A Grounded Procedure for Managing Data and Sample Size of a Home Medical Device Assessment

Simone Borsci[1], Jennifer L. Martin[2], and Julie Barnett[1]

[1] Brunel University, School of Information Systems, Computing and Mathematics, Kingston Lane, Uxbridge, Middlesex UB8 3PH, UK
[2] University of Nottingham, Department of Electrical and Electronic Engineering, University Park, Nottingham NG7 2RD, UK
Simone.borsci@brunel.ac.uk

Abstract. The selection of participants for usability assessment, together with the minimum number of subjects required to obtain a set of reliable data, is a hot topic in Human Computer Interaction (HCI). Albeit, prominent contributions through the application of different *p* estimation models argued that five users provide a good benchmark when seeking to discover interaction problems a lot of studies have complained this five-user assumption. The sample size topic is today a central issue for the assessment of critical-systems, such as medical devices, because lacks in usability and, moreover, in the safety in use of these kind of products may seriously damage the final users. We argue that rely on one-size-fits-all solutions, such as the five-user assumption (for websites) or the mandated size of 15 users for major group (for medical device) lead manufactures to release unsafe product. Nevertheless, albeit there are no magic numbers for determining "a priori" the cohort size, by using a specific procedure it is possible to monitoring the sample discovery likelihood after the first five users in order to obtain reliable information about the gathered data and determine whether the problems discovered by the sample have a certain level of representativeness (i.e., reliability). We call this approach "Grounded Procedure" (GP). The goal of this study is to present the GP assumptions and steps, by exemplifying its application in the assessment of a home medical device.

Keywords: discovery likelihood, medical device, sample size, usability testing.

1 Introduction

The current trend of technology manufacturing is to propose new concepts, shapes and functioning of devices that aim to go toward an even more integrated and simplified use of the products. As Streitz [1] states, this trend produces a physical and mental disappearance of the technologies that result in what is known as ubiquitous computing. Ubiquitous computing can be considered as a new evolutionary line of the human artifact interaction in which technology is designed to be pervasive, context-aware and adaptive [2].

This ubiquitous approach is going to re-conceptualize the everyday use of the technologies not only for common interactive devices (e.g., computers, mobile phones

M. Kurosu (Ed.): Human-Computer Interaction, Part I, HCII 2013, LNCS 8004, pp. 166–175, 2013.

etc.), but also for critical-life systems, such as medical devices. In addition, as Herman and Devey [3] have noted, there is a growing trend to transform specialized devices in home care technologies. The diffusion in our everyday environments (i.e., work places, home, etc.) of these integrated technologies forces manufacturers to strongly focus their attention on the usability of the product, especially for those products, such as medical devices, that can seriously affect the user's well-being.

The usability and use-related safety of medical devices are strongly regulated [4-9] and manufacturers are required by authorities to take a user-centred design approach, where usability is integrated into the entire development cycle. The processes recommended by IEC 62366 [4] and ANSI/AAMI HE75 [9] require manufacturers to conduct multiple cycles of design and evaluation during development in order to identify and address any serious use-related risks associated with the device use. Nevertheless, as recent research suggest [10, 11], many manufacturers do not have the necessary expertise or knowledge about usability and human factors, and therefore can delay evaluation of a product until just before the release of the device in order to confirm the effectiveness of their design.

A further issue related to usability testing of medical devices is the difficulty of using the results to make appropriate design or business decisions [12]. A possible consequence of this may be that devices reach the market that pose a risk to users or patients. The safety of medical devices is an important factor , as Heneghan [13] shows, every year a large number of device release on the market have to be recalled because of safety concerns.

One of the most important concerns of manufactures when planning a usability test is deciding on which kind, and how many participants should be included. On one hand if an insufficient number of participants are included, manufacturers run the risk of not identifying all usability issues, on the other hand, however, if manufacturers conduct more testing than is necessary they waste valuable resources.

Recently the Food and Drug Administration (FDA), in response to demands from manufacturers for more clarity about the involvement of final users in usability testing, published guidance that stated that: "The most important aspect of sampling may be the extent to which the test participants correspond to the actual end users of the device" [14]. The FDA guidance suggests that while in HCI field a set of estimation models have been created, for determining the number of users needed for usability testing, these models do not reflect the real world. On the basis of this, the guidance recommends that validation testing of medical devices should include 15 users from each major user group, basing this figure on an empirical study by Faulkner [15].

Despite the guidance highlighting the limitations of the estimation models, at the end the FDA suggest that manufacturers rely on a one-sample-size-fits-all solution (i.e., to test the device with 15 users), similar to what happened in the HCI field when in the nineties it was proposed that a sample of five users could be considered sufficient for a reliable analysis of a web site (i.e., the five-user assumption) [16-20].

The aim of this paper is to critically discuss, in light of recent HCI studies, the previous one-sample-size fits all solutions and propose a new procedure for calculating usability testing sample sizes based on the data emerging from the usability assessment.

2 From One-Size Fits All Solutions to a Procedure for Managing and Checking the Sample Behavior

The reliability of the five-user assumption for the assessment of websites was shown by Nielsen [16, 18], in tune with Virzi studies [19, 20], in the nineties through the use of an estimation model, called Return on Investment (ROI). However, this model, together with the five-users assumption, was strongly criticized as too optimistic by a large set of studies [21-25]. Today in HCI, a sample of five users is only considered as a good starting point for an usability assessment, and at least others three well-tested models that addressed the optimistic results of the ROI model have been developed by researchers for estimating the number of users needed for a usability test: the Good-Turing model (GT) [23, 26], the Bootstrap Discovery Behavior model (BDB) [21] and the Monte Carlo re-sampling method [27]. All these models aim to esteem a specific index called the *p*, which represents the average percentage of errors discovered by a user. The final number of users for an evaluation sample can be calculated by inserting the *p* into the following *Error Distribution Formula* :

$$D = 1 - (1 - p)^N \tag{1}$$

When manufacturers start the evaluation neither *p* nor *D* (the total number of usability problems) is known, although clearly given one, the other can be readily calculated. This leaves those wishing to evaluate the usability of a product or service the inverse problem of whether the number of users involved in the test (N) have identified a sufficient number of problems to ensure that a given threshold percentage (i.e, D_{th}) has been met. This threshold will vary according to the type of product: for medical devices where the risks of usability errors are much greater than the website, an appropriate threshold is likely to be 97% or even higher for particularly high-risk devices [14].

The only way to check whether the evaluation of a medical device has reached the desired threshold (i.e., D_{th}) is by estimating the *p* of the total sample and then calculating how each *p* changes when a new user is added to the sample. Every time that a user is added to a sample, the overall *p* of the cohort may increase or decrease, depending on the added user's performances in terms of identifying problems. At the same time, in tune with the new user's performances the possibilities of the manufacturers to achieve a high percentage of identified problems may decrease or decrease.

By applying the estimated *p* to the *Error Distribution Formula* (1) it is possible to construct a curve of discoverability, by examining when the threshold is reached by the sample. This allows the estimation of the minimum number of participants that can represent the ability of a larger population of final users to identify all the interaction issues during an assessment.

The most recent research [21, 24-26] agrees on the fact that there are no fixed sample sizes that may guarantee beforehand the reliability of the evaluation. In fact, the variability of the users' answers and reactions during the interaction analysis is unpredictable. Moreover, all devices are different and may have different levels of complexity. In sum, the number and the kind of the problems identified by participants may vary substantially. This means that the best size of a sample is the one that can

allow practitioners to gather the larger percentage of problems (*D*), with the minimal economic investment. Finally, for medical device manufacturers to blindly rely on a mandated size is a risk as it is not possible for them to know when it is possible to stop the assessment because the desired D_{th} is reached, or when a supplementary investment in evaluation is needed because the threshold is not achieved. The only pragmatic solution to help manufacturers take appropriate decisions during usability testing is to check in an iterative way the sample behavior, as suggested in the Grounded Procedure (GP) [28] proposed by the researchers of the Match programme (founded by the EPSRC Grants: EP/F063822/1 EP/G012393/1).

In the GP manufacturers start by assuming a specific range of *p* standard (e.g., for medical 0.40-0.50 to reach the 90-97% of the problems), and use this value as a comparator against which the behavior of the real population of subjects can be assessed. In light of this, practitioners have to compare the *p* of their actual tested sample (e.g., four or five users) to the standard to make the following two main judgments, leading to the associated decisions and actions:

- *If the sample fits the standard*: report the results to the client and determine whether the product should be re-designed or released.
- *If the sample does not fit the standard*: add more users to the sample and re-test the *p* in a cyclical way until the pre-determined percentage of problems (D_{th}) is reached.

The manufactures, by applying the GP, aim to obtain reliable evidence for deciding whether to extend their evaluation by adding users or whether they can stop the evaluation because they have sufficient information. The GP consists of three main steps:

- *Monitoring the interaction problems (step 1)*: a table of problems is constructed to analyze the number of discovered problems, the number of users that have identified each problem (i.e., the weight) and the average *p* of the sample;
- *Refining the p of the cohort (step 2)*: a range of models are applied and then the number of users required reviewed in the light of the emerging *p*;
- *Taking a decision based on the sample behavior (step 3)*: the *p* is used to apply the *Error Distribution Formula* and take a decision on the basis of the available budget and evaluation aim.

Each of these steps is now discussed using an exemplar evaluation case.

3 Description of the Evaluation Case

We conducted an evaluation of a blood pressure monitor (BPM) gathered in September 2011. We tested 12 users (6 male; Age M: 29.16; SD: 1.85) each with more than one year of experience of using different kinds of BPM. We applied a think-aloud protocol where each user was asked to verbalize the problems they experienced during the use of the device. During the test session the participants completed three simple tasks: i) Preparing the monitor for use; ii) measuring blood pressure and writing down the result; iii) Switching off the monitor.

We are not interested here in describing the quality of the device, but in demonstrating the value of the GP for conducting the evaluation and making decisions about the results. Since, the researchers did not use the GP during this study, we will discuss what their results in terms of the problem identified by their sample, as well as the additional analysis and conclusions that would have been enabled by applying the GP.

3.1 The Behavior of the Evaluation Casa Cohort

The participants identified an amount of 12 different problems. For each one of these problems we coded the users' behavior (see Table 1) as 0 when a user did not identify a problem and with 1 when a user did identify it.

Table 1. Problems identified by each participant. The individual *p* represents the number of problems discovered by each participant divided for the total problems discovered by the sample. The weight of problems represents the percentage of the sample that have identified the same problem.

	Task 1			Task 2							Task 3		Individual
Problems	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	***p***
P1	0	0	0	0	0	1	0	1	1	0	0	0	0.25
P2	0	0	0	1	1	1	1	1	1	1	1	1	0.75
P3	0	0	0	0	0	1	0	0	1	0	1	1	0.33
P4	1	1	0	0	1	1	1	1	0	1	1	1	0.75
P5	1	1	0	1	0	1	0	1	1	1	0	0	0.58
P6	1	1	1	0	1	1	1	1	0	1	0	1	0.75
P7	0	0	0	1	1	1	1	0	1	1	1	1	0.66
P8	1	1	1	1	1	0	0	1	1	1	0	0	0.66
P9	1	0	0	1	0	0	0	1	0	0	1	0	0.33
P10	0	1	0	0	0	0	0	0	0	0	1	1	0.25
P11	0	0	0	1	0	1	1	0	0	1	0	0	0.33
P12	0	1	1	1	1	1	1	1	1	1	0	0	0.75
Weight	**42%**	50%	**25%**	58%	50%	75%	50%	67%	58%	67%	50%	50%	

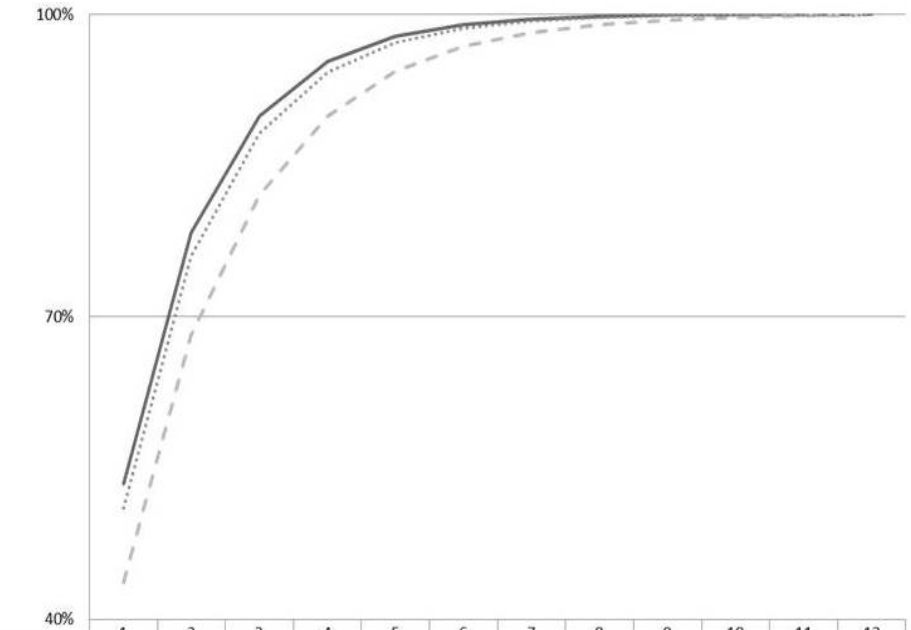
The weight of the problems can be used by manufacturers as an indicator of the sample behavior homogeneity or heterogeneity in discovering the problems. Usually in HCI, a sample can be considered heterogeneous when more than 50% of the problems are identified by only one participant. For instance, a sample of 10 users which identified a set of 10 problems, can be consider heterogeneous, whether 5 out of 10 of the identified issues are been experienced only once during the test. Nevertheless in medical device field, by looking for a most restrictive limit to increase the safety of the device, a sample can be considered heterogeneous when more than 50% of the problems are discovered by less than a half of the participants. For instance, whether 5 problems out of 10 are been identified by less than 5 users in a sample of 10.

In our evaluation case the sample is homogeneous as only two problems out of 12 are identified by less than 6 users (see Table 1). We estimate the *p* of the sample by applying three estimation models (Table 2): the ROI, the GT and the BDB. We do not report here the MC, because the results of this model are the same of the BDB one.

Table 2. Discovery likelihood of the sample (*p*) estimated by the Return of Investment (ROI) model, Good-Touring (GT) model and Bootstrap Discovery Behavior (BDB) model

	Estimation models		
	ROI	**GT**	**BDB**
P of the cohort	0.53	0.43	0.51

The sample shows a range of *p* from 0.43 to 0.53 (M: 0.49), by applying these values to the Error Distribution Formula (1), we may report that this cohort of 12 participants discovered between 98% to 99.9% problems of the device with a homogenous discovery behavior (fig. 1).



	1	2	3	4	5	6	7	8	9	10	11	12
ROI	53.47%	78.35%	89.93%	95.31%	97.82%	98.99%	99.53%	99.78%	99.90%	99.95%	99.98%	99.99%
GT	43.61%	68.20%	82.06%	89.89%	94.30%	96.78%	98.19%	98.98%	99.42%	99.67%	99.82%	99.90%
BDB	51.00%	75.99%	88.24%	94.24%	97.18%	98.62%	99.32%	99.67%	99.84%	99.92%	99.96%	99.98%

Fig. 1. Percentage of problems discovered by the sample on the base of the sample *p* changes after each participant analysis, estimated by the Return of Investment (ROI) model, Good-Touring (GT) model and Bootstrap Discovery Behavior (BDB) model

In light of our analysis, it is clear that more users do not need to be added to the evaluation as this would be a waste of resources; the probability of any new user identifying new problems whilst completing the same the three tasks is between 0.1% to 2% problems.

3.2 Save Your Investments and Guarantee the Reliability of the Assessment by the Grounded Procedure

As in the classic estimation studies [18], we can assign an arbitrary cost of £100 to each analysis and therefore conclude that to discover 12 problems the investment of the manufacturers was £ 1200.

Nevertheless, by using the average values of p of the three estimation models, we can estimate that evaluators reached 90% of the problems after the analysis of the first four users (i.e., $D(p_{ROI},p_{GT},p_{BDB})$=93.14%) and 97% after the first six (i.e., $D\ (p_{ROI},p_{GT},p_{BDB})$ =98.13%). In light of this, if the GP had been applied during the assessment of this BPM, after 6 users the manufacturers would have stopped the assessment, thereby obtaining a reliable results and saving 50% of the budget (£600).

We simulate the application of the GP steps during the evaluation case, by using a threshold percentage (D) of 97% of the total problems, as follow:

- Step 1: Manufacturers start the assessment with a sample of five users, and they compare the p of this initial sample to the standard (p=0.5) to decide whether to stop the assessment or add new users to the sample.
- Step 2: By looking at Table 3 manufacturers observe that the first five users identified 11 problems with a p ranging from 0.42 to 0.58 (M: 0.49). This discovery likelihood is close to the standard, and by applying the average p in to the *Error Distribution Formula* (1), the manufacturers may estimate that this sample of 5 users identified 96% of the problems, with an estimated range of D from 93% to 98%. Nevertheless, the sample is quite homogeneous, in fact, 6 problems out of 11 (55%) are discovered by more than 50% of the users, while the remained problems (45%) are discovered by less than a half of the sample.

Table 3. Problems identified by each participant during the analysis of the three tasks, with a sample of 5 users. This sample is quite homogeneous (55%), albeit there is a high percentage of heterogeneity (45%).

	Task 1		**Task 2**							**Task 3**		**Individual**
Problems	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	***p***
P1	0	0	0	0	1	0	1	1	0	0	0	**0.27**
P2	0	0	1	1	1	1	1	1	1	1	1	**0.81**
P3	0	0	0	0	1	0	0	1	0	1	1	**0.36**
P4	1	1	0	1	1	1	1	0	1	1	1	**0.81**
P5	1	1	1	0	1	0	1	1	1	0	0	**0.64**
Weight	**40%**	**40%**	**40%**	**40%**	100%	**40%**	80%	80%	60%	60%	60%	

- Step 3: Since the sample is quite homogeneous, manufactures could decide to stop the assessment. Nevertheless, by considering that the percentage of heterogeneity is high (45 %); practitioners may decide to add at least another user to increase the reliability of the evaluation data.

The manufacturers include a new user (i.e., number 6) in the sample. Finally, after another cycle of GP analysis (i.e., steps 1, 2 and 3), this new user increases the cohort p ($0.46<p<0.56$, M: 0.51), and as table 4 shows, a new usability problem is identified, and the sample becomes more homogeneous (2 problems out of 12 are discovered by less than 50% of the sample). On the basis of this data manufacturers have enough information to stop the assessment and report that the participants have identified a total amount of 12 problems, which represents 98.7% (97.7%<D<99.3%) of the possible issues that can be identified by a larger sample of end users interacting with the product during the three evaluation tasks.

Table 4. Problems identified by each participant during the analysis of the three tasks, when the user number 6 is added to the sample. This sample shows a behavior in tune with the cohort analyzed in table 1. The participants discover 12 problems and, the new user increases the homogeneity of the cohort; in fact, only 2 problems out 12 are identified by less than 50% of the sample.

	Task 1			**Task 2**							**Task 3**		**Individual**
Problems	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	***p***
P1	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**1**	**1**	**0**	**0**	**0**	**0.25**
P2	**0**	**0**	**0**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0.75**
P3	**0**	**0**	**0**	**0**	**0**	**1**	**0**	**0**	**1**	**0**	**1**	**1**	**0.33**
P4	**1**	**1**	**0**	**0**	**1**	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**0.75**
P5	**1**	**1**	**0**	**1**	**0**	**1**	**0**	**1**	**1**	**1**	**0**	**0**	**0.58**
P6	**1**	**1**	**1**	**0**	**1**	**1**	**1**	**1**	**0**	**1**	**0**	**1**	**0.75**
Weight	50%	50%	**17%**	**33%**	50%	100%	50%	83%	66%	66%	50%	66%	

In this case, both the overall p and the homogeneity of the sample are greatly increased when user 6 is added to the cohort. However, sometimes adding a new user may decrease both the homogeneity and the p of the cohort. This could happen for different reasons, such as selecting inappropriate users. In these cases, the manufactures have to reconsider the selection criteria, and restart the GP analysis after a new user analysis.

4 Conclusion

The GP helps manufacturers to decide when to stop the evaluation of a device when the optimal sample size is reached, thereby preventing wasting resources. Of course, the results of our evaluation case are not generalizable to the assessment of any other BPM or medical device. This is because the GP only indicates the reliability of the

data gathered during a specific evaluation process, meaning that with other participants or with other evaluation conditions, the GP outcomes will vary. As a result, there is no one single magic number of users for reliably testing a certain kind of device, and, as a consequence, the manufacturers should apply the GP for each evaluation, whether it be formative or summative. Finally, the diffusion of the GP in medical device field could significantly improve the possibility of manufacturers to release usable and safe product on the market by taking decisions during the life-cycle on the basis of the real data at hand.

References

1. Streitz, N.: From cognitive compatibility to the disappearing computer: experience design for smart environments. In: Proceedings of the 15th European Conference on Cognitive Ergonomics: the Ergonomics of Cool Interaction, pp. 1–2. ACM, Funchal (2008)
2. Soylu, A., Causmaecker, P.D., Desmet, P.: Context and Adaptivity in Pervasive Computing Environments: Links with Software Engineering and Ontological Engineering. Journal of Software 4, 992–1013 (2009)
3. Herman, W.A., Devey, G.B.: Future Trends in Medical Device Technologies: A Ten-Year Forecast. Food and Drug Administration, Center for Devices and Radiological Health (2011)
4. IEC: IEC 62366: 2007 Medical devices – Application of usability engineering to medical devices. CEN, Brussels, BE (2007)
5. ISO: ISO 14971:2000 Medical devices – Application of risk management to medical devices. CEN, Brussels, BE (2000)
6. ISO: ISO 13485:2003 Medical devices – Quality management systems – Requirements for regulatory purposes. CEN, Brussels, BE (2003)
7. ISO: ISO 14971:2007 Medical devices – Application of risk management to medical devices. CEN, Brussels, BE (2007)
8. ISO: ISO 15223-1:2012 Medical devices – Symbols to be used with medical device labels, labelling and information to be supplied – Part 1: General requirements. CEN, Brussels, BE (2012)
9. ANSI/AAMI: HE75: Human factors engineering-Design of medical devices. Association for the Advancement of Medical Instrumentation, Arlington, VA (2009)
10. Martin, J., Norris, B.J., Murphy, E., Crowe, J.A.: Medical device development: The challenge for ergonomics. Applied Ergonomics 39, 271–283 (2008)
11. Money, A., Barnett, J., Kuljis, J., Craven, M., Martin, J., Young, T.: The role of the user within the medical device design and development process: medical device manufacturers' perspectives. BMC Medical Informatics and Decision Making 11, 15 (2011)
12. Martin, J., Barnett, J.: Integrating the results of user research into medical device development: insights from a case study. BMC Medical Informatics and Decision Making 12, 74 (2012)
13. Heneghan, C., Thompson, M., Billingsley, M., Cohen, D.: Medical-device recalls in the UK and the device-regulation process: retrospective review of safety notices and alerts. BMJ Open 1 (2011)
14. Food and Drug Administration (FDA): Draft Guidance for Industry and Food and Drug Administration Staff - Applying Human Factors and Usability Engineering to Optimize Medical Device Design. U.S. Food and Drug Administration, Silver Spring, MD (2011)

15. Faulkner, L.: Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. Behavior Research Methods 35, 379–383 (2003)
16. Nielsen, J.: http://www.useit.com/alertbox/20000319.html
17. Nielsen, J.: http://www.useit.com/alertbox/number-of-test-users.html
18. Nielsen, J., Landauer, T.K.: A mathematical model of the finding of usability problems. In: Proceedings of the INTERACT 1993 and CHI 1993 Conference on Human Factors in Computing Systems, pp. 206–213. ACM, Amsterdam (1993)
19. Virzi, R.A.: Streamlining the Design Process: Running Fewer Subjects. In: Proceedings of the Human Factors Society 34th Annual Meeting, vol. 34, pp. 291–294. ACM, Santa Monica (1990)
20. Virzi, R.A.: Refining the test phase of usability evaluation: how many subjects is enough? Human Factors 34, 457–468 (1992)
21. Borsci, S., Londei, A., Federici, S.: The Bootstrap Discovery Behaviour (BDB): a new outlook on usability evaluation. Cognitive Processing 12, 23–31 (2011)
22. Caulton, D.A.: Relaxing the homogeneity assumption in usability testing. Behaviour & Information Technology 20, 1–7 (2001)
23. Lewis, J.R.: Evaluation of Procedures for Adjusting Problem-Discovery Rates Estimated From Small Samples. International Journal of Human-Computer Interaction 13, 445–479 (2001)
24. Schmettow, M.: Heterogeneity in the usability evaluation process. In: Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction, vol. 1, pp. 89–98. British Computer Society, Liverpool (2008)
25. Schmettow, M.: Sample size in usability studies. Communications of the ACM 55, 64–70 (2012)
26. Turner, C.W., Lewis, J.R., Nielsen, J.: Determining Usability Test Sample Size. In: Karwowski, W. (ed.) International Encyclopedia of Ergonomics and Human Factors, vol. 2, pp. 3084–3088. CRC Press, Boca Raton (2006)
27. Lewis, J.R.: Validation of Monte Carlo estimation of problem discovery likelihood (Tech. Rep. No. 29.3357). IBM (2000)
28. Borsci, S., Macredie, R.D., Barnett, J., Martin, J., Kuljis, J., Young, T.: Reviewing and Extending the Five–user Assumption: A Grounded Procedure for Interaction Evaluation (manuscript submitted for publication, 2013)