Increasing Payments in Crowdsourcing: Don't look a gift horse in the mouth!

Martín Varela[1], *Toni Mäki*[1], *Lea Skorin-Kapov*[2], *Tobias Hoßfeld*[3]

[1] VTT Technical Research Centre of Finland
[2] University of Zagreb, Faculty of Electrical Engineering and Computing
[3] University of Würzburg, Institute of Computer Science, Würzburg, Germany
{martin.varela,toni.maki}@vtt.fi, lea.skorin-kapov@fer.hr, tobias.hossfeld@uni-wuerzburg.de

Abstract

A commonly cited maxim states that *"you get what you pay for"*, implying that there is a strong correlation between the price paid for something, and its quality. In this paper, we examine whether this often-cited wisdom applies to using crowdsourcing for conducting subjective QoE experiments, and if so, how. As part of a large-scale user study designed to assess Web QoE, we conducted two crowdsourced campaigns to collect user ratings and study the influence of certain website design parameters related to typography and color on the overall visual appeal of the site. While the test content was exactly the same across both campaigns, the second campaign was set up to pay participants three times the reward of the first one. The goal was to analyze the impact of payment on a number of parameters, including the ratio of reliable users and obtained MOS values. With respect to QoE modeling, we found that while payment levels influenced absolute MOS values, there was no significant impact on the actual model.

Index Terms: crowdsourcing, quality assessment, incentives

1. Introduction

Crowdsourcing has gained momentum in the QoE research community as a means to both expedite and reduce the cost of conducting subjective user assessments, while allowing end users to perform tasks in their real-world settings. The idea is to outsource a job (in this case subjective quality assessment) to an anonymous crowd of users in the form of an open call. In the case of existing commercial Internet crowdsourcing platforms such as Amazon Mechanical Turk and Microworkers, "employers" submit certain tasks, while "workers" (referring to widespread Internet users) may complete such tasks for an announced payment. In the context of subjective user studies, such an approach may significantly reduce the time and costs of conducting lab tests and offer access to a large and diverse panel of internationally spread users. These benefits come with a price attached, in terms of the quality of the results, and potential instrumentation difficulties, which make the approach unsuitable for certain types of assessment (e.g., cases when special equipment/devices or controlled end user settings are needed). When conducting Web QoE studies, crowdsourcing seems like a potentially good approach to assess large numbers of test conditions. While the reliability of user ratings needs to be accounted for (mechanisms to detect unreliable users are needed), along with uncontrolled user environments, the results of such studies have shown to deliver results similar to traditional testing in the lab environment [1]. We note that a *reliable* user is considered to be one that expresses true feelings regarding perceived quality, while *unreliable* users may be found to assign random or constant grades when conducting quality assessment, look to finish the assessment as quickly as possible, or not complete all steps related to a given task.

The results presented herein were derived from two large scale crowdsourcing campaigns [2] (>350 users each) designed to assess the impact of certain design factors on the visual appeal of web sites. The first campaign was set up to pay 0.2US$ for each user (worker), and the second one paid 0.6US$ for performing exactly the same task. Given the large gap in incentives, we wanted to study the following:

1. do higher payments attract more unreliable users resulting in lower quality of work?
2. do higher payments influence user ratings and lead to an increase in QoE values reported by users (as opposed to users taking a more conservative rating approach)?

Related studies have found that while participation rates are increased with an increase in payment, data quality (e.g., in terms of reliability, accuracy) seems to be virtually independent from payment levels [3–5]. The authors in [4] suggest the latter to be related to the effect of the user's perception of the quality of their work as related to the level of payment - regardless of payment, user's felt they were payed less than deserved, and thus were not motivated to perform better. In contrast, others have found that financial incentives may encourage improved quality [6], e.g., if a bonus is offered in the case of accurate results [7]. With regards to the quantity of work performed, studies have found that subjects generally worked less when the payment was lower [4, 8]. Other studies that have addressed worker motivation have found that in addition to extrinsic motivation (e.g., financial incentive), intrinsic motivation to complete a task (e.g., enjoyment, social contacts) often plays a key role [9]. In contrast to related work, we focus on incentives and their consequences for QoE testing and modeling.

The rest of the paper is organized as follows. We briefly describe the experimental setup in Section 2. In Section 3 we perform a statistical analysis of the results. Section 4 provides a SWOT analysis of the crowdsourcing approach for Web QoE assessment. We conclude the paper with Section 5.

2. Setup of Crowdsourcing Experiments

The experimental setup was designed to determine the influence of commonly-cited best practices in design related to typography and color on the visual appeal of a website. We considered four different web pages, and for each of those, the number of typefaces used (in three levels) and their suitability — or *good-*

ness to the content and use (also in three levels). Likewise, we considered the number of colors in the palette used, and their suitability to the content (also in three levels each). Best practices in design (e.g. [10, 11] for typography and use of color, respectively) suggest minimizing the number of typefaces and colors used in a design, and making sure that the typefaces are used correctly (i.e., do not use a display type when writing an article, do not use a modern typeface to typeset a medieval text, etc.), and that the color palettes used are harmonic (i.e., smart use of a color wheel).

2.1. Test Content Preparation

Of the four contents we used, three were professional designs, and the fourth was a re-implementation of a simplified version of an Austrian newspaper site (originally made available in [12]). We then used a simplified version of a professionally designed site for the fourth content, and proceeded to "deface" these designs in a controlled way by varying the number and *goodness* of both fonts and color schemes. We created several versions of each test content with different palette size and typeface combinations, treating each factor independently. For color, the "number of colors" is not the total number of colors *per-se*, but an ordering on the size of the palette used.

Regarding the *goodness* of both color palettes and typographic choices, it is an inherently subjective factor, which cannot be easily quantified or determined by rules. There are, however, certain characteristics that are usually desirable in a design. For instance for the typographic aspects, these could be the fonts chosen, their legibility, the compatibility between the typefaces used, congruence with the text contents, etc. For color, there are similar considerations to take into account (although somehow simpler, as color theory is a very well-established field). An example of color manipulation is illustrated in Figure 1 displaying two versions of the same page with the original colours (assumed to be good) and the manipulated colours.

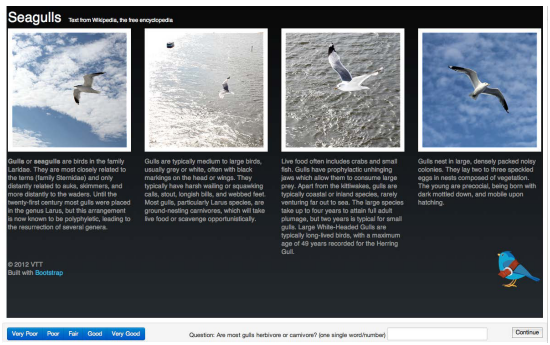
We developed instrumentation for the automatic generation of the test contents, by simply modifying certain variables in the CSS stylesheets (the actual process is slightly more involved, but conceptually the same). Having considered four parameters with three possible values each, we had 81 possible conditions to test. In order to limit the length of the test, we decided to divide those into smaller groups, in which all possible combinations of two parameters were varied, and the other parameters (including the content) were drawn randomly. This approach yielded six such groups of nine conditions, and we repeated it three times, ending up with 18 groups in total. Given that for each condition some of the parameters were randomly chosen, we only covered 72 out of the 81 possible ones. If we further consider the contents separately, there were 128 unique conditions. Each group of conditions was tested by at least 20 subjects. Given the large number of conditions to be tested, the benefits of employing a crowdsourcing approach as opposed to conducting traditional lab experiments are clear, with faster and cheaper access to a large number of test participants.

2.2. Experimental Setup

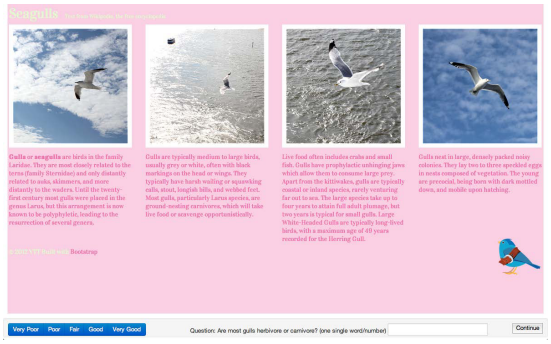
The assessment was carried out by means of a custom-built web application. The users were greeted with a short description of the experiment, in which they were told that for each page displayed, they should rate the design of the site on a 5-point MOS-like scale. We opted for having a simple question (or rather an instruction given at the beginning of the experiment) instead of a more traditional aesthetics questionnaire, as we were not sure about whether the subjects would understand the differences between the different aesthetics dimensions. The rating scale was labelled and not numerical. The possible values were "very poor", "poor", "fair", "good" and "very good". There was no training for the users, and the users had to rate 10 conditions, of which one was repeated, to check for consistency in the ratings. Each page to rate was presented in an HTML iframe and the ratings were collected with radio-buttons below it. There was also a question about the text in the page, which the users had to answer with a single word or number. These questions were meant to insure that users were actually reading the text. The design approaches and statistical methods to quantify reliability and to identify unreliable users are based on [13].

Upon completion of the test, the users were provided with a unique token for claiming their payment. For the tests, the Microworkers.com crowdsourcing platform was used.

Two assessment campaigns were carried out, roughly one month apart, and using two different task compensations (0.20 US$ in the first campaign, and 0.60 US$ in the second one). For a complete analysis of the results of these campaigns from the visual appeal perspective, we refer the reader to [2].



(a) Good color palette



(b) Bad color palette

Figure 1: Evaluation of content "Seagulls" with test application.

3. Analysis of Payments in User Studies

In the following, the two different crowdsourcing campaigns are compared in order to analyze the impact of payments on the reliability (Section 3.2) and the user assessments (Section 3.3).

3.1. Demographics of Test Users

One of the advantages, and also problems, of crowdsourcing is the fact that the population of users is global, and has a certain bias towards developing countries. This allows for a varied userbase, but also may result in a test population not representative of the intended userbase. Table 1 presents an overview of the demographics information for both campaigns. It can be seen that Asian users account for a very large portion of the results, and that the second campaign included a smaller number of countries (roughly 66% that of the first campaign). In C_2, a larger ratio of Asian users from low-wage countries was observed than in C_1, who may have been attracted by the higher payments. However, this change in demographics of the subjects may be simply caused by the time when the campaign was launched and its much shorter duration, due to day-night activities of users and time shift in countries [14]. C_1 started at 31-Aug-2012 07:20:18 CEST, C_2 at 09-Nov-2012 13:49:56 CET. The most likely reason that C_2 had a shorter duration was that following announcement of the campaign, users were quicker to pick the job from the crowdsourcing platform, resulting in quicker completion of the campaign (described further in the following section). We note that in both cases, most subjects self-identified themselves as naïve when asked if they had previous experience in performing this type of task.

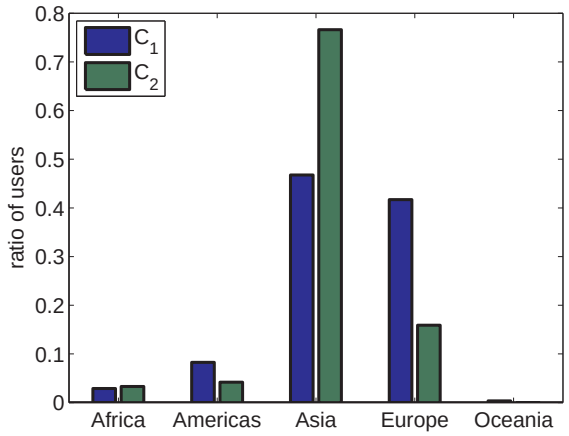


Figure 2: Ratio of users from a certain continent in campaigns C_1 and C_2 with payments P_1 and $P_2 = 3 \cdot P_1$.

Table 1: Demographic information for both campaigns.

Feature	C_1	C_2
Mean age	27.02	25.35
Ratio of female subjects	22.08 %	16.07 %
Number of countries	45	30
Ratio of Asian Users	46.77 %	76.60 %
Ratio of naïve subjects	73.24 %	62.44 %
Ratio of users with vision issues	21.90 %	15.94 %
Ratio of colorblind users	2.18 %	6.24 %

3.2. Reliability and Unreliable Users

In the current implementation of the Microworkers platform, an employer announces a campaign (e.g., to conduct the web QoE test), the features required for a worker (e.g., origin country) and the number N of required workers. Then, any user (fulfilling those features) may pick the job from the platform, and following successful completion enter an obtained payment code to prove his work. The campaign is closed after N different workers have submitted their payment codes, i.e., subsequent workers cannot submit their proof and hence will not be paid. This first-in-first-out (FIFO) principle is reasonable for short microjobs being conducted within one minute (which were the target business for Microworkers.com). However, for more complex and time-consuming tasks like QoE assessments this may cause problems due to unpaid workers. Currently, additional features like task-locking mechanisms are implemented by Microworkers.com for supporting complex tasks and to avoid unpaid work.

Table 2: Comparison of objective measures between campaigns C_1 and C_2 with payments P_1 and $P_2 = 3 \cdot P_1$.

Measure	C_1	C_2
Requested Users	350	450
Ratio of Completed Tests	90.26 %	89.34 %
campaign completion time	173.05 h	2.74 h
Avg. Time for Completed Tasks	8.21 min	9.14 min
Avg. #Correct Content Questions	8.27	7.48
Avg. Consistency	0.32	0.24
Ratio of Reliable users	71.54 %	66.10 %

Table 2 compares the campaign C_1 and C_2 with payments P_1 and $P_2 = 3P_1$ in terms of task completion time, number of correct content questions, consistency of users, as well as origin country of subjects. Due to the FIFO implementation, it may be expected that workers try to complete C_2 faster to ensure getting the reward. However, the crowdsourcing platform shows the workers the number of open and completed jobs for a campaign. Thus, there is no hurry for most of the users. On one hand, the average time for completing the task was significantly higher in C_2 than in C_1. This suggests that high paid users are working more dutiful. Figure 3 shows the cumulative distribution function (CDF) of the task completion time per user. On the other hand, the number of correct content questions was statistically significantly lower in C_2 than in C_1 which provokes the conclusion that the result of the work is less accurate in high paid campaigns. However, there may be different reasons like language problems to correctly understand the content questions. In C_2, a larger ratio of Asian users from low-wage countries was observed than in C_1, which may be attracted by the higher payments. However, this change in demographics of the subjects may be simply caused by the time when the campaign was launched due to day-night activities of users and time shift in countries [14].

Next, the consistency of voting in the campaigns is compared. To this end, we define a consistency value which is the difference $Z_{i,j}$ of the QoE assessment by an individual subject j for the same test condition in campaign i. The probability that the consistency value is larger than 2 is $P(Z_1 \geq 2) = 3.14\%$ and $P(Z_2 \geq 2) = 1.85\%$, respectively. For both campaigns, the mean consistency values per campaign averaged over all users are low and in the same order. In particular, the corresponding 95 % confidence intervals of the mean values are $[0.0184; 0.0498]$ and $[0.0099; 0.0314]$ and therefore overlapping. Thus, there is no statistically significant difference in terms of consistency for the two campaigns and thus no impact of the payments on the consistency of user ratings.

In the following, a worker j and his ratings in campaign i are defined to be reliable, if the consistency value is less than 2,

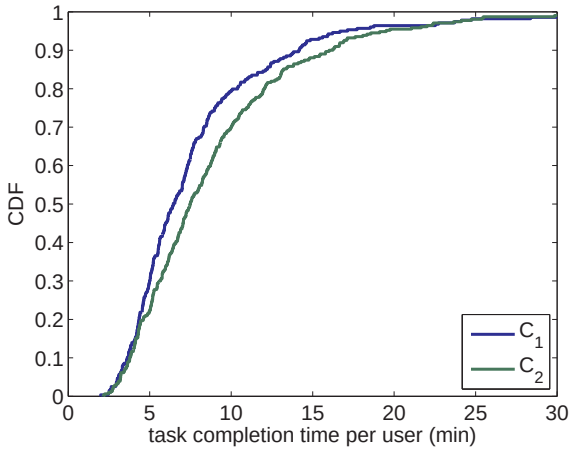


Figure 3: Task completion times of reliable users in campaigns C_1 and C_2 with payments P_1 and $P_2 = 3 \cdot P_1$.

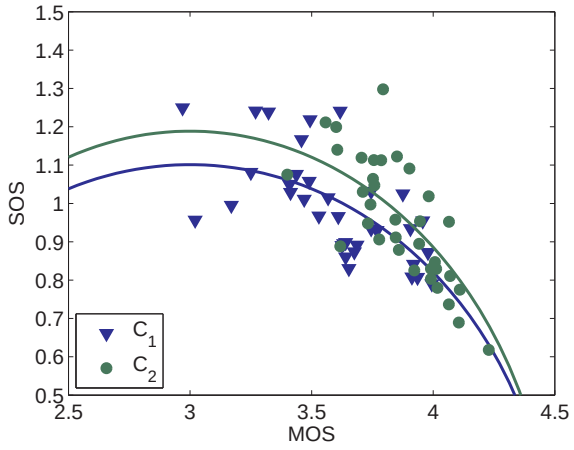


Figure 4: For each test condition, the standard deviation of opinion scores vs. MOS values over subjects is plotted for campaigns C_1 and C_2 with payments P_1 and $P_2 = 3 \cdot P_1$. The solid lines show a square fitting function according to SOS hypothesis [17].

i.e. $Z_{i,j} \leq 1$, *and* the number $K_{i,j}$ of correct content questions is larger than 7, i.e. $K_{i,j} \geq 7$. Table 2 shows that the reliability is lower in C_2 which is mainly caused by $K_{i,j}$. In particular, the Spearman rank correlation coefficient indicates no correlation between reliability and consistency (-0.0306), but a strong correlation between reliability and content questions (0.6955). In summary, higher payments do not increase reliability of workers and therefore data quality, i.e., *"you don't get what you pay for"* (which is in line with previously cited work).

3.3. User Ratings and Impact on QoE

For our further analysis of user ratings, we will only consider reliable users according to the definition above. By means of ANOVA, we identified the same key influence factors in both campaigns which lead to similar p-values. In particular, the origin country, the type of website, the color goodness C_g and the font goodness F_g were identified as key influence factors determining visual appeal (VA) QoE assessment. Other factors like the age of the subjects, the number of fonts or colors, etc. turned out not to be relevant for the VA assessment.

The impact of those parameters on the mean opinion score (MOS) is quantified as main effect plot in Figure 5. In particular, the MOS value and the corresponding 95 % confidence interval are plotted for both campaigns C_1 and C_2 depending on the identified main influence factor. First, the payments reflected by the different campaigns clearly lead to different MOS value and non-overlapping confidence intervals for those parameters (and also for non-major influence factors like age of subjects). ANOVA clearly shows that the amount of payments is also a main influence factor on VA QoE with a p-value below $1e-10$. Second, the MOS values in the higher paid campaign C_2 are significantly larger than in C_1. There are two possible explanations for this. Users in the second campaign have a different understanding of VA QoE or the meaning of rating scales (e.g. due to their origin country and language [15]). Since the origin country is an identified key influence factor on VA QoE [16] and the demographics are different in both campaigns, different absolute MOS values are the consequence.

Another possible explanation may be that the users in the second campaign wanted to 'satisfy' the employer thus improving the chances of getting their work approved. This may be caused for instance if the workers are not sure what is the main purpose of the study, which in turn may depend on some demographic aspects (familiarity with English language, etc.). Figure 4 shows the standard deviation of opinion scores (SOS) res vs. mean opinion scores (MOS). For each test condition, the SOS and the MOS value is computed over all subjects and plotted for campaigns C_1 and C_2 with a marker. The solid lines show a square fitting function according to SOS hypothesis [17]. It can be seen that the users in C_2 show higher SOS values than the users in C_1 for similar MOS ratings.

Nevertheless, since ANOVA identified the same key influence factors in both campaigns, but the absolute MOS values differ depending on the payments, the consequence is to use normalized user ratings for QoE analysis and modeling. Figure 6 shows the main effect plots where the user ratings are normalized by the average user rating over all subjects and test conditions. As the average user rating is higher in C_2 than in C_1, this normalization will lead to the same normalized average user ratings. The results show that there is for example a significant impact of the country and the age that leads to non-monotonic results. Both main effects are difficult to consider in a model because of the non-monotonic behavior and the quantification or grouping of countries. Intuitively, the visual appeal model should be independent of user demographics. Therefore, we try a different normalization scheme which normalizes the ratings per user in order to avoid those effects.

Figure 7 shows now the main effect plots for normalized user ratings. In particular, we used standard scores (or Z-scores) per user. For a user with ratings y_j for the different test conditions $j = 1, \cdots, M$, the Z-score is defined as $z_j = \frac{y_j - E[y]}{STD[y]}$ with the mean value $E[y]$ and the standard deviation $STD[y]$ for this user over the test conditions. This kind of normalization is used to avoid rating scale effects, and indeed it can be seen that it overcomes them, as well as linear shifts due e.g. to the incentives or demographics. Figure 7 shows that the normalized user ratings are very similar in both campaigns and that the same key influence factors are still observed (website, color goodness, font goodness). We see further that the origin country is not a major influence on the Z-scores. Hence, VA QoE

is affected by those three factors, while factors like country or payments simply shift the upper bound of VA QoE.

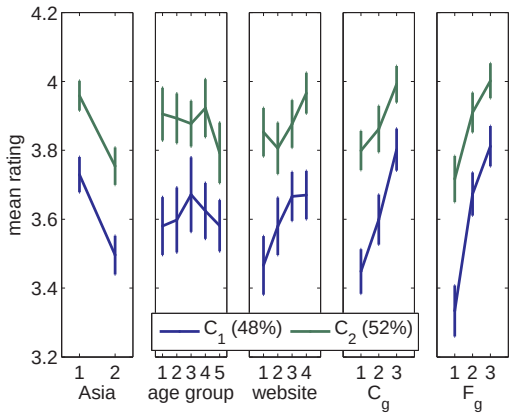


Figure 5: Impact of payments P_1 and $P_2 = 3 \cdot P_1$ in campaigns C_1 and C_2 on user ratings.

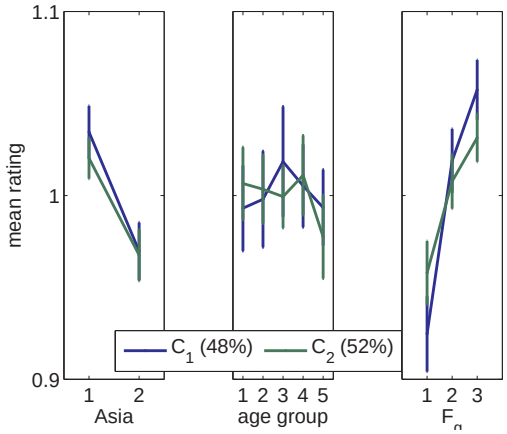


Figure 6: User ratings are normalized by overall average user rating in campaigns C_1 and C_2.

4. SWOT Analysis of QoE Crowdtesting

A SWOT (Strengths, Weaknesses, Opportunities, Threats) analysis of crowdsourced Web QoE experiments can be found in Figure 8. The main strengths of the approach are mostly related to its low cost and possibilities of conducting very large-scale tests in a short time, with a varied userbase. Concerning weaknesses, there are several. Laboratory-based protocols cannot be used in this context for several reasons, including instrumentation details, test duration, lack of moderator, and test subjects who are, by and large, prone to cheating. These, in turn, requires careful instrumentation of the test campaign, and statistical filtering of the results.

Compared to lab-based tests, which tend to have very localized and homogeneous subject populations, crowdsourcing opens new opportunities with regards to understanding the impact of demographic and other contextual factors, for example. As observed in the campaigns described in this work, these factors can have an important impact on ratings. Concerning

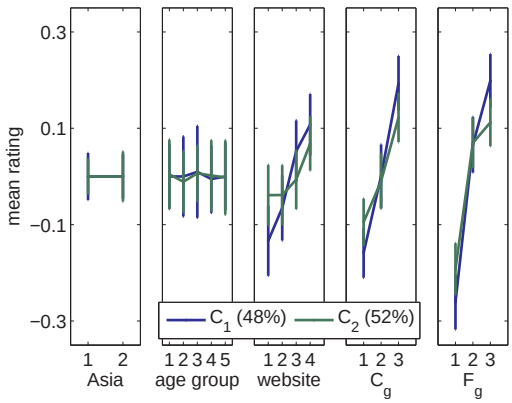


Figure 7: Z-scores of user ratings are considered to quantify the impact of payments P_1 and $P_2 = 3 \cdot P_1$ in campaigns C_1 and C_2 on VA assessment.

threats, crowdsourcing introduces some factors (e.g. display type, internet connection speed) which cannot be always controlled or accounted for. Careful instrumentation of the tests allows to mitigate (and in some cases possibly eliminate) some of these factors, but it is not possible to have a completely controlled test setup.

5. Conclusions and Outlook

The use of a crowdsourcing-based approach for testing the visual appeal of websites resulted in a number of interesting lessons learned. Firstly, it became clear that crowdsourcing does provide a valuable mechanism for quickly and cheaply conducting these types of experiments while still obtaining meaningful results. In that sense, the results obtained are encouraging. On the other hand, a number of issues were also noticed. Firstly, and contrary to possible expectations, an increase in payments will not necessarily lead to better results. In fact, in our results it led to an increase in the number of unreliable users, most likely due to increased financial incentive to participate. Taking this into account, it is clear that additional incentives (e.g., gamification) and careful statistical analysis are required to avoid poor quality results.

Another apparent impact of the increased payments was the much faster completion of the test campaign. While this is in some cases desirable, it also results in a narrower variety of users in terms of demographics (due, for example, to the influence of time-zones). It might be worth taking this into account when proposing the campaigns, and possibly throttling their execution in order to obtain more representative population samples. The effects of time-zone differences also affects the reproducibility of the results, as it is hard, if not impossible to obtain similar demographics distributions in different test runs.

In terms of the actual scores we notice that while payment level influences absolute MOS values for given assessment tests, it does not influence qualitative relations (i.e., main effects, interactions, shape of curves). Thus there does not appear to be a severe impact on models built from the campaign data (if such models exist). However, user ratings may have to be normalized to cope with the payment effect and to merge data from different studies (with different payments).

We note that at this point we are not able to really make gen-

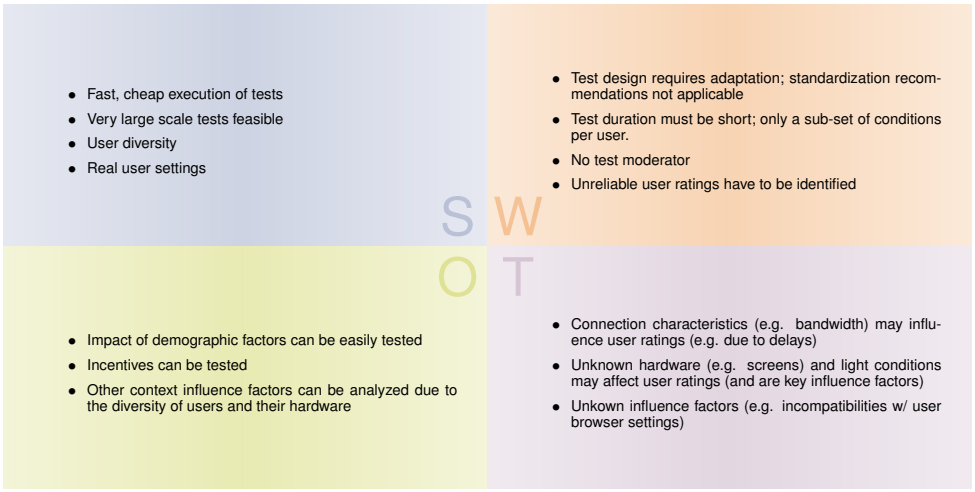


Figure 8: SWOT analysis of crowdsourcing for Web QoE

eralized claims based on these results, as they strongly depend on the actual crowdsourcing platform. However, it seems clear that the payment aspects need to be considered when setting up crowdsourcing campaigns in the context of QoE assessment and modeling. It would be also interesting to investigate the effects of decreasing the payment levels, both with regards to the campaign completion times, and results' quality, as well as comparing them to other ways to do crowdsourcing "for free" (e.g. with student groups, or via social platforms such as Facebook).

6. Acknowledgements

This work was partly funded by the COST Action IC1003 — QUALINET. The authors alone are responsible for the content. M. Varela and T. Mäki's work was partially funded by Tekes the Finnish agency for research innovation, in the context of the CELTIC+ project QuEEN. L. Skorin-Kapov's work was supported by the Ministry of Science, Education and Sports of the Republic of Croatia projects no. 036-0362027-1639 and 071-0362027-2329.

7. References

[1] C. Keimel, J. Habigt, C. Horch, and K. Diepold, "QualityCrowd — a Framework for Crowd-Based Quality Evaluation," in *Picture Coding Symposium (PCS), 2012*, 2012, pp. 245–248.

[2] M. Varela, T. Mäki, L. Skorin-Kapov, and T. Hoßfeld, "Towards an Understanding of Visual Appeal in Website Design," in *Proceedings of QoMEX 2013*, Klagenfurt, Austria, Jul. 2013.

[3] M. Buhrmester, T. Kwang, and S. D. Gosling, "Amazon's mechanical turk a new source of inexpensive, yet high-quality, data?" *Perspectives on Psychological Science*, vol. 6, no. 1, pp. 3–5, 2011.

[4] W. Mason and D. J. Watts, "Financial incentives and the performance of crowds," in *Proceedings of the ACM SIGKDD workshop on human computation*. ACM, 2009, pp. 77–85.

[5] M. Marge, S. Banerjee, and A. I. Rudnicky, "Using the amazon mechanical turk for transcription of spoken language," in *Acoustics Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 5270–5273.

[6] A. Aker, M. El-Haj, M.-D. Albakour, and U. Kruschwitz, "Assessing crowdsourcing quality through objective tasks," in *Proceedings of LREC*, vol. 12, 2012.

[7] C. Harris, "You're hired! an examination of crowdsourcing incentive models in human resource tasks," in *WSDM Workshop on Crowdsourcing for Search and Data Mining (CSDM)*, 2011, pp. 15–18.

[8] J. J. Horton and L. B. Chilton, "The labor economics of paid crowdsourcing," in *Proceedings of the 11th ACM conference on Electronic commerce*. ACM, 2010, pp. 209–218.

[9] N. Kaufmann, T. Schulze, and D. Veit, "More than fun and money. worker motivation in crowdsourcing–a study on mechanical turk," in *Proceedings of the Seventeenth Americas Conference on Information Systems*, 2011, pp. 1–11.

[10] R. Bringhurst, *The Elements of Typographic Style*, 2nd ed. Hartley & Marks Publishers, 2002.

[11] L. Eisemann, *Pantone's Guide to Communicating with Color*. HOW Books, 2000.

[12] S. Egger and R. Schatz, "Interactive Content for Subjective Studies on Web Browsing QoE: A Kepler Derivative," in *ETSI STQ Workshop on Selected Items on Telecommunication Quality Matters*, Vienna, Nov. 2012.

[13] T. Hoßfeld, C. Keimel, M. Hirth, B. Gardlo, J. Habigt, K. Diepold, and P. Tran-Gia, "CrowdTesting: A Novel Methodology for Subjective User Studies and QoE Evaluation," University of Würzburg, Tech. Rep. 486, Feb. 2013.

[14] M. Hirth, T. Hoßfeld, and P. Tran-Gia, "Anatomy of a Crowdsourcing Platform - Using the Example of Microworkers.com," in *Workshop on Future Internet and Next Generation Networks (FINGNet)*, Seoul, Korea, Jun. 2011.

[15] Z. Cai, N. Kitawaki, T. Yamada, and S. Makino, "Comparison of MOS evaluation characteristics for chinese, japanese, and english in IP telephony," in *Universal Communication Symposium (IUCS), 2010 4th International*, 2010, pp. 112–115.

[16] D. Cyr, M. Head, and H. Larios, "Colour appeal in website design within and across cultures: A multi-method evaluation," *International Journal of Human Computer Studies*, vol. 68, pp. 1–21, 2010.

[17] T. Hoßfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!" in *QoMEX 2011*, Mechelen, Belgium, Sep. 2011.