Towards Industrial-Strength Usability Evaluation

Martin Schmettow
Passau University
Informations Systems II
94032 Passau, Germany
schmettow@web.de

ABSTRACT
Usability professionals may face strict economic demands on the usability process in near future. This position paper outlines a research agenda to make usability evaluation a predictable and highly efficient engineering process.

Categories and Subject Descriptors
H.5.2 User Interfaces (e.g. HCI) Evaluation/methodology

Keywords
Usability Evaluation, Measurement, Process Quality

1.MOTIVATION
Usability professionals are never tired to stress the economic impact of good usability. And indeed, there are several compelling arguments: The first may be derived from the ISO norm 9241-11: Efficiency is regarded as one of the three main criteria of usability and can directly be converted into a bargain. For example, a very efficient interface to an enterprise information system makes users do their tasks more quickly which increases overall throughput. The second argument is specific to web usability. Web users are known to be very impatient with web sites having poor usability, especially with online purchasing; consequently usability directly affects the conversion rate of e-commerce companies. The third argument is from the perspective of software development. It is a widely accepted law, that defect fixing costs overlinearly depend on how early a defect was introduced and how late it was found. This is a justification for doing intensive usability evaluation early in system development.

But, many usability professionals still act under the paradigm of discount usability. In a broad sense this denotes: usability evaluation as a best effort strategy and conducted iteratively by experts who just know what they are doing. What, if clients or employers of usability professionals start taking the above economic arguments seriously? For example: What, if a start-up company has an innovative product idea and plenty of venture capital, but usability is mission-critical and they have only one shot? Will they rely on discount usability? Will they accept the good reputation of a usability company as the only guarantee? It is more likely, that they want objective preconditions, like a proven and certified evaluation plan. And maybe they even want quantitative guarantees and proven contract fulfillment, like: There is no show stopper left in the system and at least 90% of serious problems are identified. The paradigm of discount usability is inappropriate in such cases.

Research on the usability evaluation process has seen two major debates (research agendas, respectively): The Five-Users-Is-Not-Enough debate and the Damaged Merchandise debate. The Five Users debate is about how to reliably plan and control usability evaluation studies, whereas the Damaged Merchandise debate treats the topic of how to compare evaluation methods in fair and valid way. In the following, I will argue why we must continue these research agendas, in order to make usability evaluation a well understood and highly optimized engineering activity. But, I will also claim that we have to put off some blinders.

2.WHY TO CONTINUE THE “FIVE USERS” DEBATE
The five users debate goes back to Nielsen and Landauers suggestion to model the progress of evaluation studies as a geometric series [9]. Unfortunately, the debate was primarily carried by an oversimplification of Nielsen, who trivialized his own findings in stating that testing five users is enough in industrial practice [8]. This is, by the way, an excellent example of the discount usability paradigm, which may turn out obsolete.

In contrast, several researchers went deeper into the theoretical impact of this model: The phenomenon of variance in the process was discovered [3], good task design was found to be a major impact factor [6] and basic stochastic assumptions of the model were questioned [2]. A recent contribution was the proof that the geometric model is inherently flawed by falsely assuming that usability defects are equally visible and sessions equally effective [10]. Instead, the beta-geometric model, accounting for heterogeneity, was shown to better predict the process.

But, this is still an oversimplification that does not comprise all impact factors found in industrial studies. For example, recently I tried to fit the data reported from the CUE-4 study with the beta-geometric model – with disappointing results: The model could not sufficiently explain the overwhelming number of defects that were detected only once [7]. In consequence, there is still no reliable estimation of how many defects were left undetected. For the first, there are two options for enhancing the model in order to better fit the data and reliably plan and control usability studies: First, the study progress has to be tracked on the finer grained level of single tasks presented in a usability test (or imagined by usability inspectors). Specifically, this may help identify when a certain set of tasks is “exhausted” and replace it by new tasks that make further defects observable. Second, the current models do not handle the problem of false alarms in evaluation studies. These may well be liable for the misfit reported above. Currently, we are working on an enhanced model to incorporate the occurrence of false alarms and varying task sets. This hopefully enables us to better estimate the number of remaining defects (misses) and to give a probability for a reported defect being a false alarm. The latter may prevent wasting development resources on would-be defects and thus has direct economic impact.

3.BEYOND "CHASING THE HE"

The Damaged Merchandise debate arouse by the harsh critique of Gray and Salzmann on the poor validity of experiments on UEMs [5]. However, my main point here is not validity, but the observation that research on designing UEMs has not made much substantial progress. Even recent well designed studies are still very restricted in their contribution to understanding the cognitive or contextual factors of finding usability defects. Instead, they make more or less marginal adaptions to common inspection methods and compare this in a two conditional experimental design to the Heuristic Evaluation (HE). The observed effectiveness gains are in many cases marginal (e.g. [4]) or non-existent [11]). This "Chasing the HE" approach has the severe drawback of restricted insight. It lets us only know which of two procedures is (slightly) better. It does not inform about the specific interplay of impact factors granting effective defect identification. But, this is a precondition to design (much) better procedures, provide adequate training and adjust the evaluation process to business goals.

Only few studies have paid attention to successful versus unsuccessful cognitive-behavioral strategies of usability experts. To give an example for a rarely recognized work that has done better: Perspective based reading is a well known technique in software inspections and raises effectiveness by reducing cognitive load. Zhang et. al. have transferred this technique to usability inspection and have found likewise improvements [13]. Another positive example is how Woolrych et. al. analyzed the knowledge resources involved in usability inspections [1]. (They also made some points on how false alarms arise.)

These are interesting and relevant results, as they may lead to methods and training concepts for increased effectiveness of usability experts. But, there still is a lack of quantitative research on such topics. Especially, defects are likely having qualitative properties that make a difference with respect to behavioral strategies and knowledge resources. Frøkjær and Hornbæk have found differing detection profiles for two inspection methods after classifying defects with the User Action Framework [4]. Another promising way to go is to search for defect classes in the raw data from evaluation processes and derive an empirically valid classification Advanced statistical exploration techniques, like differential item functioning from item response theory [12] or binary cluster analysis probably apply well to this problem, in contrast to ordinary variance analysis. The strength of these techniques is that they to not require manipulating independent variables. Instead, they can reveal latent variables in existing data sets, including results from industrial studies.

These approaches may be used to profile methods according to their effectiveness regarding certain types of defects. In industrial settings this is useful for selecting a method appropriate to the development context. For example, we may purposefully choose a method for identification of task related defects early in development. Later in the development process another method may serve identification of superficial design issues. Another possibility is aligning the evaluation focus to business goals, e.g. evaluating for efficiency in case a system is primarily aimed at experts.

4.CONCLUSION

Modern software engineering is well regarding economic demands: efficiency of development processes, early defect discovery and aligning software qualities to business goals. The usability profession is still dragging a little behind, but may sometimes face their customers' claims for process approval, efficiency and guarantees. The aim of this paper was to point out valuable research agendas in the past, but to also identify future directions of research: Quantitative research with refined experimental designs and advanced statistical techniques may reveal relevant properties on several levels of the usability evaluation process. Knowing the properties on process level results in better approaches to plan and control studies towards given business goals. Knowing the properties on the cognitive-behavioral level are a precondition to significantly raise effectiveness and appropriateness of evaluation processes. Much can be achieved with advanced statistical techniques on existing data sets. The minimum to get is specific and well grounded hypotheses that will inspire for well designed and elaborate experimental studies to deeply understand the anatomy of usability evaluation.

5.REFERENCES

[1] Alan Woolrych, Gilbert Cockton, and Mark Hindmarch. Knowledge Resources in Usability Inspection. In *Proceedings of the HCI 2005*, 2005.

[2] David A. Caulton. Relaxing the homogeneity assumption in usability testing. *Behaviour & Information Technology*, 20(1):1–7, 2001.

[3] Laura Faulkner. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods, Instruments & Computers*, 35(3):379–383, 2003.

[4] Erik Frøkjær and Kasper Hornbæk. Metaphors of human thinking for usability inspection and design. *ACM Trans. Comput.-Hum. Interact.*, 14(4):1–33, 2008.

[5] Wayne D. Gray and Marilyn C. Salzman. Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-Computer Interaction*, 13(3):203–261, 1998.

[6] Gitte Lindgaard and Jarinee Chattratichart. Usability testing: What have we overlooked? In *CHI '07: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 1415–1424, New York, NY, USA, 2007. ACM Press.

[7] Rolf Molich and Joseph S. Dumas. Comparative usability evaluation (CUE-4). *Behaviour & Information Technology*, 27(3), 2008.

[8] Jakob Nielsen. Why you only need to test with 5 users. Jakob Nielsens Alertbox, March 19 2000. http://www.useit.com/alertbox/20000319.html.

[9] Jakob Nielsen and Thomas K. Landauer. A mathematical model of the finding of usability problems. In *CHI '93: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 206–213, New York, NY, USA, 1993. ACM Press.

[10] Martin Schmettow. Heterogeneity in the usability evaluation process. In David England and Russell Beale, editors,

Proceedings of the HCI 2008, volume 1 of *People and Computers*, pages 89–98. British Computing Society, 2008. in print.

[11] Martin Schmettow and Sabine Niebuhr. A pattern-based usability inspection method: First empirical performance measures and future issues. In Devina Ramduny-Ellis and Dorothy Rachovides, editors, *Proceedings of the HCI 2007*, volume 2 of *People and Computers*, pages 99–102. BCS, September 2007.

[12] Martin Schmettow and Wolfgang Vietze. Introducing item response theory for measuring usability inspection processes. In *CHI 2008 Proceedings*, pages 893–902. ACM SIGCHI, April 2008.

[13] Zhang Zhijun, Victor Basili, and Ben Shneiderman. An empirical study of perspective based usability inspection. Technical report, University of Maryland, Human-Computer Interaction Lab, 1998.