Evaluating a Web-Based Tool for Crowdsourced Navigation Stress Tests

Florian Meier[1], Alexander Bazo[2], Manuel Burghardt[2], and Christian Wolff[2]

[1] Information Science Group, University of Regensburg, Regensburg, Germany
[2] Media Informatics Group, University of Regensburg, Regensburg, Germany
{florian.meier,alexander.bazo,manuel.burghardt,
christian.wolff}@ur.de

Abstract. We present a web-based tool for evaluating the information architecture of a website. The tool allows the use of crowdsourcing platforms like *Amazon's MTurk* as a means for recruiting test persons, and to conduct asynchronous remote navigation stress tests (cf. Instone 2000). We also report on an evaluation study which compares our tool-based crowdsourced approach to a more traditional laboratory test setting. Results of this comparison indicate that although there are interesting differences between the two testing approaches, both lead to similar test results.

Keywords: remote usability testing, crowdsourcing, MTurk, information architecture, navigation stress test.

1 Introduction: Web-Based Usability Testing

Extensive usability evaluations conducted in a laboratory setting are very cost-intensive and time-consuming (Nielsen 2009). As a consequence, web-based asynchronous usability testing is becoming increasingly popular (Sauro 2011), thus fulfilling the prediction of Hartson et al. (1996), who described "the network as an extension of the usability laboratory". Throughout the literature we find many examples for research on the efficiency and effectiveness of web-based usability tests in comparison to tests conducted in a laboratory setting. There are three main threads of research that are relevant for our study:

Asynchronous Usability Testing

Bruun et al. (2009) conducted an extensive review of literature to identify several papers that compare asynchronous usability testing methods to laboratory-based approaches. Among the asynchronous methods that are described in the literature are three main classes of tests: (1) *reporting methods*, like for instance the *critical incident method* (Castillo, Hartson & Hix 1998) or the *diary-based user reporting* (Thompson 1999), (2) *web analytics methods* that make use of logfiles and other quantitative user behavior data, and (3) *tool-based approaches*, which allow a specific

A. Marcus (Ed.): DUXU/HCII 2013, Part IV, LNCS 8015, pp. 248–256, 2013.
© Springer-Verlag Berlin Heidelberg 2013

and customized evaluation of predefined user tasks such as finding/clicking a certain link, or creating/updating/deleting a user account (Bolt & Tulathimutte 2010).

Analog vs. Digital Usability Testing

A second thread of related research is dedicated to the comparison of analog versus digital versions of usability testing methods, like for instance *card sorting*. In the case of card sorting, no significant differences have been found between the results of analog, supervised card sorting tests and results of digital card sorting tests, which can be conducted with web-based tools such as *Netsorting* (Bussolon, Del Missier & Russi 2006).

Crowd-Sourced Usability Testing

The area of research most relevant for our work is concerned with usability tests of websites which make use of crowdsourcing platforms like *Amazon Mechanical Turk (MTurk)* (cf. Amazon Mechanical Turk, 2009). Although there are many services[1] that offer and support crowdsourced usability testing, only little effort has been dedicated to the evaluation of the efficiency of crowdsourcing platforms as a recruiting strategy for participants of user studies, not to mention the obvious combination with asynchronous remote usability tools. Among the scarce research in this area is a study that investigates the ability of experts and crowdsourced workers to assess the quality of Wikipedia articles (Kittur, Chi & Suh 2008). Results show that crowdsourced workers in general do worse than experts, as many of them do not give serious and reliable judgments, but rather tend to *gaming* and *spamming*. However, the study also revealed that the use of control mechanisms (for instance CAPTCHA-like questions) and an adaption of the test questions as well as individual task design can raise the success rate of the crowdsourcing group significantly. Franco et al. (2010) compare a traditional lab usability test with user studies conducted on *MTurk* by evaluating the website *workintexas.com*: The results of both approaches reveal a broad consensus on identifying the most severe usability problems of the site. Moreover, the MTurkers made extensive use of the comment function, thus providing useful feedback on the site's usability issues. In a more recent study Liu et al. (2012) compare traditional lab usability tests with crowdsourced usability tests. They observed several differences concerning the number of participants, the demographics, the time spent on tests and the actual cost, but also found that the number of identified usability problems was quite the same. The authors also point out advantages and disadvantages of laboratory-based and crowdsourcing tests (cf. Table 1), suggesting a cyclic combination of both approaches throughout the evaluation process.

[1] `http://www.trymyui.com/` | `www.easyusability.com/`

Table 1. Advantages and disadvantages of crowdsourced usability tests over lab usability tests according to Liu et al. (2012)

Advantages	Disadvantages
More Participants	Lower Quality Feedback
High Speed	Less Interaction
Low Cost	Spammers
Various Backgrounds	Less Focused User Groups

2 Research Agenda

The study presented in this article is following up the different threads of research described in the previous section: It describes a special case of asynchronous, remote usability testing, which is realized via a web-based tool that makes use of the advantages of crowdsourced recruiting platforms.

The testing method that is the subject of our study is Instone's (2000) *navigation stress test* (NST), which focuses on the evaluation of a website's *information architecture* (Toub 2000). Information architecture is "the art and science of shaping information products and experiences to support usability and findability" (Morville & Rosenfeld, 2006, p. 4), which means it is a specific aspect of usability as a whole. However, information architecture is difficult to test (and measure), as it lies beneath the surface of a website's visual and technical design, and thus requires specifically designed testing methods such as the NST. As the NST was originally designed for use in a paper and pencil setting, we evaluate its efficiency and effectiveness for both, a traditional, synchronous laboratory context, as well as an asynchronous, crowdsourced web context.

On our research agenda for this paper are the following main objectives:

1. Develop a web-based tool prototype that can be integrated with an existing crowd-sourcing platform to enable asynchronous remote-usability testing of the information architecture of a specific website.
2. Design an evaluation study that allows to compare our crowdsourced remote approach to a laboratory usability test setting
3. Enhance the functionality of the tool prototype in a way the tool can be used for testing generic websites.

3 Study Design

To conduct a NST, a random page of a complex website is selected by the experimenter. We chose a subpage of the *Media Informatics Group* website at the University of Regensburg as the subject of investigation. The website has a rather traditional layout (cf. Figure 1) with a local navigation for each page on the left side, and a global navigation in the footer area that is visible at every page. The global navigation in the footer area can be dynamically hidden or displayed as required. The respective page is printed in black and white, and the participants of the evaluation study are asked to answer basic questions (cf. Table 2) concerning the site navigation by marking up the relevant elements on the printed page. Table 2 shows the adapted questions (cf. Instone 2000) for our study.

Table 2. Overview of the eight navigation questions (cf. Instone 2000) and respective markup used in the NST-study of a university web page

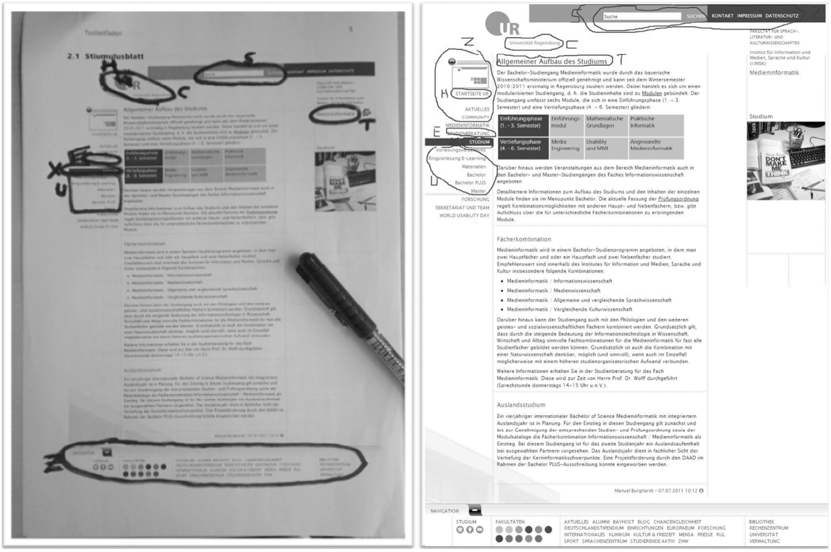
	Navigation question	**Recommended mark up on the paper**
1	What website is this?	Circle the website name and mark it with the letter '**C**'
2	What is the title of this very page?	Circle the page title and mark it with the letter '**T**'
3	Where is the search function?	Circle the search function and mark it with the letter '**S**'
4	How do you get to the home page of this website?	Circle the link and mark it with the letter '**H**'
5	Which link corresponds to this very page?	Circle the subpage link and mark it with the letter '**X**'
6	Which link gets you one level up in the site hierarchy (i.e. to the parent page)?	Circle the link and mark it with the letter '**E**'
7	Which group(s) of links get(s) you one level down in the site hierarchy (i.e. to further sub pages)?	Circle the group(s) of links and mark it/them with the letter '**U**'
8	Which group(s) of links do you think is/are on every page of this web site (i.e. they belong to the main navigation)?	Circle the group(s) of links and mark it/them with the letter '**Z**'

We have developed a web-based tool[2] that allows to conduct asynchronous remote NST by annotating the image of a static website within the web browser, making use of the HTML5 canvas element (W3C 2012). The annotated web pages can be interpreted asynchronously by the creator of the remote evaluation. Figure 1 shows that the web-based NST adaption looks very similar to the pen and paper version. The test candidates for the actual evaluation were recruited via *MTurk*, a crowdsourcing platform that is designed to create and publish small manageable tasks which are known as *micro-tasks*[3] in common crowdsourcing terminology. Due to fiscal issues that do not allow non-US-based requesters to post tasks directly on MTurk, the intermediary

[2] The tool is available at `http://pc57724.uni-regensburg.de/~flo/stresstest.html`. The tested web page and tasks were originally formulated in German.

[3] *Amazon* has a special name for micro-tasks on *MTurk*: they are called *Human Intelligence Tasks* (HITs), emphasizing the fact that the tasks are meant to be solved by the crowd, i.e. by human workers.

platform *Crowdflower* was used for the distribution of the test on MTurk. In our case the micro-tasks consisted of the actual NST questions (cf. Table 2) as well as questions about the subjective assessment of the task difficulty. We also asked the evaluators to comment on the information architecture of the tested page in general.



Suche SUCHEN KONTAKT IMPRESSUM DATENSCHUTZ

UR Universität Regensburg

FAKULTÄT FÜR SPRACH-, LITERATUR- UND KULTURWISSENSCHAFTEN

Institut für Information und Medien, Sprache und Kultur (I:IMSK)

Medieninformatik

STARTSEITE UR
AKTUELLES
COMMUNITY
MEDIENINFORMATIK
STUDIENBERATUNG
STUDIUM
Vorlesungsverzeichnis
Ringvorlesung E-Learning
Materialien
Bachelor
Bachelor PLUS
Master
FORSCHUNG
SEKRETARIAT UND TEAM
WORLD USABILITY DAY

Allgemeiner Aufbau des Studiums

Der Bachelor-Studiengang Medieninformatik wurde durch das bayerische Wissenschaftsministerium offiziell genehmigt und kann seit dem Wintersemester 2010/2011 erstmalig in Regensburg studiert werden. Dabei handelt es sich um einen modularisierten Studiengang, d. h. die Studieninhalte sind zu Modulen gebündelt. Der Studiengang umfasst sechs Module, die sich in eine Einführungsphase (1. – 3. Semester) und eine Vertiefungsphase (4. – 6. Semester) gliedern:

Einführungsphase (1. - 3. Semester)	Einführungs-modul	Mathematische Grundlagen	Praktische Informatik
Vertiefungsphase (4. - 6. Semester)	Media Engineering	Usability und MMI	Angewandte Medieninformatik

Darüber hinaus werden Veranstaltungen aus dem Bereich Medieninformatik auch in den Bachelor- und Master-Studiengängen des Faches Informationswissenschaft angeboten

Detailliertere Informationen zum Aufbau des Studiums und den Inhalten der einzelnen Module finden sie im Menüpunkt Bachelor. Die aktuelle Fassung der Prüfungsordnung regelt Kombinationsmöglichkeiten mit anderen Haupt- und Nebenfächern, bzw. gibt Aufschluss über die für unterschiedliche Fächerkombinationen zu erbringenden Module.

Fächerkombination

Medieninformatik wird in einem Bachelor-Studienprogramm angeboten, in dem man zwei Hauptfächer und oder ein Hauptfach und zwei Nebenfächer studiert. Empfehlenswert sind innerhalb des Institutes für Information und Medien, Sprache und Kultur insbesondere folgende Kombinationen:

- Medieninformatik / Informationswissenschaft
- Medieninformatik / Medienwissenschaft
- Medieninformatik / Allgemeine und vergleichende Sprachwissenschaft
- Medieninformatik / Vergleichende Kulturwissenschaft

Darüber hinaus kann der Studiengang auch mit den Philologien und den weiteren geistes- und sozialwissenschaftlichen Fächern kombiniert werden. Grundsätzlich gilt, dass durch die steigende Bedeutung der Informationstechnologie in Wissenschaft, Wirtschaft und Alltag sinnvolle Fachkombinationen für die Medieninformatik für fast alle Studienfächer gebildet werden können. Grundsätzlich ist auch die Kombination mit einer Naturwissenschaft denkbar, möglich (und sinnvoll), wenn auch im Einzelfall möglicherweise mit einem höheren studienorganisatorischen Aufwand verbunden.

Weitere Informationen erhalten Sie in der Studienberatung für das Fach Medieninformatik. Diese wird zur Zeit von Herrn Prof. Dr. Wolff durchgeführt (Sprechstunde donnerstags 14–15 Uhr u.n.V.).

Auslandsstudium

Ein vierjähriger internationaler Bachelor of Science Medieninformatik mit integriertem Auslandsjahr ist in Planung. Für den Einstieg in diesen Studiengang gilt zunächst und bis zur Genehmigung der entsprechenden Studien- und Prüfungsordnung sowie der Modulkataloge die Fächerkombination Informationswissenschaft / Medieninformatik als Einstieg. Bei diesem Studiengang ist für das zweite Studienjahr ein Auslandsaufenthalt bei ausgewählten Partnern vorgesehen. Das Auslandsjahr dient in fachlicher Sicht der Vertiefung der Kerninformatikschwerpunkte. Eine Projektförderung durch den DAAD im Rahmen der Bachelor PLUS-Ausschreibung konnte eingeworben werden.

Manuel Burghardt – 07.07.2011 10:12

Studium

NAVIGATION

STUDIUM

FAKULTÄTEN

AKTUELLES ALUMNI BAYHOST BLOG CHANCENGLEICHHEIT DEUTSCHLANDSTIPENDIUM EINRICHTUNGEN EUROPAEUM FORSCHUNG INTERNATIONALES KLINIKUM KULTUR & FREIZEIT MENSA PRESSE RUL SPORT SPRACHENZENTRUM STUDIERENDE AKTIV ZHW

BIBLIOTHEK
RECHENZENTRUM
UNIVERSITÄT
VERWALTUNG

Fig. 1. The left side shows a photograph of the paper NST, the right side shows a screenshot of our annotated web version

In order to compare the crowdsourced approach to the traditional pen and paper approach, we conducted two NSTs for the same web page (cf. Figure 1). The results of the evaluation of both variants are presented in the next section.

4 Discussion of Results

In order to compare both NST-variants to each other, we recorded if a task was successfully achieved (*task success*) and how long it took the evaluator to achieve the task (*time on task*). Figure 2 shows part of the sample solution for successful tasks as well as a heat map-like visualization of the aggregated crowdworker annotations.

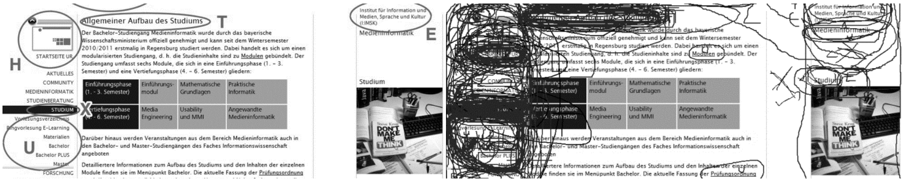


Fig. 2. Sample solution for successful tasks in the NST (left) and heat map of the aggregated mouse interactions from all crowdworkers (right)

Although the number of participants was – expectedly – quite different in the two NST tests (crowdsourcing n=28, laboratory n=10), the two groups of participants were very similar with regard to demographic aspects. We found that the total task success (including all eight subtasks as defined in the NST) was almost identical in the traditional (70%) and in the crowdsourced test setting (69%). Also, the total time to achieve all tasks did not differ significantly (p=0,296). The average time to achieve all tasks in both NST-settings took around 4-5 minutes.

Table 3. Average time on task, standard deviation and the result of a t-test for the two test conditions

	time on task (average)	**SD**	**t-test**
crowdsourced NST	4min 12sec	120 sec	p=0,296
laboratory NST	4min 44sec	44 sec	

Figure 3 gives a detailed overview of the specific success rates for each task in the respective test scenarios. Strikingly, task 1 (*mark page name*), task 3 (*mark search functionality*) and task 4 (*mark link to home page*) have a 100% success rate in the crowdsourcing scenario. This shows that annotating areas of a webpage by using the mouse does not seem to be a problem for the evaluators. Wrong or missing annotations occur for both test scenarios, but seem to be connected to deficits of the site's information architecture rather than to the tool's usability (cf. task 6). Furthermore, a chi-square test showed that there are no significant differences in *success per task.*

Further Findings

- Although we were afraid of a high percentage of spammers in the crowdsourced variant, our test showed that the quality assurance mechanisms of *MTurk* worked pretty well: Actually, 75% of all applicants could be used for the test.
- Despite the fact that 25% of the *MTurk* evaluations had to be excluded due to quality aspects, crowdsourcing platforms as a means of recruiting test participants are still cheaper and faster than recruiting test persons for laboratory scenarios.
- The average time needed for completing the tasks was slightly higher for the crowdsourced variant, which seems plausible, as evaluators do the tests in private and have no one who is watching them.
- *MTurk* evaluators generally assess the information architecture more positively than the analogous evaluators.
- *MTurk* evaluators don't give many additional comments (in written form) while analogous evaluators do comment their annotations a lot (orally). It must be noted, though, that the verbal comments were mainly connected to general usability issues rather than to additional aspects of information architecture.

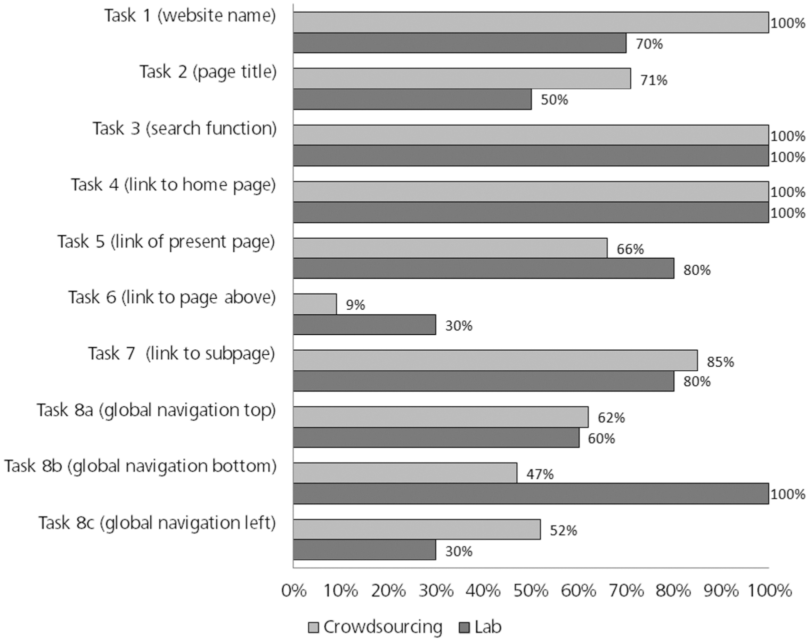


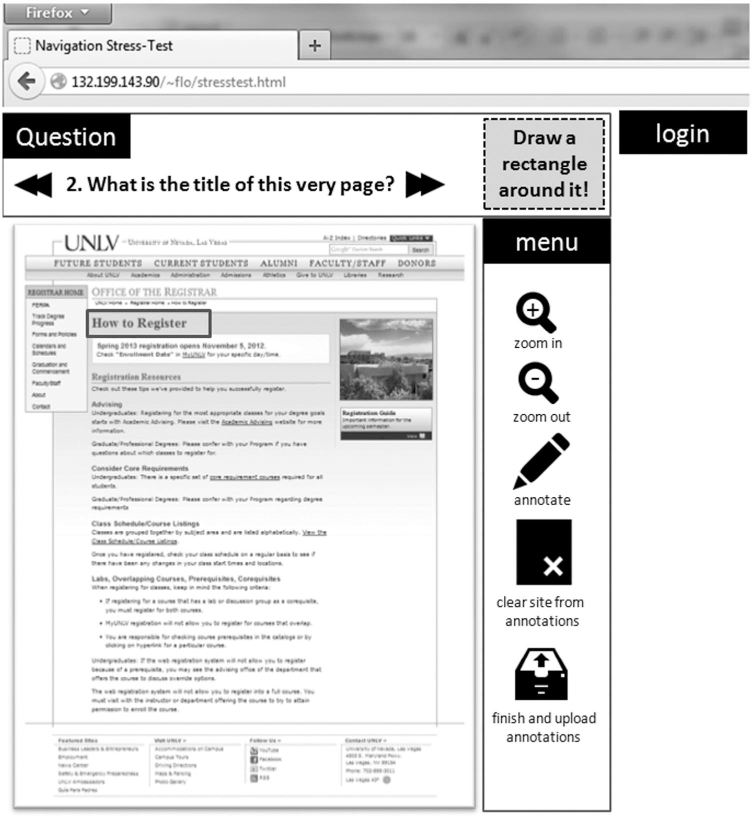
Fig. 3. Overview of the total task success per task for crowdsourced and lab NST

The results show that the navigation stress test can be implemented as a digital version, making use of crowdworkers as test participants. The benefits are reduced costs (no laboratory and equipment required, cheap crowdworkers) and increased flexibility for the experimenter, due to unsupervised, asynchronous test sessions.

5 Outlook

As the quality and efficiency of a tool-based digital NST was be evaluated positively, we are planning to evolve the tool, which at this stage has the character of a prototype, into a configurable testing tool that maybe be used by others, too. Experimenters will be able to upload an individual screenshot of a website that is to be tested as well as a set of individual questions. We suggest to stick to the basic questions as defined by Instone 2000 if possible, and modify them only if necessary.

In the new NST-tool (for a first draft of the interface cf. Figure 4), we will display the questions in a sequential order, while in the original NST-version all questions are displayed at once. This change will make annotating much simpler and clearer, as there will not be multiple overlapping annotations on one single canvas, but rather one annotation per task/question on a separate canvas. As a side effect, the evaluators need not index their annotations with capital letters, as the annotations are already explicitly related to the different tasks. Also, this allows to implement automatic, task-specific analysis features such as time per task. We will also test a rectangular selection tool, which renders freehand annotations unnecessary.


Firefox
Navigation Stress-Test
132.199.143.90/~flo/stresstest.html
Question
login
menu
zoom in
zoom out
annotate
clear site from annotations
finish and upload annotations
UNLV
How to Register

2. What is the title of this very page?

Draw a rectangle around it!

Fig. 4. Preview of the new NST-tool

The new NST-tool will be available at http://www.crowdsourcing-tools.com/nst once it has successfully passed a first evaluation round and reached a decent level of maturity.

References

1. Amazon Mechanical Turk: Requester User Interface Guide (2008), `http://s3.amazonaws.com/awsdocs/MechTurk/latest/amt-ui.pdf` (last accessed on February 28, 2013)
2. Bolt, N., Tulathimutte, T.: Remote Research. Rosenfeld Media, New York (2010)
3. Bussolon, S., Del Missier, F., Russi, B.: Online card sorting: as good as the paper version. In: Proceedings of the 13th European Conference on Cognitive Ergonomics (ECCE), Zürich (2006), `http://www.oat.ethz.ch/news/presentations/bussolon` (last accessed on February 28, 2013)

4. Bruun, A., Gull, P., Hofmeister, L., Stage, J.: Let your users do the testing: a comparison of three remote asynchronous usability testing methods. In: Proceedings of the 27th International Conference on Human Factors in Computing Systems (CHI 2009), pp. 1619–1628. ACM, New York (2009), http://doi.acm.org/10.1145/1518701.1518948 (last accessed on February 28, 2013)
5. Castillo, J.C., Hartson, H.R., Hix, D.: Remote usability evaluation: Can users report their own critical incidents? In: Proceedings of CHI 1998, pp. 253–254. ACM, New York (1998)
6. Liu, D., Lease, M., Kuipers, R., Bias, R.: Crowdsourcing Usability Testing. In: Computing Research Repository, CoRR (2007), http://arxiv.org/abs/1203.1468
7. Franco, S., Herbstritt, S., Johnson, E., Schumacher, S., Van Zandt, L.: Assessing Users' Needs for Usability on the Job Search Features of WorkInTexas.com (2010), http://susieherbstritt.com/downloads/witcomux.pdf (last accessed on February 28, 2013)
8. Hartson, H.R., Castillo, J.C., Kelso, J., Neale, W.C.: Remote evaluation: the network as an extension of the usability laboratory. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems: Common Ground (CHI 1996), pp. 228–235. ACM, New York (1996)
9. Instone, K.: Navigations Stress Test (2000), http://instone.org/navstress (last accessed on February 28, 2013)
10. Kittur, A., Chi, E.H., Suh, B.: Crowdsourcing user studies with Mechanical Turk. In: Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems (CHI 2008), pp. 453–456. ACM, New York (2008)
11. Morville, P., Rosenfeld, L.: Information Architecture for the World Wide Web. Designing Large-Scale Web Sites. O'Reilly, Sebastopol (2006)
12. Nielsen, J.: Discount Usability: 20 Years. Alertbox, September 14 (2009), http://www.useit.com/alertbox/discount-usability.html (last accessed on February 28, 2013)
13. Sauro, J.: The Methods UX Professionals Use (2011), http://www.measuringusability.com/blog/ux-methods.php (last accessed on February 28, 2013)
14. Toub, S.: Evaluating Information Architecture. A Practical Guide to Assessing Web Site Organization. Argus Center for Information Architecture Whitepaper (2000), http://argus-acia.com/white_papers/evaluating_ia.pdf (last accessed on February 28, 2013)
15. Thompson, J.: Investigating the Effectiveness of Applying the Critical Incident Technique to Remote Usability Evaluation. Master thesis, Virginia Polytechnic Institute and State University (1999), http://scholar.lib.vt.edu/theses/available/etd-121699-205449/unrestricted/thesis.pdf (last accessed on February 28, 2013)
16. W3C. W3C Working Draft: HTML5. A vocabulary and associated APIs for HTML and XHTML (2012), http://www.w3.org/TR/html5/the-canvas-element.html#the-canvas-element (last accessed on February 28, 2013)