Optimizing Usability Studies by Complementary Evaluation Methods

Martin Schmettow
University of Twente
Enschede, The Netherlands
m.schmettow@utwente.nl

Cedric Bach
Bertin Technologies
Toulouse, France
bach@bertin.fr

Dominique Scapin
INRIA
Rocquencourt, France
dominique.scapin@inria.fr

This paper examines combinations of complementary evaluation methods as a strategy for efficient usability problem discovery. A data set from an earlier study is re-analyzed, involving three evaluation methods applied to two virtual environment applications. Results of a mixed-effects logistic regression suggest that usability testing and inspection discover rather disjunctive sets of problems. A resampling analysis reveals that mixing inspection and usability testing sessions in equal parts finds 20% more problems with the same number of sessions.

Usability Evaluation, Effectiveness, Virtual Environments, Logistic Regression, Mixed-effects Linear Model

1. INTRODUCTION

Finding usability problems is a key activity in user-centered design. Finding usability problems also is costly and, seemingly, it is very difficult to find (almost) all relevant problems in a system. Since introduction of the first usability evaluation methods, there was a long quest for the most effective and efficient way to find usability problems. In general, this has been approached under two perspectives: first, different evaluation methods have been compared against each other to find the most efficient one, and second, models have been devised for predicting effectiveness from the number of independent experts or participants in an evaluation study (i.e. the sample size). In the present paper, we address another way to increase effectiveness and efficiency of usability evaluation, which is to use combinations of complimentary methods. A case study¹ is presented

where three evaluation methods have been applied to the same interfaces. Two of them showed almost the same efficiency. It is shown through resampling analysis that effectiveness can be increased without modification of the method or increasing the sample size, but alone through mixing evaluation sessions from two or more different methods.

1.1 Sample size in usability evaluation

The obvious strategy to find more UPs is to increase the number of experts or test participants, albeit this being more costly. This also raises the question of how many experts or users is enough to reach a preset target, say 85% of all UPs. Nielsen & Landauer (1993) were among the first to attempt a mathematical approach, aiming to bring costs and values into balance. They conceived usability evaluation as a random experiment where the detection of a usability problems is the basic

¹ This present paper is a re-analysis of a data set first presented by Bach & Scapin (2010). The aim of the original study was to compare effectiveness of a novel inspection method to two other evaluation methods. The main result of this earlier analysis was a significant difference between document inspection (DI) and expert inspection (EI) in terms of effectiveness. Furthermore, the average efficiency of

usability testing (UT) and DI proved to be rather similar.

A preliminary version of the paper focused solely on the efficiency gain through complementary methods (Schmettow et al. 2010). The present paper elaborates on the theoretical links to recent mathematical models on sample size estimation in usability studies, and introduces a solid statistical methodology (logistic mixed-effects regression).

stochastic event. They modelled this process as Poisson distribution, which implicitly assumes that problems are equally likely to be discovered. Under the same mathematical assumption, the progress of problem discovery follows a geometric series, with the percentage of discovered problems D depending on the basic probability of discovery p and sample size N as:

$$D = 1 - (1 - p)^N \quad (1)$$

The geometric series model is also known as the curve of diminishing returns, as with increasing sample size the progress in discovering new problems decelerates. Obviously, this complicates matters when wanting to balance effort and costs.

Furthermore, several authors have pointed out that the assumptions of the geometric series model are not correct. Most notably, it seems unrealistic that all problems have the same probability p to be discovered (Kanis 2010; Schmettow 2012). In contrast, usability problems vary in visibility and this can severely decelerate the progress of discovery. Consequently, larger samples are required than is suggested by the so called “magic numbers” claims, like five (Nielsen 2000) or 8-12 (Hwang & Salvendy 2010). For example, a usability testing study on a novel medical infusion pump interface reportedly found 88% of the problems with a sample size of 34 users (Schmettow et al. 2013), which is way beyond suggested magic number. Because infusion pumps are comparably simple devices for a rather homogenous user group, the authors argue that testing more complex systems with diverse users calls for even bigger samples. Furthermore, they question that the common 85% rule (Nielsen 2000) is sufficient for critical systems. In consequence, effective usability evaluation may be much more costly than has been assumed in the past. While theoretically, effectiveness can always be improved through larger samples, this is practically limited due to the asymptotic nature of the process (the curve of diminishing returns).

1.2 Method effectiveness

Another strategy for improving effectiveness is to improve usability evaluation methods (UEM) themselves. In fact, countless studies have devised novel or modified procedures for usability evaluation (see Gilbert Cockton, Lavery, & Woolrych (2003) for an overview on expert-based evaluations).

Interestingly, several comparative studies also concluded on qualitative differences between evaluation methods. Frøkjær & Hornbæk (2008) compared a novel UEM based on psychological metaphors to usability testing (UT) and Heuristic

Evaluation (HE). In terms of average efficiency, the novel method did not stand out against HE. However, a posteriori comparison, involving a classification of UPs and severity ratings, revealed several qualitative differences between the methods. Some UPs were better visible with HE, others with the novel method.

Fu, Salvendy, & Turley (2002) showed qualitative differences between UT and HE. Noteworthy, these authors predicted qualitative differences from the model of action control by Rasmussen (1986). Indeed, they found that expert evaluations are better at uncovering UPs on the skill- or rule-based level of control, while usability testing is more efficient at knowledge-based UPs.

While Frøkjær & Hornbæk (2008) did not find an improvement in pure efficiency, they still concluded their novel method to be superior as it uncovered more severe problems. Going one step further, Fu et al. (2002) emphasized that methods have different strengths and weaknesses and thus may play their roles in different phases of the development cycle.

In this study, we further investigate qualitative differences between evaluation methods, and show that the combination of qualitatively different methods is beneficial for evaluation efficiency. The next section conveys our primary theoretical argument, capitalizing on recent theoretical findings on the relationship between visibility variance and evaluation efficiency.

1.3 Benefit of complementary methods

The majority of studies that compared evaluation methods, focus on the improvement in average visibility of usability problems, represented as p in the geometric series model (Eq 1).

As said previously, this model is inappropriate as it ignores that usability problems may differ in how easy they are discovered, which is called *visibility*. Recent findings suggest that it is inappropriate to ignore visibility variance, as progress of discovery is decelerated (Schmettow 2009). In other words: two evaluations that have the same average problem visibility will not necessarily make the same progress in discovering UPs. If one method has a more pronounced variance in problem visibility, discovery will proceed at a considerably lower rate, requiring larger sample sizes (Schmettow 2012).

A third strategy towards effective problem discovery may therefore be the *reduction of visibility variance*. One could approach this strategy by revising existing evaluation methods, for example adding new heuristics to HE. Here we examine another way that does not require modification of established methods: when two methods are sensitive for different subsets of usability problems, combining these methods

Table 1 Example showing the beneficial effects of method complementarity on visibility variance

	A	B	C	A & B	A & C
UP1	1	2	6	2	4
UP2	2	1	5	0	4
UP3	7	8	1	6	8
UP4	9	7	0	8	6
Visibility Variance	14.9	12.3	8.7	13.3	3.7

should effectively reduce visibility variance, resulting in more efficient problem discovery.

For an illustration, consider the following scenario: three evaluation methods A, B and C were applied to the same system, with a sample size of ten, each. Altogether, four usability problems were discovered, but with different effectiveness, as shown in Table 1. For example, UP1 was discovered six times with method C, but less frequently with methods A (1) and B (2). On the opposite, UP4 was 9 times found with A, but omitted completely with method C. Overall, it appears that those problems effectively discovered with A and B are difficult to discover with C, and vice versa.

The two columns to the right show the outcome when running each five sessions of A and B, A and C, respectively. As shown in the right-most columns, in the A & C evaluation process UPs have an almost uniform frequency of discovery. One can imagine that, perhaps, all four problems were readily discovered with half the sample size. This is very different to A & B, where UP2 is omitted completely. These patterns are directly linked to visibility variance. While the combination of similar methods A and B, does not significantly change visibility variance (13.3), it is strongly reduced when combining methods A and C (3.7), as these are complementary. Since visibility variance decelerates the progress of discovery (Schmettow 2012), we can expect the combination of A and C to be more efficient in discovering the four usability problems, as compared to the pure conditions A or C.

2. EXPERIMENTAL ANALYSIS

The present study compared three UEMs for desktop virtual environments. Although this particular application domain is not the primary stake of this paper, we give a short overview on this topic.

2.1 Evaluation of Virtual Environments

Virtual Environments (VE) are becoming widely used and have expanded to cover an extensive range of activities. An example of this expansion is the availability of applications such as Google Earth that allow computer-based access to 3D satellite maps. Although these applications have been adapted for office computers, in many new contexts of use their keyboard/mouse/screen-based interactions are not sufficient from a usability point of view. Advanced, enriched, even ubiquitous interactions using large display screens with remote interaction devices (e.g. laser pointers, oriented sound flows, gesture recognition) are more likely to be used (Dubois et al. 2008).

Actually, several studies have highlighted specific usability problems associated with VEs (Gabbard & Hix 1997). Stanney, Mollaghasemi, Reeves, Breaux, & Graeber (2003) have shown that the designers of VE systems cannot rely solely on the methods developed for standard 2D graphical user interfaces (GUIs) since their interaction styles and the use of 3D are radically different from standard GUIs.

Accordingly, a number of studies are concerned with the adaptation of existing UEMs such as cognitive walkthrough (Sutcliffe & Kaur 2000), usability questionnaires (Kalawsky 1999), heuristic evaluation (Sutcliffe & Gault 2004); and user testing (Tromp et al. 2003). Conducting user testing to evaluate VEs seems to be more difficult than testing GUIs or websites. Bowman, Gabbard, & Hix (2002) reveal a set of difficulties when conducting user testing studies on VEs: physical environment issues, evaluator issues, and user issues. This suggests that efficient user testing to evaluate complex VEs remains a challenge. This could be a reason explaining the lack of available results in the literature.

Several authors (Bowman et al. 2002; Sutcliffe & Gault 2004) claim that with regard to sample size and efficiency evaluation methods for VE are similar to the results of Nielsen & Landauer (1993). However, these claims are not sufficiently supported by empirical results and, as explained above, the commonly used geometric series estimator for required sample sizes is optimistically biased.

2.2 Research Questions

First, we hypothesize that *visibility of a particular problem depends on the employed evaluation method (RQ1)*. If this turns out to be the case, then *mixing two methods should result in lower visibility variance (RQ2)*, if these have complementary problem discovery profiles. In effect, *the combination makes*

the evaluation process more efficient (RQ3) – more problems are discovered with the same sample size.

3. METHOD

In the following, we briefly present the empirical setup of the study, which is a typical comparison of usability evaluation methods (UEM). A comprehensive description of the study can be found in the original publication (Bach & Scapin 2010).

3.1 Material

Three usability evaluation methods (UEM), user testing (UT), document-based inspection (DI) and expert inspection (EI), were separately used to evaluate two VEs: an educational software (a 3D video game tutorial, referred to as EDU) and a 3D map of a mountain valley (a landscape in the Alps, referred to as MAP).

EDU follows a rather constrained scenario, which requires carrying out the tasks progressively in order to move from one task to the next. The scenario provides 35 tasks at various levels of difficulty. The system can simply require the participant to press a key or to carry out a complex task requiring planning, sub-objectives to reach and movements.

MAP allows a user to freely explore a 3D view of the mountain valley generated from high definition geographical data (aerial pictures and/or satellites). It allows the user to collect tourist information about the valley through information panels or links to websites.

3.2 Sample

Ten participants took part individually in user testing and 19 junior experts took part in inspections (10 in document-based inspections and 9 in expert inspections). The group of participants in user testing consisted of 5 men and 5 women, 19 to 24 years old, the average being 21.8 years ($\sigma = 1.5$). All participants' sight and hearing abilities were normal or corrected-to-normal. All participants used regularly a traditional computer (i.e., GUI, screen, keyboard, mouse) at the university. Initially, participants sought for this study were those familiar with classic computer equipment but not with VE applications.

The 19 participants in the two inspection conditions (DI and EI) were all fifth year students in work psychology, also trained in software ergonomics. The training was mainly theoretical and did not cover the ergonomic criteria for GUIs, which the DI method is based upon. Neither did the participants had practical experience in usability inspection, nor did they have previous experience with the two VE applications. The

Table 2 Experimental conditions

Condition	# participants	# problems	
		EDU	MAP
Document inspection	10	79	88
Expert inspection	9	39	52
Usability testing	10	76	84
Combined	29	127	147
# tokens		3683	4263

participants were randomly assigned to the two inspections conditions: 10 students for DI (five female; age, 24.5 years, $\sigma = 2.5$) and 9 students for EI (six female; age: 26 years, $\sigma = 7$).

3.3 Design and procedure

Participants were assigned to one of the three method conditions and had to evaluate both VE systems. Each experimental session was one hour long (30 minutes to evaluate each VE). Each experimental condition produced a set of usability problem observations. Table 2 shows the number of problems successfully discovered in each condition. For the data analysis a total of 3686 dichotomous tokens ("hit" or "miss") in the EDU condition and 4263 in MAP were recorded.

3.4 Data analysis

Twenty-nine hours of usability evaluation activity performed in a laboratory context were recorded and analyzed. In the following it is briefly described how the raw observations were classified and aggregated into usability problems. Then details on the quantitative data analysis are given.

3.4.1 Classification of UPs

A strict procedure was used for documenting problems and matching them under a common format (Hornbæk & Frøkjær 2008). The method comparisons were first carried out using the problem classification based on Ergonomic Criteria, which has already been demonstrated to be effective (Bach et al. 2003). Ergonomic Criteria allows two levels of classification, eight primary criteria and 20 secondary criteria.

The documenting step corresponds to the individual description of usability problems by evaluators, using a structured format. It sometimes also included some notes on severity. Such a description differs depending on the UEM. The documenting step involved data collection, organization and

homogenization of the problems diagnosed and documented directly by the participants in the inspections. For user testing, participants' interactions and comments were recorded to facilitate direct and post experiment interpretation. During the interpretation of the evaluation results, problems were analyzed by experimenters as they were expressed in the context of their first appearance, by replaying the application and checking the participants' comments from recorded videos. There, the issue is to distinguish between real problems and false alarms. This was achieved through consensus of two experts.

The matching step is usually conducted to compare sets of usability problems and to identify duplicate problems. Ergonomic Criteria as well as the recommendations by Cockton & Lavery (1999) were used to link observations to usability problem descriptions. While matching observed tokens to problems, special care was given to checking the equivalence in description and granularity between inspection-based problems and user testing-based problems. For each identified usability problem, an ergonomic criterion was assigned in order to build an organized map showing the distribution of the usability problems. This allowed an assessment of the diversity of the problems. Problem instances have been considered a match when the problem identification context, the interaction object concerned, and/or the interaction consequences (observable or inferable state changes) are similar (Cockton & Lavery 1999). This procedure allowed us to make a coherent set of data for conducting further statistical analysis.

3.4.2 Logistic regression

Statistically, usability evaluation studies can be conceived as a series of independent attempts (sessions) to discover a set of usability problems (Schmettow & Vietze 2008). A single session (expert or test user) is a random experiment, where any existing problem is either encountered or missed. Hence, the outcome on every problem is either a "hit" or a "miss". In the statistical literature, this is often referred to as *dichotomous* data or as *presence-absence* data (in ecology).

Most past studies that compared evaluation methods used classic statistical techniques, such as linear regression or ANOVA. However, one assumption of ANOVA is that the outcome variable has the range $[-\infty; \infty]$. Obviously, that does not match the situation where the outcome is the probability of success, being strictly in the range $[0; 1]$. Another issue is the distribution of error terms, which for ANOVA needs to be Gaussian and homoscedastic. In contrast, counts of dichotomous events (miss or success) typically result in binomially distributed residuals. This differs

from the Gaussian error term in two respects: it typically is not symmetrically bell-shaped and variance is not constant, but depends on the mean p as:²

$$\sigma^2 = np(1 - p) \quad (2)$$

For presence-absence data, the appropriate method is logistic regression, a member of the generalized linear models (GLM) family (Hardin & Hilbe 2007). Logistic regression renders the relationship between successes in a number of trials and deliberate metric or categorical predictors. Coefficients in logistic regression models are on a logit scale, which is the inverse of the logistic function, hence the name.

3.4.3 Random effects

In most empirical studies, where researchers are interested in the effect of a treatment or predictor, hypotheses are almost exclusively stated as a linear relationship (with continuous predictors) or difference in means (in factorial designs). Most of the time, variance is viewed as just a nuisance parameter; strong variance makes it necessary to increase the sample size (or use more expensive instruments) to reach a certain level of precision, but it does not convey any interesting information. In the present study, explicit modeling of variance is crucial for two reasons:

First, we are interested in how the visibility of individual problems differs between methods. Note that this is totally different to *mean* visibility changing between methods, as this is variation due to a manipulated variable. This is commonly called a *fixed effect*, whereas unexplained variation in the sample is referred to as a *random effect*.

Second, when using logistic regression one has to take special care of modelling residual variance correctly³. With Binomial distribution, variance is strictly tied to the probability parameter (review Eq. 2), without any additional scaling parameter (as σ^2 in Gaussian distributions). If variance of residuals in logistic regression is larger than nominal, one speaks of *over-dispersion*, which is a sign of visibility variance (Schmettow 2009). For the data analysis, we use the method of *mixed-effects logistic regression*⁵ to deal with over-dispersion and make inference about

² In Binomial distribution, variance is largest around $p = .5$ and decreases towards $p = 0$ and $p = 1$.

³ A common misconception is that generalized linear models relax assumptions of linear models. The opposite is the case, as they all make similar strict, but different, assumptions about range, residual distribution and variance structure.

⁵ To avoid confusion: *mixed effects* is totally unrelated to the concept of *mixing complementary methods*.

variance. Two types of random effects go into the regression model: visibility variance within methods takes the form of a so-called *intercept random effect*, whereas the variability of visibility between methods is modeled as a *slope random effect*. A strong slope random effect indicates that the visibility of individual problems changes unsystematically between methods. This is taken as primary indicator for method complementarity. Lastly, another intercept random effect was introduced for subjects, thereby accounting for individual differences in identifying usability problems.

Markov Chain Monte Carlo sampling was used to estimate the mixed effects model (Hadfield 2010). All statistical analysis was performed with the statistical programming environment R (R Development Core Team 2011).

4. RESULTS

First, it is examined how the visibility of individual problems varies by method, using a mixed-effects logistic regression (RQ1). Subsequently, by a resampling analysis it is demonstrated how mixing complementary UEMs decreases variance of problem visibility (RQ2), resulting in improved efficiency (RQ3). For the sake of brevity, all analysis steps were performed on both applications, MAP and EDU, merged, resulting in a total 274 usability problems. This is legitimate as application was a within-subject factor.

4.1 Problem visibility by method

A first indication for complementarity of UEMs is the number of problems that are discovered with one method, but not with another. Figure 1 shows the

Figure 1 Overlap of usability problems as found in the three conditions DI, EI und UT

Table 3 Results of mixed effects logistic regression

Effect	Coef	low-CI 95%	up-CI 95%
<i>Random effects</i>			
Intercept (DI)	1.44	1.07	1.85
Method EI	.62	.20	1.40
Method UT	6.27	4.53	8.09
<i>Fixed effects</i>			
Intercept (DI)	-2.03	-2.22	-1.81
Method EI	-1.45	-1.82	-1.16
Method UT	-.46	-.82	-.10

intersection between the three method conditions. Strongest separation were observed between DI and UT: 87 problems (32%) were found in at least one UT session, but were totally overlooked in the DI condition. A similar number of 94 problems (34%) has been discovered by DI experts, but was not encountered by any UT participants. While the intersection between UT and EI is similarly small, there seems to be quite some commonality between the two inspection methods.

A logistic linear mixed-effects regression is estimated with UEM as a fixed factor, accounting for systematic changes in mean visibility between methods. Two random effects are introduced to the model: an intercept random effect for variance in problem visibility in the reference group DI, and a slope random effect for visibility changing between methods. A third intercept random effect accounts for individual differences in subjects.⁶

As shown in Table 3 (fixed effects), DI has the highest average discovery rate of the three methods. UT performs only slightly below DI, whereas EI performs poorly ($p \leq .001$). The intercept random effect is clearly above zero; problems differ considerably in visibility when identified by the DI method. The slope random effect for EI is comparably small, the lower 95% credibility limit nearly approaches zero. Except for the systematic lower discovery rate, problems have similar relative visibility in both inspection methods.

In contrast, the slope random effect between DI and UT is very pronounced (6.27). The relative change in visibility is more than four times stronger than visibility

⁶ Weak priors were used to obtain estimates similar to maximum likelihood. MCMC samples was set to 1,000,000, with a burn-in of 500,000. Convergence was checked on a time series plot. 95% credibility intervals were obtained using the highest posterior density intervals on the sampled posterior distribution.

variance within the DI condition (1.44). Seeing that UPs strongly change visibility from DI to UT is a clear indicator of method complementarity.

4.2 Reducing variance by mixing methods

The mixed-effects analysis showed that the methods DI and UT have similar average detection capabilities, but visibility differed strongly on the level of individual problems. According to RQ2 we expect that such complementarity reduces visibility variance.

In a resampling experiment similar to (Schmettow & Niebuhr 2007), mixed groups ($n = 1000$) of ten sessions are drawn from either two conditions. These groups varied in proportion from 1/9 to 9/1, with the UT/DI condition and the two pure groups, 10/0 and 0/10. For each composed sample, variance is recorded of how often problems are discovered. As the top graph in Figure 2 shows, the pure DI condition has a lower variance ($\sigma^2 = 4.47$) compared to UT ($\sigma^2 = 7.26$). The variance of mixing DI and UT at a 7/3 proportion is considerably lower compared to both pure groups ($\sigma^2 = 3.67$). The middle and bottom graph show mixed evaluations involving EI. Irrespectively whether one mixes EI with UT or DI, the lowest variance is found with the maximum number of nine EI members. Adding members from UT or DI always inflates variance. In fact, this result is not very surprising: Eq.2 expresses the relationship between p and the variance. In all three conditions, p is smaller than 0.5. Most likely, the variance of the EI group must be smaller due to the smaller p .

4.3 Benefits of mixing methods

DI and UT were shown to have quite different profiles and visibility variance was effectively reduced in mixes. Therefore, these two methods are promising candidates for a complementary-method strategy (RQ3). In contrast, EI is similar to DI, but overall inferior. Still, as EI is complementary to UT; we may expect some benefit of UT/EI mixes as well.

Table 4 Effectiveness and benefit of optimal DI/UT mixes by sample size

Group size	4	6	8	10
Optimal Mix	2/2	3/3	4/4	5/5
Effectiveness (# problems)				
Pure DI	114	137	154	167
Pure UT	115	134	147	160
Optimal Mix	136	164	185	203
Benefit (# problems)	18%	20%	20%	22%

To assess the potential benefits of mixing methods, the results from the resampling experiment are analyzed once again. For each sampled group, effectiveness is recorded as the number of identified problems. As shown in Figure 3, mixing complementary methods increases effectiveness. This is most apparent in the upper graph, showing DI/UT mixes. All mixed proportions are on average more effective than both pure groups. The optimal proportion is to have DI and UT sessions in equal parts (5/5), yielding 202 problems on average. The optimal mixed strategy is substantially more effective as the pure DI (167) and pure UT (160) strategies.

The previous analysis has shown that EI is overall inferior in problem discovery. Still, even adding EI sessions to an UT process is of some benefit (Figure 3, middle). Adding three EI sessions yields six more problems (166) as compared to a pure UT strategy. In contrast, there is no benefit in combining EI and DI, confirming that complementarity of methods is creating the benefit.

5. DISCUSSION

Three evaluation methods were compared on two virtual environment systems. While one method, expert inspection (EI), performed generally poor, the two methods, document inspection (DI) and usability testing (UT) showed similar overall performance. However, there was very little consistency in visibility of problems between the methods and a large proportion of problems went undiscovered through either method alone. In a way, DI and UT *do different things equally well*.

In the present study, complementary methods seem to counterbalance each other's weakness; in effect, more problems are discovered with less effort. Many previous attempts aimed at improving a single method's effectiveness; the effects were often small to marginal. In contrast, the benefit of optimally mix of methods is considerable: 20% better effectiveness at discovering problems and cost savings of up to 40%.

When empirical data is lacking, we believe that one can also identify complementary methods by common sense alone. For example, the method of Cognitive Walkthrough for the Web (Blackmon et al. 2002) semi-automatically assesses the appropriate labeling of links to measure the 'information scent', but ignores other relevant features, like layout and graphical appearance. This method is a possible candidate to complement with other methods, like usability testing or inspection with guidelines.

Figure 2 Effects of different mixes of methods on visibility variance

Figure 3 Effects of different mixes of methods on problem discovery effectiveness

Fu et al. (2002) conclude that usability researchers should first run expert inspections to eliminate skill and rule-based errors in early design phases and subsequently turn to usability testing. We disagree for two reasons: first, it seems plausible that knowledge-based problems are often related to essential user requirements, such as mapping of domain concepts and workflow. Often, these kind of problems are deeply rooted in a system's architecture, for example the data model. In Software Engineering it is well known that costs for fixing defects are higher, the earlier a defect had been introduced and the later it was discovered (Boehm & Basili 2001). Second, Fu et al. (2002) seem to assume that running both methods in one phase or iteration comes at greater costs. Our results indicate the opposite: using a mix of complementary methods can result in cost reduction.

As another remark to Software Engineering, the concept of perspective-based reading is well regarded in software inspection (Shull et al. 2000). The underlying idea is that inspection of engineering artifacts is most effective when several experts each focus on one specific quality aspect. Zhang et al. (1999) successfully transferred this idea to usability inspection.

To conclude, in our study we saw a particularly unsettling effect: all three evaluation methods were almost blind for a substantial subset of usability problems. Apparently, usability problems are too diverse to catch them with one approach. More generally, Cairns & Thimbleby (2003) characterize usability as a *diverse* concept, arguing that diversity of approaches in HCI is necessary to maximize usability by the principle of complementarity. While Cairns and Thimbleby mostly capitalize on the philanthropic spirit of HCI as a discipline, we argued rather economically: diversity of methods counterbalances the diverse nature of usability by the principle of complementarity, adding value and saving effort.

6. REFERENCES

- Bach, C. et al., 2003. Adaptation of Ergonomic Criteria to Human-Virtual Environments Interactions. In *Proceedings of Interact 03*. Amsterdam: IFIP, IOS Press, pp. 880–883.
- Bach, C. & Scapin, D., 2010. Comparing Inspections and User Testing for the Evaluation of Virtual Environments. *International Journal of Human-Computer Interaction*, 26(8), pp.786–824.
- Blackmon, M.H. et al., 2002. Cognitive walkthrough for the web. In *Proceedings of the SIGCHI conference on Human factors in computing systems Changing our world, changing ourselves - CHI '02*. New York, New York, USA: ACM Press, p. 463.
- Boehm, B.W. & Basili, V.R., 2001. Software Defect Reduction Top 10 list. *IEEE Computer*, 34(1), pp.135–137.
- Bowman, D.A., Gabbard, J.L. & Hix, D., 2002. A Survey of Usability Evaluation in Virtual Environments: Classification and Comparison of Methods. *Presence: Teleoperators and Virtual Environments*, 11(4), pp.404–424.
- Cairns, P. & Thimbleby, H., 2003. *The diversity and ethics of HCI*.
- Cockton, G. & Lavery, D., 1999. A framework for usability problem extraction. In *Proceedings of Interact 99*. Amsterdam: IOS Press, pp. 344–352.
- Cockton, G., Lavery, D. & Woolrych, A., 2003. Inspection-based Evaluations. In *The human-computer interaction handbook: fundamentals, evolving technologies and emerging applications*. Lawrence Erlbaum Associates, Inc., pp. 1118–1138.
- Dubois, E., Bach, C. & Truillet, P., 2008. Comparing Mixed Interactive Systems for Navigating 3D Environments in Museums. In T. N. C. Graham & P. Palanque, eds. *Proceedings of DSVIS 2008*. Springer Verlag, pp. 15–28.
- Frøkjær, E. & Hornbæk, K., 2008. Metaphors of human thinking for usability inspection and design. *ACM Trans. Comput.-Hum. Interact.*, 14(4), pp.1–33.
- Fu, L., Salvendy, G. & Turley, L., 2002. Effectiveness of user testing and heuristic evaluation as a function of performance classification. *Behaviour & Information Technology*, 21(2), pp.137–143.
- Gabbard, J.L. & Hix, D., 1997. *A taxonomy of usability characteristics in virtual environments*. Virginia Tech, Blacksburg.
- Hadfield, J., 2010. MCMC methods for multi-response generalized linear mixed models: the MCMCglmm R package. *Journal of Statistical Software*, 33(2), pp.1–22.
- Hardin, J.W. & Hilbe, J.M., 2007. *Generalized Linear Models and Extensions* 2nd ed., stata press.
- Hornbæk, K. & Frøkjær, E., 2008. Comparison of techniques for matching of usability problem descriptions. *Interacting with Computers*, 20(6), pp.505–514.
- Hwang, W. & Salvendy, G., 2010. Number of people required for usability evaluation: the 10±2 rule. *Commun. ACM*, 53(5), pp.130–133.

- Kalawsky, R.S., 1999. VRUSE--a computerised diagnostic tool: for usability evaluation of virtual/synthetic environment systems. *Applied ergonomics*, 30(1), pp.11–25.
- Kanis, H., 2010. Estimating the number of usability problems. *Applied ergonomics*, 42(2), pp.337–347.
- Nielsen, J., 2000. Why you only need to test with 5 users. Available at: <http://www.useit.com/alertbox/20000319.html> [Accessed April 11, 2016].
- Nielsen, J. & Landauer, T.K., 1993. A mathematical model of the finding of usability problems. In ACM Press, ed. *Proceedings of INTERCHI 1993*. New York, NY, USA: ACM Press, pp. 206–213.
- R Development Core Team, 2011. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.
- Rasmussen, J., 1986. *Information Processing and Human-Machine Interaction: An Approach to Cognitive Engineering*. New York, NY, USA: Elsevier Science Inc.
- Schmettow, M., 2009. Controlling the usability evaluation process under varying defect visibility. In *BCS-HCI '09: Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology*. Swinton, UK: British Computer Society, pp. 188–197.
- Schmettow, M., 2012. Sample size in usability studies. *Communications of the ACM*, 55(4), p.64.
- Schmettow, M., Bach, C. & Scapin, D.L., 2010. Effizientere Usability Evaluationen mit gemischten Prozessen. In J. Ziegler & A. Schmidt, eds. *Mensch und Computer 2010: Interaktive Kulturen*. München: Oldenbourg Verlag, pp. 271–280.
- Schmettow, M. & Niebuhr, S., 2007. A Pattern-based Usability Inspection Method: First Empirical Performance Measures and Future Issues. In D. Ramduny-Ellis & D. Rachovides, eds. *BCS HCI '07: Proceedings of the 2007 British Computer Society Conference on Human-Computer Interaction*. People and Computers. pp. 99–102.
- Schmettow, M. & Vietze, W., 2008. Introducing item response theory for measuring usability inspection processes. In *Proceeding of the twenty-sixth annual CHI conference on Human factors in computing systems - CHI '08*. New York, New York, USA: ACM Press, p. 893.
- Schmettow, M., Vos, W. & Schraagen, J.M., 2013. With how many users should you test a medical infusion pump? Sampling strategies for usability tests on high-risk systems. *Journal of biomedical informatics*, 46(4), pp.626–641.
- Shull, F., Rus, I. & Basili, V., 2000. How perspective-based reading can improve requirements inspections. *Computer*, 33(7), pp.73–79.
- Stanney, K.M. et al., 2003. Usability engineering of virtual environments (VEs): identifying multiple criteria that drive effective VE system design. *International Journal of Human-Computer Studies*, 58(4), pp.447–481.
- Sutcliffe, A. & Gault, B., 2004. Heuristic evaluation of virtual reality applications. *Interacting with Computers*, 16(4), pp.831–849.
- Sutcliffe, A.G. & Kaur, K.D., 2000. Evaluating the usability of virtual reality user interfaces. *Behaviour & Information Technology*, 19(6), pp.415–426.
- Tromp, J.G., Steed, A. & Wilson, J.R., 2003. Systematic Usability Evaluation and Design Issues for Collaborative Virtual Environments. *Presence: Teleoperators and Virtual Environments*, 12(3), pp.241–267.
- Zhang, Z., Basili, V. & Shneiderman, B., 1999. Perspective-based usability inspection: An empirical validation of efficacy. *Empirical Software Engineering*, 69, pp.43–69.