5th ISCA/DEGA Workshop on Perceptual Quality of Systems (PQS 2016)
29-31 August 2016, Berlin, Germany

Size does Matter. Comparing the Results of a Lab and a Crowdsourcing File Download QoE Study.

Andreas Sackl[1], *Bruno Gardlo*[1], *Raimund Schatz*[1]

[1]AIT Austrian Institute of Technology
Giefinggasse 2, 1210 Vienna, Austria
{andreas.sackl|bruno.gardlo|raimund.schatz}@ait.ac.at

Abstract

Over the last couple of years, crowdsourcing has become a widely used method for conducting subjective QoE experiments over the Internet. However, the scope of crowdsourced QoE experiments so far has been mostly limited to video and image quality testing, despite the existence of many other relevant application categories.

In this paper we demonstrate the applicability of crowdsourced QoE testing to the case of file downloads. We conducted several campaigns in which participants had to download large (10-50MB) media files (with defined waiting times) and subsequently rate their QoE. The results are compared with those of a lab-based file download QoE study featuring an equivalent design.

Our results show that crowdsourced QoE testing can also be applied to file downloads with a size of 10 MB as rating results are very similar to the lab. However, beyond user reliability checks and filtering, we found the study design to be a highly critical element as it exerted strong influence on overall participant behavior. For this reason we also present a discussion of valuable lessons learned in terms of test design and participant behavior.

Index Terms: quality of experience, crowdsourcing, downloads, waiting times

1. Introduction

In QoE research, various empirical research methods can be applied to gather subjective quality perception assessment data, methods like laboratory studies [1], field trials [2] and crowdsourcing campaigns [19]. During the last years, crowdsourcing studies have been increasingly used to complement and/or substitute more traditional methods like laboratory QoE studies. Nevertheless, for some contexts and applications it is not obvious whether crowdsourcing approaches are applicable or create similar results to a laboratory setting. In the proposed paper we address this issue by comparing and discussing the results of four empirical QoE user studies (one lab study LAB, three crowdsourcing studies CS1, CS2 and CS3) to evaluate whether it is possible and reasonable to conduct Web QoE file download studies via crowdsourcing campaigns. To this end, we will answer the following research questions in the paper:

- RQ1: For file downloads, are subjective network quality rating results similar between lab and crowdsourcing studies?
- RQ2: How does explicitly displaying file size impact participant behavior in a crowdsourcing study?
- RQ3: Are crowdsourcing users willing to download large files in the context of a testing campaign?

2. Related Work

The investigation of the utilization of crowdsourcing for QoE assessment started roughly five years ago, primarily motivated by the possibility to generate large volumes of subjective testing data sourced from a global pool of participants by conducting Web-based studies via the Internet [3]. Since then, crowdsourcing has enjoyed rapidly growing interest by QoE researchers not only as new cost-effective way of conducting subjective tests, but also as object of methodological study itself [4]. In particular, the remote, less-controlled nature of crowdsourcing poses a number of challenges to its application to QoE assessment. These challenges relate to fundamental differences between crowd-based and laboratory-based QoE evaluation studies with regard to conceptual, technical and motivational aspects [7] which might compromise reliability and validity of the results generated.

In terms of applications, crowdsourced QoE assessment so far has been primarily used for testing online video scenarios (e.g. [6, 8, 20, 7]), not only because web-based evaluation works very well for testing short media clips without requiring special hardware, but also because since test execution setup and context very well match the actual use case (online video watching). Beyond video streaming, the method has also been frequently applied to the domains of image quality assessment (e.g. [9, 10, 11, 12]) as well as audio and (synthetic) voice quality testing (e.g. [13, 14, 15]). In contrast, Crowdsourcing has been much less used for Web QoE research (e.g. in the context of Web browsing quality [16]). In particular, the method so far has not been applied to evaluating the QoE impact of waiting times in the context of file downloads, despite this scenario's relevance and previous lab and field studies [17, 18, 1]. This gap in existing research motivates the investigation of the feasibility and applicability of QoE crowdsourcing to the case of file downloads by conducting crowdsourcing studies and comparing the results with the lab.

3. Study Design

In our download QoE crowdsourcing studies, participants had to click on video links in an individual order (overall 6 music videos were available), thereby triggering a (simulated) download. This download was accompanied by a visual progress bar indicating the remaining download time (2 to 29 seconds) before the short videos (duration was 20 seconds) could be watched (Note: actually, the videos were already downloaded at the beginning of the campaign, i.e., the download duration was simulated). Figure 1 depicts the download progress bar which was displayed in CS1, CS2 and CS3. The download progress bar guarantees that the task is perceived as a "down-

10.21437/PQS.2016-27

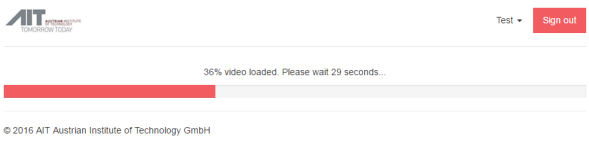

Figure 1: Screenshot of the download progress bar used for the crowdsourcing studies CS1, CS2 and CS3.

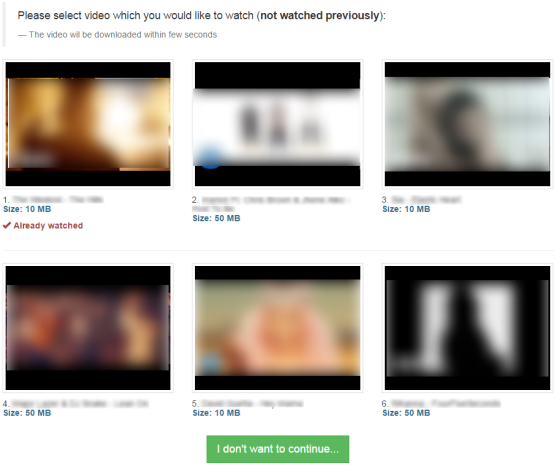

Figure 2: Screenshot of the video selection page of the crowdsourcing studies CS2 and CS3.

load task" instead of an "initial delay" of video playback. Participants were able to select the order of the downloads individually and they were able to abort the campaign anytime resulting in a lower amount of paid out money of course. Figure 2 depicts the video selection screen used in the crowdsourcing campaigns CS2 and CS3. All videos are labeled with file size information and during any time of the trial the participant could abort the study by clicking on the button "I don't want to continue", which of course results in a lower amount of paid out money. The conducted crowdsourcing studies took place in three campaigns (CS1, CS2 and CS3), differing regarding the presence of a *label* denoting the file size (10 or 50 MB) of each video and regarding some additional questions at the end of the study, see Table 1 for an overview.

We used the well established crowdsourcing platform microworkers.com, which is frequently used in various experiments in the field of QoE [20, 22]. To avoid any bias due to varying expectations among globally sourced workers, for all three campaigns we selected only workers from Europe and United States. In the first campaign CS1 65 workers entered our web-application, out of which 45 successfully finished the whole experiment. In the second campaign CS2 50 out of 77 workers successfully finished the same task. In the third campaign CS3 51 out of 70 workers successfully finished. In the crowdsourcing campaign CS3 after the participants selected and consumed the last video (or if the user aborts the study), following three questions were asked:

1. "Is the Internet connection you are using now your private connection?" with answering options "Private" and "Not private"
2. "Do you consider it as a problem, that you had to download these video files in order to watch it?" with answering options "yes" and "no"
3. "What was the size of the last video?" with answering options "1 MB", "5 MB", "10 MB", "30 MB", "50 MB" and "Dont know".

Study	File size labeling	Videos selectable	additional questions
CS1	no	yes	no
CS2	yes	yes	no
CS3	yes	yes	yes

Table 1: Overview of crowdsourcing studies CS1–CS3.

File Size	DL Time	Equal?	Significance
10 MB	2s	same	H = 0.192, p= 0.662
10 MB	6s	same	H = 0.827, p= 0.366
10 MB	20s	same	H = 0.174, p= 0.678
50 MB	11s	diff	H = 5.076, p= 0.027
50 MB	14s	same	H = 1.451, p= 0.232
50 MB	29s	same	H = 3.725, p= 0.057

Table 2: One-way ANOVA test results comparing lab LAB and crowdsourcing studies CS1 and CS2 (confidence level: 95%).

The lab study (LAB) was conducted with 52 participants (26 female, 26 male) with a mean age of 37.7 years (median=31 years). In our laboratory, the users had to click on a link on a website on a Laptop to download and watch several video files (10 and 50 MB) via various downlink bandwidths (4, 14, 30, 45 Mbit/s), resulting in different download durations (2 to 29 seconds). Similar to the crowdsourcing campaigns, after each download and video consumption the users had to evaluate the perceived network speed via a continuous 5 point ACR scale [23], asking about the annoyance of the user regarding the download time with answering options ranging from 5=not annoying to 1=very annoying.

4. Empirical Results

Figure 3 shows the resulting MOS values separated by the file size (10 MB and 50 MB). Note that the ratings of the crowdsourcing campaigns CS2 and CS3 can be combined, because the difference in the test designs did not affect the ratings. For 10 MB files, there are no differences between the laboratory and the crowdsourcing setting, whereas the participants of the crowdsourcing setting evaluated slightly more negatively when 50 MB files were downloaded. Our one-way ANOVA tests conducted confirm this observation: lab and crowdsourcing results are similar for conditions featuring 10 MB files with high confidence, while the lab–crowd rating differences for the 50 MB conditions are borderline in terms of statistical significance (see Table 2).

Figure 4 provides a reasonable explanation for the (small) differences between the lab and the crowdsourcing experiments: Obviously, during the first three user decisions most of the participants selected smaller files (10 MB), whereas towards the end of the task the larger file sizes were selected more often (50 MB). So, undesired side effects like fatigue or boredom could have caused the lower ratings for larger file sizes, i.e., crowdsourcing test participants were more and more annoyed during

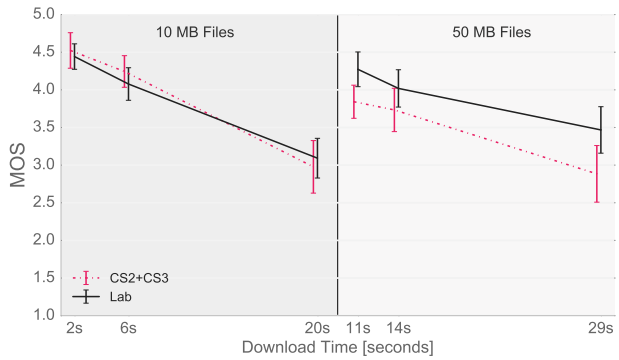

Figure 3: QoE results for for different download times and file sizes (10 MB, 50 MB) in the studies CS2 and CS3 (5=not annoying to 1=very annoying).

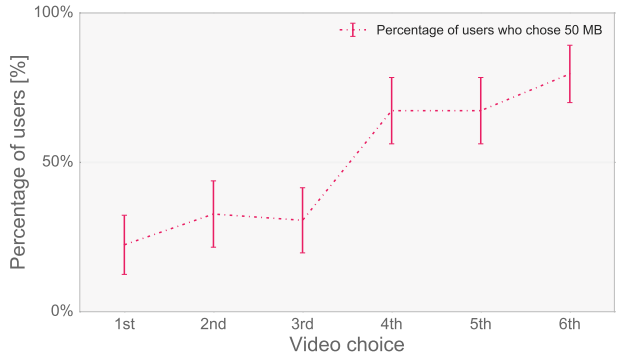

Figure 4: Percentage of users who selected a 50 MB file in campaign CS3 over the course of a test session.

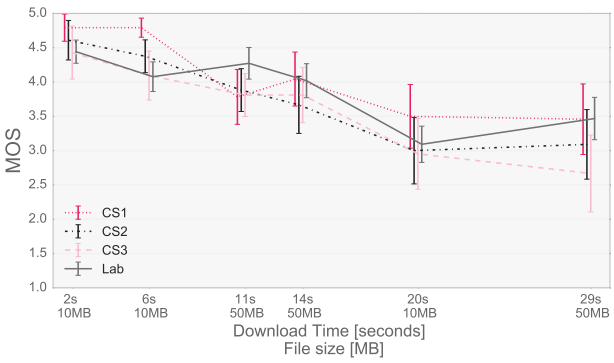

Figure 5: QoE rating results for different download times in the studies CS1, CS2, CS3 and LAB (Please note the different file sizes, answers: 5=not annoying to 1=very annoying).

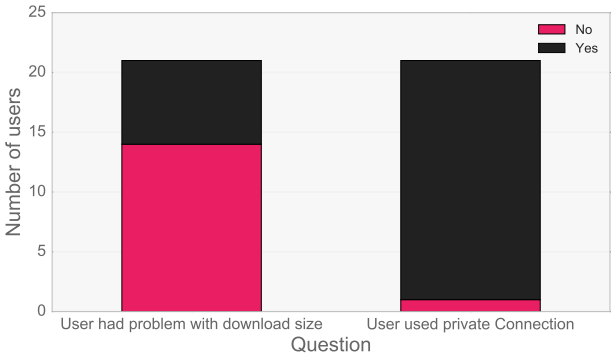

Figure 6: Survey results of crowdsourcing campaign CS3 regarding usage of private data connection and type of connection.

the campaign. Nevertheless, because of the small differences of the resulting MOS values for larger file sizes, we recommend to conduct further studies which further evaluates the mentioned side effects. Note that the depicted crowdsourcing MOS values are based on ratings of participants who downloaded five or six videos, i.e., users who only downloaded four or less videos were excluded from the dataset. According to our analysis, users who downloaded four or less videos were less reliable than users with 5 or 6 downloads, also the ratings are more reasonable for this group.

To provide an overview of all ratings, Figure 5 shows the MOS values and the 90% confidence intervals for the LAB and the crowd studies CS1, CS2 and CS3. So, according to our results the presence of labels regarding the file size is crucial for valid subjective quality assessment tasks, i.e., the MOS values of the study CS1 are not comparable with the lab results, whereas the results of the crowdsourcing studies CS2 and CS3 are comparable to the lab results.

Figure 6 shows the results of the survey which was displayed at the end of the crowdsourcing study CS3 to the participants. Almost all users used their private connection instead of an Internet connection at home or in an Internet cafe. Hence, large file downloads reduce the available, private data plan of the participants. Nevertheless, two-thirds of the participants had no problem with the fact to download large files during the study. So, there is a relatively large acceptance regarding downloading file in crowdsourcing campaigns. Figure 7 depicts how these two user groups ("concerned about file downloads" vs. "not concerned") assess the resulting download durations (please consider the low amount of ratings per user group): For short download durations (up to 14 seconds), there are no differences between these two user group. For longer download durations (20 and 29 seconds), users who are concerned about file downloads provide lower ratings compared to the unconcerned user group. Please note that for the 20 seconds duration a 10 MB file was downloaded, and for the 29 seconds duration a 50 MB file was downloaded. Hence, the file size was not the reason for the different rating behavior, i.e., only the download duration was perceived differently.

Figure 8 shows the abortion rate for each of the three crowdsourcing campaigns CS 1,2 and 3. The bars for the campaigns 1 and 3 display the expected user behavior: Some users abort the trial after one or two selected videos. The longer the duration is of the experiments, the smaller is the cancellation rate, i.e., only very few users abort the trial after 4 or 5 videos, which is reasonable because the highest reward is paid out after consuming all 6 videos. Nevertheless, in the crowdsourcing study CS2 (black bars in figure 8) this behavior did not occur, e.g. a high amount of users aborted the trial after 5 of 6 videos, which is

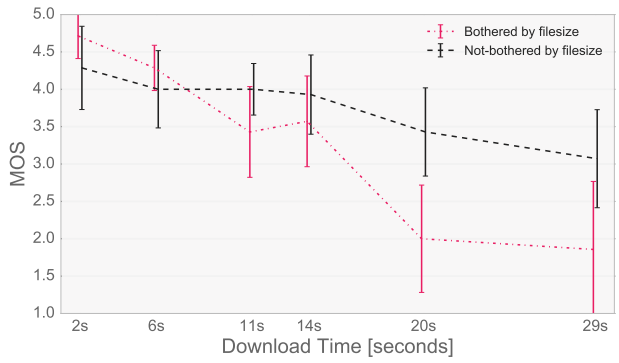

Figure 7: Split of QoE rating results from CS3 according concerned and not concerned users (Based on survey results, see Fig. 6).

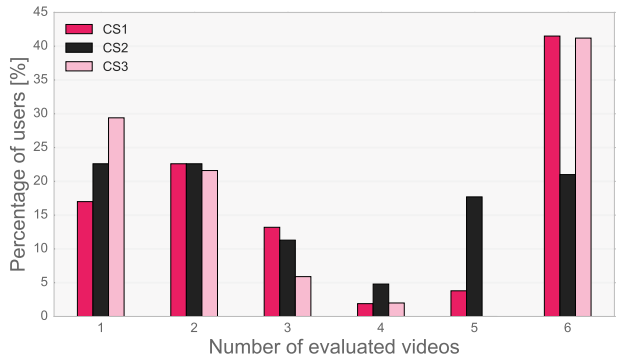

Figure 8: Distribution of users across total number of videos selected and consumed for campaigns CS1–3.

rather unreasonable.

Considering question 3 about the size of the last downloaded video in campaign CS3, only 6 out of 21 participants were able to answer it correctly and 12 answered with "I don't know". Hence, for further studies the labeling should be presented in a better way.

5. Conclusions & Outlook

In this paper, we demonstrated that it is possible and reasonable to conduct crowdsourcing campaigns that investigate how file download time impacts the subjective quality perception regarding Web QoE. So, research question RQ1 can be positively answered — given that the setup is appropriate (i.e. file sizes are labeled) and small file sizes are used.

Regarding research question RQ2, the presence of file size information is mandatory for obtaining valid results. In the described crowdsourcing setup, our participants were able to make two essential decisions: a) in which order the videos are downloaded and watched and b) possible cancellation of the trial after each selected video. Ad b) except for the second crowdsourcing study CS2, the observed cancellation behavior was as expected, i.e., the cancellation rate decreased towards the last video because of the related high reward after finishing all six videos. The cancellation rate has its maximum at the beginning of the trial, i.e., in further campaigns participants need to be motivated especially at the beginning to continue with the experiment. Ad a) As shown in Figure 4, the crowdsourcing participants downloaded first the smaller 10 MB videos before downloading the 50 MB videos. This selection order needs to be considered in further crowdsourcing test designs to avoid unwanted side-effects. Please see also [24] for the influence of individual choices on QoE.

Regarding RQ3, approx. two thirds of our crowdsourcing participants stated that downloading files during the task was not a problem (Note: almost all participants used their private Internet connection, i.e., their download volume was used during the campaign). Regarding the resulting MOS values, for download durations up to 14 seconds there was no difference between these two groups (having an issue with download an not having an issue). For longer download durations, participants which had a problem with file downloads evaluated the download duration more negatively compared to the other group. Because of the small amount of users per group a final conclusion is not possible, but for further experiments we recommend to ask the participants if file downloads are problematic for them to have the option for further filtering.

6. References

[1] A. Sackl, S. Egger and R. Schatz, "The influence of network quality fluctuations on Web QoE," Quality of Multimedia Experience (QoMEX), 2014 Sixth International Workshop on, Singapore, 2014, pp. 123-128.

[2] P. Casas, B. Gardlo, M. Seufert, F. Wamser and R. Schatz, "Taming QoE in cellular networks: From subjective lab studies to measurements in the field," Network and Service Management (CNSM), 2015 11th International Conference on, Barcelona, 2015, pp. 237-245.

[3] K. Mao, L. Capra, M. Harman, and Y. Jia, "A Survey of the Use of Crowdsourcing in Software Engineering," RN, vol. 15, p. 1, 2015.

[4] T. Hoßfeld and C. Keimel, "Crowdsourcing in QoE Evaluation," In: "Quality of Experience: Advanced Concepts, Applications and Methods," Ed. by S. Möller and A. Raake, Springer: T-Labs Series in Telecommunication Services, ISBN 978-3-319-02680-0, Mar. 2014.

[5] C. Keimel, J. Habigt, and K. Diepold, "Challenges in crowd-based video quality assessment," in 2012 Fourth International Workshop on Quality of Multimedia Experience (QoMEX), 2012, pp. 13-18.

[6] K.-t. Chen, C.-j. Chang, A. Sinica, C.-c. Wu, Y.-c. Chang, and C.-l. Lei, "Quadrant of Euphoria: A crowdsourcing platform for QoE assessment," IEEE Network, April, pp. 28-35, 2010.

[7] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "Video quality evaluation in the cloud," In 19th International Packet Video Workshop (PV), pp. 155-160, 2012.

[8] B. Gardlo, M. Ries, M. Rupp, and R. Jarina, "A QoE Evaluation Methodology for HD Video Streaming Using Social Networking," in 2011 IEEE International Symposium on Multimedia (ISM), 2011, pp. 222-227.

[9] F. Ribeiro, D. Florencio, and V. Nascimento, "Crowdsourcing subjective image quality evaluation," in IEEE International Conference on Image Processing, September 2011.

[10] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," IEEE Transactions on Image Processing, vol. 25, no. 1, pp. 372-387, January 2016.

[11] Q. Xu, Q. Huang, and Y. Yao, "Online crowdsourcing subjective image quality assessment," in ACM International Conference on Multimedia, 2012, pp. 359-368.

[12] J. Redi and I. Povoa, "Crowdsourcing for Rating Image Aesthetic Appeal: Better a Paid or a Volunteer Crowd?," 3rd International

ACM workshop on Crowdsourcing for Multimedia (CrowdMM 2014), Orlando, FL, USA, Nov 2014.

[13] J. Parson, D. Braga, M. Tjalve, and J. Oh, "Evaluating Voice Quality and Speech Synthesis Using Crowdsourcing," Text, Speech, and Dialogue, in: I. Habernal and V. Matouek (Eds.) Springer Berlin Heidelberg, pp. 233-240, 2013.

[14] B. Naderi, T. Polzehl, A. Beyer, T. Pilz, and S. Möller, S., "Crowdee: Mobile Crowdsourcing Micro-task Platform for Celebrating the Diversity of Languages," Proc. of the 15th Annual Conference of the International Speech Communication Association (Interspeech), ISCA, 1496–1497, 2014.

[15] T. Volk, C. Keimel, M. Moosmeier, and K. Diepold, "Crowdsourcing vs. Laboratory Experiments - QoE Evaluation of Binaural Playback in a Teleconference Scenario," Computer Networks, vol. 90, no. C, pp. 99-109, October, 2015.

[16] M. Varela, T. Mäki, L. Skorin-Kapov and T. Hoßfeld, "Increasing payments in crowdsourcing: dont look a gift horse in the mouth", In 4th international workshop on perceptual quality of systems (PQS), Vienna, Austria, 2013.

[17] P. Reichl, S. Egger, R. Schatz, and A. DAlconzo, "The Logarithmic Nature of QoE and the Role of the Weber-Fechner Law in QoE Assessment," in Communications (ICC), 2010 IEEE International Conference on, 2010, pp. 1-5.

[18] R. Schatz and S. Egger, "Vienna Surfing - Assessing Mobile Broadband Quality in the Field," in Proceedings of the 1st ACM SIGCOMM Workshop on Measurements Up the STack (W-MUST), 2011.

[19] B. Gardlo, S. Egger, M. Seufert and R. Schatz, "Crowdsourcing 2.0: Enhancing execution speed and reliability of web-based QoE testing," 2014 IEEE International Conference on Communications (ICC), Sydney, NSW, 2014, pp. 1070-1075.

[20] T. Hoßfeld, M. Seufert, M. Hirth, T. Zinner, P. Tran-Gia and R. Schatz,"Quantification of YouTube QoE via Crowdsourcing," *Multimedia (ISM), 2011 IEEE International Symposium on*, Dana Point CA, 2011, pp. 494-499.

[21] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, and K. Diepold, "Best practices for QoE crowdtesting: QoE assessment with crowdsourcing," IEEE Transactions on Multimedia, vol. 16, no. 2, 2014, pp. 541-558.

[22] T. Hoßfeld and J. Redi, "Journey through the crowd: Best practices and recommendations for crowdsourced QoE," *Quality of Multimedia Experience (QoMEX), 2015 Seventh International Workshop on*, Pylos-Nestoras, 2015, pp. 1–2.

[23] International Telecommunication Union, "Methodology for the Subjective Assessment of the Quality of Television Pictures," ITU-R Recommendation BT.500, 2012.

[24] A. Sackl and R. Schatz, "The influence of user decisions on subjective quality assessment ratings", IEEE Workshop on Quality of Experience-based Management for Future Internet Applications and Services (QoE-FI) - in conjunction with IEEE ICC 2015, London, UK, Sep 2015.