tracker was used in conjunction with a laptop. The eye tracker provides measures for fixations and saccades. Fixations are when the eye stabilises over a region for a given period of time and the individual can be said to be consciously aware of what they perceive. It is assumed that attention is directly linked to an individual's gaze and so inferences about their cognitive processes can be made. Saccades are the movements between fixations during which information processing of the scene is suppressed (Bojko, 2005; Duchowski, 2007; Holmqvist et al., 2011).
Areas of interest (AOI) are generated within the eye tracking software to denote boundaries between significant regions of the scene. In this experiment AOIs were created around each section of the paper. Fixations occurring within an AOI can then be analysed to provide metrics about the distribution of engagement across the paper's sections. Sections of the paper that spanned multiple pages had individual AOIs with the fixations aggregated post-testing to provide a single engagement value for each section.

2.3 Stimuli

The paper that formed the subject of the peer-review was chosen for two reasons: the topic covered was deemed suitably broad and the paper had been successfully published in the proceedings of IChoRD'13 (Snider, Culley, & Dekoninck, 2013); the variety of content type in the paper was wide-ranging (McAlpine, Hicks, Huet, & Culley, 2006) in terms of distribution of tables, text and figures. Identifying information from the paper was removed to prevent impartiality but the content was otherwise unaltered. The paper comprised 18 distinct sections that are listed in Table 6.

2.4 Procedure

Participants were informed of the Design Society review criteria (see Table 2 taken from Design Conference review form) before being allowed to view the paper. The paper was presented double-paged on a laptop screen with the exception of the first and last pages that were shown individually. A maximum of fifteen minutes was allowed to review the paper though participants could finish earlier if they wished. Participants could navigate the paper freely, using the keyboard to scroll forwards and backwards between pages. Functions such as Find were disabled.
The experiment was conducted in an open area of the conference facility but was relatively free of noise. The equipment was directed against a blank wall and participants were provided with noise-cancelling headphones if they required. Adjacent to the testing area was a large natural light source. This introduced a degree of variability in the ambient lighting, but it was considered acceptable. Before testing began each participant filled in a short questionnaire and was then calibrated for the eye tracker.

2.5 Equipment

A Tobii X2-30 eye-tracker recording at 30Hz on a Dell Precision M4800 Mobile Workstations running Tobii Studio 3.2 was used to perform the experiment. All analysis has been conducted using the Tobii Studio software using an IV-T fixation filter (default settings), Excel and SPSS V21.

3 RESULTS

Eye tracking technology has dramatically improved in recent years becoming more reliable and robust in dealing with head movement. Portable systems are small and lightweight facilitating in-situ experimentation. However, eye trackers still encounter issues in recording data and there is a limit to their capability. The X2-30 Eye Tracker samples an individual's gaze at 30 Hz and is designed to be used in conjunction with a laptop. The model of tracker used and its set up are such that data capture reliability is exchanged for portability and so data must first be checked for completeness. Using the Tobii studio recording sample rate, a measure of the percentage duration for which the tracker successfully tracked the participant's gaze, a threshold of 80% was set for inclusion in analysis. On this basis two of the seventeen participants were excluded from gaze metric analysis.

3.1 Review responses

After viewing the paper participants were instructed to enter their review via the experimental software. The review criteria were multiple choice with an explicit score attached to each statement (Table 2).

Table 2. Review questions and scoring criteria for Design Society conference proceedings

Q01 - Quality of content:	**Q02 - Significance for theory or practice:**
10 - Excellent work and a significant contribution 08 - Good work, significant 06 - Solid work 04 - Weak content 02 - Only an insignificant contribution 0 - Questionable work	10 - Very significant 08 – Significant 06 - Not bad 04 - Low significance 02 - Only of marginal significance 0 - Outdated work
Q03 - Originality and innovativeness:	**Q04 - Quality of presentation:**
10 – Ground breaking 08 - A pioneer work 06 - One step forward 04 - Better works on the same topic exist 02 - This has been said several times 0 - Outdated work	10 - Excellently written 08 - Well written 06 – Legible 04 - Needs some revision 02 - Requires considerable work 0 - Not acceptable
Q05 - Overall quality	**Q06 - Familiarity of topic: ***
10 - Very high quality 08 – Good quality 06 - Borderline quality 04 - Low quality 02 - Minor quality 0 - Has no merit	10 - Very familiar with the topic, my area of expertise 08 - Good knowledge 06 - More or less familiar 04 - Only marginally familiar 02 - Not really familiar 0 - Completely new to me

*Familiarity of topic is not used to compute the paper's score, merely to indicate the reviewer's knowledge of the subject area.

Review responses have been segregated according to the character traits of individual reviewers:

Table 3. Review scores by review experience

Review experience group	**Average score**	**Standard deviation**	**No. of responses**
Expert	32/50	9.3	7
Advanced	30/50	5.7	2
Intermediate	30.5/50	8.7	4
Novice	32/50	3.7	4

Table 3 shows the average review scores of individuals as grouped by the level of their review experience. The relatively high standard deviation for the more experienced reviewers indicates a degree of inconsistency within the groups with some individuals rating as high as 44/50 and others as low as 16/50.

Table 4. Review scores by topic familiarity

Subject experience group	**Average score**	**Standard deviation**	**No. of responses**
Very familiar with the topic, my area of expertise	0.0	0	0
Good knowledge	29.1/50	6.6	7
More or less familiar	31.5/50	16.8	4
Only marginally familiar	34.5/50	7.2	4
Not really familiar	33.0/50	1.4	2

The review scores are grouped and averaged for all participants by their degree of familiarity with the paper's topic (Table 4). Again the relatively high standard deviation for reviewers with a higher degree of knowledge of the paper's subject indicates a lack of consensus on the paper's quality.

Table 5. Review for expert reviewers with good subject knowledge

Subject and review experience group	Average score	Standard deviation	No. of responses
Expert reviewer, good subject knowledge	31.5/50	5.3	4

Reviewers possessing expert review experience and good familiarity of the paper's topic are averaged. Once again, there is little difference from previous results, though this is to be expected to a degree as they are drawn from the same small population (Table 5).

3.2 Engagement distribution

The distribution of the reviewer's engagement over the sections of the paper, measured by dwell time, is used to indicate the relative importance placed on sections within the paper to determine quality (Table 6). Dwell time is the total time of all fixations spent looking within a section.

Table 6. Engagement distribution across paper sections as measured by dwell time for review expertise groups

				Review expertise group			
Paper section	**AOI Identifier**	**Word count**	**Expected Dwell***	**Expert**	**Advanced**	**Intermediate**	**Novice**
Title	1	10	0.2%	1.1%	1.8%	1.5%	1.6%
Abstract	2	157	3.6%	13.2%	11.4%	9.5%	10.4%
Introduction	3	760	17.2%	11.2%	17.2%	19.1%	22.4%
Section 2	4	31	0.7%	1.0%	0.8%	1.2%	0.7%
Section 2.1	5	491	11.1%	15.1%	12.9%	18.0%	11.9%
Section 2.2	6	448	10.2%	6.5%	6.4%	6.3%	8.0%
Section 2.3	7	135	3.1%	3.1%	3.0%	2.4%	1.8%
Section 2.4	8	104	2.4%	4.4%	3.3%	3.5%	3.4%
Section 2.5	9	133	3.0%	4.8%	4.9%	3.8%	4.7%
Section 2.6	10	401	9.1%	6.5%	5.4%	6.3%	4.4%
Section 3	11	97	2.2%	2.4%	2.4%	1.2%	1.5%
Section 3.1	12	264	6.0%	7.1%	4.7%	3.7%	8.5%
Section 3.2	13	489	11.1%	7.1%	6.6%	6.8%	8.5%
Section 3.3	14	94	2.1%	1.4%	1.1%	2.7%	1.5%
Section 4	15	148	3.4%	5.4%	5.6%	3.2%	3.8%
Conclusions	16	298	6.8%	7.3%	8.6%	8.7%	4.2%
Acknowledgements	17	21	0.5%	0.3%	0.4%	0.5%	0.4%
References	18	326	7.4%	2.2%	3.4%	1.6%	2.3%

*Expected dwell is calculated by determining the proportion of information, as measured by word count, within each section of the paper.

Dwell proportions in excess of the expected dwell are assumed to indicate a higher degree of importance placed on that section by the reviewer. A one-way ANOVA test was performed to determine the effect of reviewer expertise and subject familiarity on the proportion of dwell for each section of the paper (Table 7).

Table 7. One-way ANOVA results for the effect of review experience and subject familiarity on the engagement distribution for sections of the paper

Test Variable	Paper section	Section content	P-value
Review experience	Abstract	Summary of entire paper	0.042
	Section 3.3	Discussion of the results	0.002
Familiarity	Section 2.6	Coding procedure, paper's contribution	0.013

The paper's lead author was consulted to confirm the contribution of each section (see column three, section content, of Table 7) to the whole of the paper. Review experience had a significant effect on the dwell proportion of the Abstract and section 3.3. Subject familiarity had a significant effect on the dwell proportion of section 2.6. All other sections for both test variables showed no significance.

3.3 Engagement sequence

The sequence of sections that were fixated upon can be plotted as a time-series in which the x-axis represents the time at which a fixation occurred, and the y-axis represents the linear order of the paper's sections. Each section is demarcated as an area of interest (AOI) and given a numerical identifier with the main title starting at 1 and proceeding through the paper till the reference section (AOI 18). The AOI identifiers are listed for each section of the paper in Table 6.

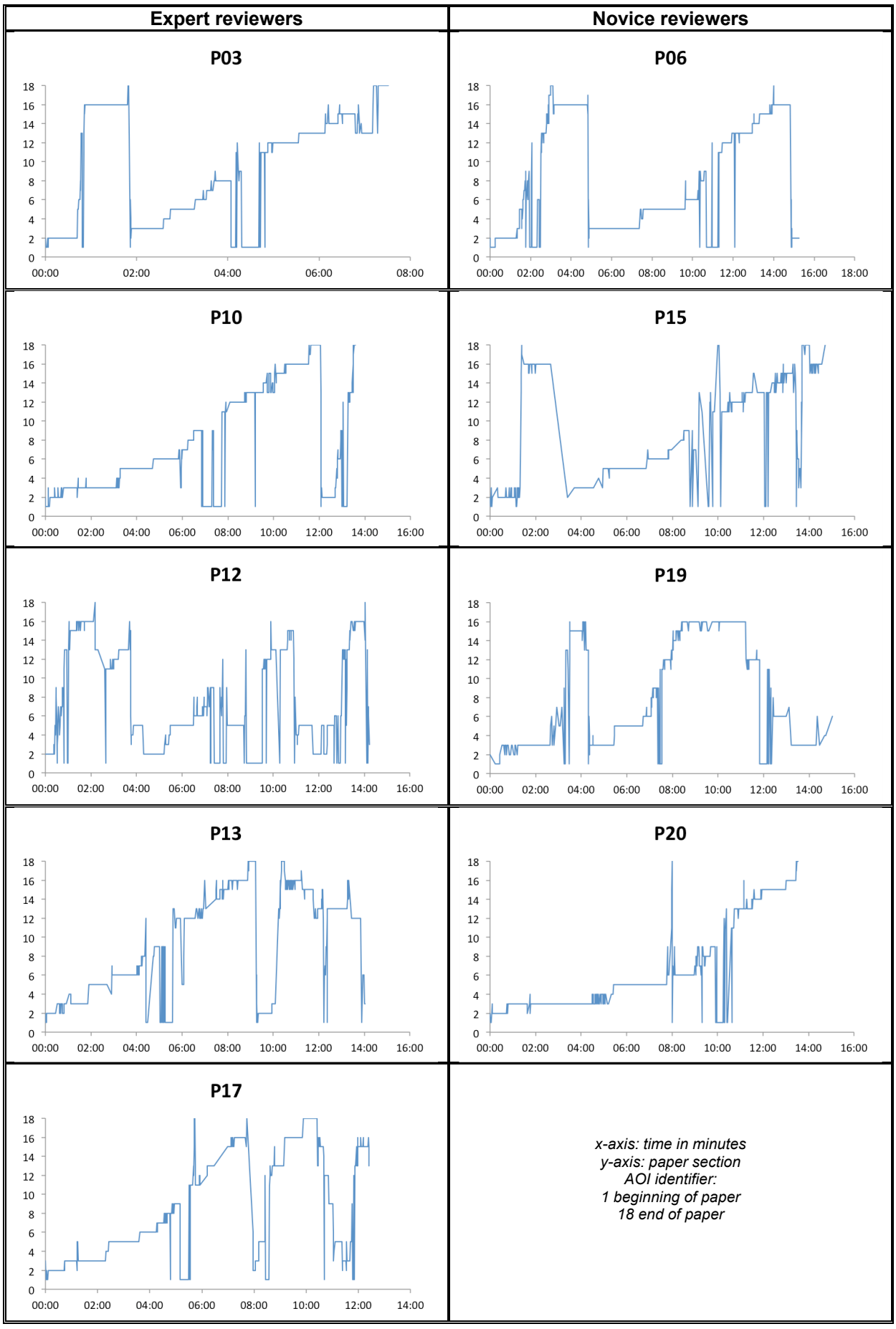

Figure 1. Time-series of section sequence of engagement for expert and novice reviewers

Sections that span multiple pages have had the dwells calculated for the entire section by aggregating fixations from both pages. A fixation that is located within an AOI is then plotted on the y-axis. Flat sections of the time-series plots are where successive fixations occur within the same AOI. A linear progression through the paper, reading from start to finish without returning to a previously read section, would be represented as a line increasing with step changes between each section. The time-series plots of expert and novice reviewers are shown in Figure 1.

4 DISCUSSION

This section discusses the results of the experiment for three areas. The review scores are first compared to those conducted in Birkhofer and Zhao's (2010) experiment. The results from the One-Way ANOVA tests are discussed, and finally, observations on reviewer strategies are made.

4.1 Review responses

From Birkhofer and Zhao (2010) the following hypotheses are re-examined:

- *"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."*

4.1.1 H1 – The effect of the reviewer's background on review outcome

Cultural effects on the outcome of the review have not been investigated in this experiment. However, the subject familiarity, an element of the reviewer's background, has. Birkhofer and Zhao's findings are corroborated with reviewers more familiar with the subject topic tending to be more critical of the paper than those less knowledgeable of the area. However, the standard deviation for reviewers more familiar with the subject area remains relatively high compared to less knowledgeable reviewers suggesting that inconsistency remains an issue.

4.1.2 H2 – The effect of the reviewers experience on review outcome

A reasonable expectation may be that experience in reviewing would develop an individual's internal rating criteria and attune them to look for specific aspects within a paper. Birkhofer and Zhao found no clear evidence of this and the findings from this experiment corroborate this. Again, the standard deviation for reviewers across all experience groups remains relatively high suggesting inconsistency is independent of review experience (Table 3). Research question 2 is therefore concluded: review experience has no apparent effect on review score.

4.2 Engagement distribution

The proportion of the total word count within each section can be used as an estimate for how much time can be expected by the reviewers to engage with that section. The relative importance placed on each section for each expertise group can be inferred from measuring the difference between the actual and expected dwell time (Table 6).

For the Abstract section individuals with a higher degree of experience spent a greater proportion of their time engaging with that section. Conversely for the Introduction section the opposite is true, with Novices spending the most time engaging with the section.

For all expertise groups, sub-sections 2 and 3 (AOIs 4-14) were engaged with at a level of engagement would be expected given the word count.

In contrast, section 4, a discussion section, and the Conclusions all showed a moderate positive correlation between engagement proportion and review experience. More experienced reviewers spent more time on these sections suggesting a prioritisation of them for review purposes.

The one-way ANOVA results (Table 7) have been performed with individual participant's dwell distributions and not the averaged proportions used in Table 6. The results show fewer sections with significant effects but support the conclusion that experts spend a greater proportion of time on the abstract and discussion sections. As to be expected, the degree of reviewer subject familiarity affected the significance of dwell distribution for a section salient to the paper's contribution (as confirmed by the original author).

4.3 Engagement sequence

Review behaviour visualisation allows for the ready comparison of the individual processes demonstrated by each reviewer. Comparing only experts versus novices in Figure 1, certain traits can be observed. When designing the study it was considered likely by the authors that more experienced reviewers would exhibit a greater degree of pairwise comparisons between sections as well as a non-linear approach to the sequence of sections engaged. Similarly, novices were hypothesised to adopt a linear approach to reviewing the paper, following the natural structure.

Three out of four novices show a similar pattern in review sequence, firstly reading the abstract and then the conclusions. They then proceed to read the paper in a linear manner with occasional regressions to previous sections. Rapid oscillations between sections is as a result of navigating through the paper and making short target locating fixations across a number of sections. Pairwise comparisons would manifest as relatively small amplitude oscillations with flat sections between to indicate a degree of information processing.

Experts make more regressions back to the abstract supporting the findings that the abstract has a significant effect on determining the review score for the paper. In general the experts demonstrate a more non-linear approach to reviewing the papers and it is suggested that this is as a result of the individuals building up a piecemeal understanding of the paper, independent of the original author's presented structure. The logical order of the paper and the order of salient sections to the review score do not necessarily correspond and it would appear that the behaviour of expert reviewers reflects this.

Experts spend substantially less time per section than novices do, with smaller plateaus. These shorter but more frequent visits could be indicative of a top-down approach to review in which the fundamental aims of the paper are sought out as a priority.

An alternative and equally plausible conclusion is that review strategy is an independent personal characteristic. A number of potential strategies are apparent; a first-pass strategy in which the whole paper is examined in a short period of time with successive repeat visits for more detailed examination; a strictly linear examination of the paper as dictated by the formal structure; and a 'flip-flop' strategy of large amplitude shifts of engagement between distant sections of the paper's formal structure. However, the small sample size and constraint of the original research questions make further investigation of these potential strategies outside of the scope of this paper. Future research could be conducted to test for the existence of review strategies and how they may be affected by paper topic and reviewer background characteristics.

5 CONCLUSIONS

Character traits of reviewers have an affect on the process by which they review papers in terms of sequence and distribution of engagement of sections (research question 2). This is most strongly exhibited when comparing groups of individuals on their degree of review experience. However, there is no clear association between character traits and review scores, supporting findings from earlier work (Birkhofer & Zhao, 2010). Review scores still show a high degree of inconsistency between individuals, regardless of experience, suggesting little progress has been made since the issue was last addressed. The argument for different review strategies existing is strong (research question 1), though how strategies correlate to individuals may not be best described by review experience.

Review behaviour visualisations show that experienced academics prioritise reading papers by abstract and conclusions, supporting the proposition that different review strategies exist (research question 1). Further investigation is required to determine the characteristics of different strategies and to correlate them to reviewer characteristics. Given that experienced reviewers spend a great proportion of time on the abstract and conclusions a stronger emphasis should be placed in Design Society guidelines about the importance of these sections for authors and reviewers alike.

Of particular interest is the higher number of regressions in the paper structure that experienced academics make, suggesting a behaviour that is determining the paper's internal consistency across sections. It may also arise as a result of a greater proportion of attention capacity devoted to checking content, as familiarity with the review task and goals requires less conscious effort. This would be a worthwhile avenue of further investigation by incorporating retrospective think-a-loud protocols into the tracking study as well as more in-depth review scoring for individual sections.

It is plausible that the inconsistency in review scores is a result of a wide variation in review behaviours as demonstrated in Figure 1. While general trends can be described from the review

behaviour visualisations, there is still substantial difference between each time-series plot. Each academic paper is unique but there are elements that should be universally present. Therefore, it is reasonable to expect that any reviewer should actively confirm and assess these elements. A recommended paper structure, or a checklist for key elements and a paper's internal consistency, would begin to address this. Providing reviewers with more detailed review criteria would also help and it is strongly recommended that the Design Society actively disseminate expectations for the contents of submitted papers and these requirements are harmonised with the review criteria. Ambiguity from the broad assessment criteria inevitably leads to variation in how they should be interpreted. Clearer, more detailed assessment criteria, as well as simple binary checklists could contribute to a more transparent review process for the society. Training at all levels of reviewer experience could also be used to improve consistency, though this may be harder to implement. Benchmarking, either against standard papers, or a reviewers history of reviews for the society may also provide a means of moderating review variability.

No assumption is made as to the superiority of one review strategy over another at this stage. It is uncertain as to whether the apparent range of strategies employed by reviewers leads to inconsistency in the associated review scores, though it is expected that it does not. Determination of distinct review strategies requires an appreciation for the breadth of content and how it can be represented within academic papers. As eye movements are highly idiosyncratic, even for a single person, potential strategies for review based on eye movements may be numerous. Whilst the findings from this paper are in alignment with previous studies further research would necessitate a larger sample to adequately describe review strategies.

The health of peer review is at the core of any research community. The Design Society is well established and enjoys respect for the quality of its proceedings. However, much can be improved and both authors and reviewers could benefit from a fairer and more rigorous peer assessment system.

This short experiment would benefit from being conducted periodically at Design Society conferences with a broader spectrum of individuals and papers used. It is hoped that doing so would contribute to the continuing success of the society.

ACKNOWLEDGEMENTS

This research has been conducted with the support of the EPSRC UK as part of the Language of Collaborative Manufacturing research programme.

REFERENCES

BIRKHOFER, H., & ZHAO, S. (2010). The long-running issue of Review Quality–findings from an empirical study amongst international reviewers. *DS 60: Proceedings of DESIGN 2010, the ...*, 1–10.

BOJKO, A. (2005). Eye tracking in user experience testing: How to make the most of it. In *Proceedings of UPA '05*. Montreal, Canada.

DUCHOWSKI, A. (2007). *Eye Tracking Methodology Theory and Practice* (2nd ed., p. 360). London: Springer London. Retrieved from

HOLMQVIST, K., NYSTRÖM, M., ANDERSSON, R., DEWHURST, R., JARODZKA, H., & WEIJER, J. VAN DE. (2011). *Eye Tracking: A comprehensive guide to methods and measures* (p. 560). Oxford University Press.

MCALPINE, H., HICKS, B. J., HUET, G., & CULLEY, S. J. (2006). An investigation into the use and content of the engineer's logbook. *Design Studies*, *27*(4), 481–504. doi:10.1016/j.destud.2005.12.001

SNIDER, C. M., CULLEY, S. J., & DEKONINCK, E. A. (2013). relative quality for the study of creative design output . In : ICoRD ' 13 International Conference on Research into Design ,