Heterogeneity in the Usability Evaluation Process

Martin Schmettow
University of Passau
Information Systems II
Innstr. 43
94032 Passau, Germany
schmettow@web.de

ABSTRACT

Current prediction models for usability evaluations are based on stochastic distributions derived from series of Bernoulli processes. The underlying assumption of these models is a homogeneous detection probability despite of it being intuitively unrealistic. This paper contributes a simple statistical test for existence of heterogeneity in the process. The compound beta-binomial model is proposed to incorporate sources of heterogeneity and compared to the binomial model. Analysis of several data sets from the literature illustrates the methods and reveals that heterogeneity occurs in most situations. Finally, it is demonstrated how heterogeneity biases the prediction of evaluation processes. Open research questions are discussed and preliminary advice for practitioners for controlling their processes is given.

Categories and Subject Descriptors

H.5.2 [*User Interfaces (e.g. HCI)*]: Evaluation/methodology

General Terms

Measurement, Human Factors

Keywords

Usability Evaluation, Five Users Debate, Evaluation, Stochastic Models, Overdispersion, Process Prediction

1. INTRODUCTION

Usability evaluation for finding usability defects is a vital activity in the development of complex interactive products. According to a well known law of Software Engineering the costs of fixing a defect is an overlinear function of how early a defect was introduced and how late it was uncovered [1]. Thus, when usability is critical for business success a rational choice was to set a very ambitious goal, say 99% of defects detected. This topic was discussed by Jared Spool (CEO of User Interface Engineering) in his keynote at the HCI 2007 conference. He argued that with e-commerce and peer-to-peer business models usability is a crucial quality which may cost or save a company millions of dollars a day (the examples were Amazon and Ebay).

For usability consulting companies (or departments) this poses considerable problems on planning and managing usability evaluation studies. The first problem is to accurately calculate projects. At the time of negotiation with the customer there is little known on how many defects are in the system, how easy they turn out to be detected and how many test persons are needed. Instead, the contracting manager has to rely on her expectations based on past experiences. This is problematic, as reflected by the "Five users is not enough" debate, where it turned out that evaluation studies vary a lot in the required sample size. The risks at this stage are to underestimate the effort and run into a project deficit. On the other hand, when the calculation is too gratious, the customer may choose the tightly calculated offer of a competitor (who may then struggle with project deficits).

Assumed that an evaluation study started with a reasonable budget, it is still essential to track the progress, meaning the rate of usability defects found at each stage of the study. If Spool's prevision is taken seriously, customers of usability consulting companies may in future even claim a guarantee for a certain rate of detected defects. This implies that both – customer and usability company – have means to accurately estimate what proportion of defects were truly revealed by the study.

Since the early nineties, this is the motivation for studying models describing and predicting the evaluation process. The principal problem is that the usability evaluation process is stochastic. That is, there is always random variation in how defects are detected. Consequently, estimators like the detection rate are always uncertain. Fortunately, the preciseness of estimators usually increases with larger sample sizes. For planning and controlling the evaluation process this has two implications: First, the more projects and their outcomes are known to the contracting manager the better will be her guess of required effort. Second, at the beginning of the evaluation process the estimators are very unreliable. This changes when the study proceeds and more data sets become available. The closer the study approaches the defined goal, the more reliable will be the estimator, which allows for an informed decision of whether to test another couple of participants or not.

But, in order to finally have reliable estimators it is essential to apply an *appropriate* statistical model for the process. If the wrong model is chosen, the random variation will still decrease, but finally the estimator will either under- or overestimate the real situation. Both is harmful.

In this paper I will at first give a brief overview on the past and recent approaches to predict the evaluation process. Subsequently, the approaches and their underlying arguments will undergo a more thorough statistical discussion. In particular, this will show that a basic assumptions of these models is violated: the probability of detecting defects is not homogeneous, but *heterogeneous*. It differs across inspectors (or test participants) and defects. For revealing the heterogeneity in the process, a statistical test is developed which is simple enough to be applied by practitioners without a comprehensive statistical software at hand.

Furtheron, the matter of compound distributions is introduced, which is a more accurate description of what is going on in the

Reprint of a paper originally published by the British Computing Society. Cite as: 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.

process. The compound beta-binomial model allows for estimating the factors introducing the extra variance. Finally, I will investigate to what extent incorporating heterogeneity improves monitoring and predicting the evaluation process.

All models and procedures introduced here are applied to five real data sets previously published. Two of these are from think-aloud testing studies, three are inspection-based evaluations. Please note, that from the mostly mathematical perspective expressed throughout the paper the evaluation method is not essential. Therefore, the term *session* will always refer to both, individual test participants and expert evaluators (inspectors). Accordingly, the term *process size* denotes the number of independant sessions conducted at a certain stage in the evaluation process.

2. RELATED WORK

One of the first authors examining the stochastic nature of the usability evaluation process was Virzi [24]. The process of finding usability defects in a series of think-aloud testing sessions is proposed to follow the cumulative function of the geometric distribution (CGF, also known as the *curve of dimishing returns*). The CGF plots the relative outcome (i.e. percentage of defects detected so far) of the process as a function of the process size s (i.e. number of independent testing sessions or inspections) and the detection probability p.

$$P(\text{identified at least once}|n\,\text{Inspectors}, p) = 1-(1-p)^n \quad (1)$$

Virzi ran a series of usability tests and used a Monte Carlo procedure to estimate the amount of defects found with each process size s. Noteworthy, Virzi was already aware of heterogeneity in the process: He estimated different curves for defects grouped by a severity ranking . But finally, the average detection probability was estimated to be $p = .35$ and it was concluded that under these circumstances five sessions suffice to find approximately 80% of the defects.

Nielsen and Landauer continued the topic. They claimed that under the assumption of independence of single evaluation sessions the margin sum of defects found per session follows a Poisson distribution [15]. Again, these authors are aware of heterogeneity in the evaluation process. In addition to the detectability of defects they mention the heterogeneity of test persons or inspectors and the possibility that usability problems co-occur, which violates the assumption of stochastic independance.

This study also revealed that the basic probability varies a lot between studies [15]. Consequently, one cannot rely upon an a priori value for p for controlling an evaluation study. A way out is to estimate the basic probability from the early part of the study in order to extrapolate the rest of the process towards the given process goal. A number of procedures have been tried to estimate p for a particular study. Nielsen and Landauer used a least-square fitting procedure for estimating the basic probability. Later, Lewis introduced another approach in simply estimating p from the response matrix of detection events happened so far. As he had to admit later, this procedure is inherently flawed and leads to an overestimation of p [10].

A main goal of this contribution is to show, that the problem of unprecise process prediction is not due to the procedures for estimating p, but a fundamental false assumption with the CGF model. The CGF model assumes that there is a homogeneous probability for any defect and session throughout the process. Whereas Virzi and also Nielsen and Landauer argue on differences between sessions and defects regarding detection probability, Lewis seems to mostly ignore this topic .

In fact, a few other authors have treated the problem of heterogeneity in a variety of ways. Caulton extended the CGF model for heterogeneity of groups regarding defect detectability in usability testing [5]. It was assumed that there exist distinct subgroups of test persons that are sensitive to certain subsets of usability defects. This model may be applicable for situations with highly diverse user groups. A stricter formulation of this model was to assume that users or evaluators differ in their sensitivity or skills to detect defects. This can be expressed with the following modification to the CGF model, where p_i denotes the sensitivity or skill of an individual person:

$$P(\text{identified at least once}|p_1 \dots p_n) = 1-\prod_{i=1}^{n}(1-p_i) \quad (2)$$

Woolrych and Cockton [26] re-examined data of a previous study and considered the CGF model as risky, because it does not account for heterogeneous detection probability of defects. They suggest to replace p with a kind of density distribution.

Recently, Schmettow and Vietze suggested the Rasch model for measuring the impact factors of defect detectability and evaluator skills [22]. Under certain assumption this model provides a density distribution for p with evaluator skills s_i and defect difficulty-to-detect d_j as parameters.

$$P(\text{j identified by i}|s_i, d_j) = \frac{e^{s_i-d_j}}{1+e^{s_i-d_j}} \quad (3)$$

They suggested, that these individual measures may be used for prediction of process outcome . This appears promising, given that the model turns out to fit the data from real processes.

Faulkner observed a large variance of outcome with small sample size studies [7]. This was not directly assigned to heterogeneinity in the process, but it stresses an important point for cases where a guaranteed defect detection rate is at stake: For applications where usability is mission-critical it does not suffice to size the sample according to the point estimates of the CGF. Instead, confidence intervals have to be taken into account. Expressed as a statement of guarantee this appears like "the study will reveal 99% of existing defects with a *probability of 95%*". Determining the confidence interval for the outcome estimators again requires an appropriate stochastic process model.

3. DATA SETS

The following statistical approaches to deal with heterogeneity will be showcased with an analysis on five published data sets. Three of the data sets were previously used by Lewis to assess the adjustment terms for $\hat{p}$ for small sample sizes [10]: The two data sets MANTEL and SAVINGS stem from a publication assessing the performance of the Heuristic Evaluation [16]; the set MACERR is the result of a usability testing study [10]. The data set WC01 [26] is from a usability testing study that was conducted to verify the results of a previous assessment of the Heuristic Evaluation – thus, it is a special case in that most usability defects were already known (a so-called falsification testing study). Finally the data set UPI07 is from a small scale experiment where a novel inspection method (called Usability Pattern Inspection) was compared to the Heuristic Evaluation [20]. The results from both experimental conditions have been merged for the analysis here. This data set can be obtained from the accompanying website [19] or on request. Note, that in this experiment both groups of participants performed virtually equal, so session heterogeneity is not the first guess. For an overview on the process sizes, number of defects and average detection probability in the data sets see the table 1.

4. TESTING FOR HETEROGENEITY

Several authors have expressed their doubts that a homogeneous p, as assumed by the CGF model, is appropriate. However, no

Table 1. Usability evaluation data sets

Data Set	Type	p	sessions	defects	Ref
MACERR	User Test	.16	15	145	[10]
MANTEL	HE	.36	76	30	[16]
SAVINGS	HE	.26	34	44	[16]
UPI07	HE, UPI	.30	10	35	[20]
WC01	User Test	.43	12	16	[26]

studies have proven heterogeneity with proper inferential statistics so far. In the following, a simple statistical test for heterogeneity is developed and applied. It bases on certain properties of the binomial distribution.

4.1 Binomial Sums and Mixtures

Assumed there is a homogeneous p for all detection events the basic process can be regarded as a series of equal Bernoulli processes. Under this assumption the CGF model applies for predicting the outcome after a number of sessions. Another way to look at the data is to plot a data matrix with two dimensions – sessions and defects – and denote every successful detection as 1 and a missed defect as 0. Under homogeneous p the margin sums of the data matrix shall follow a binomial distribution. For the small example in figure 1 this applies to the number of times a defect is found $S^D \sim B(4,0.3)$ and the number of found defects in each session $S^I \sim B(5,0.3)$. The binomial distribution has a mean of $\mu = np$ and a variance of $\sigma^2 = np(1-p)$. Thus we can expect $\mu = 1.2$ and $\sigma^2 = 0.84$ for S^D. Indeed, observed values for S^D come close to this with $\hat{\mu} = 1.4$ and $\hat{\sigma}^2 = 0.8$.

Nielsen & Landauer followed this idea, but approximated the margin sums as a Poisson distribution. This is common practice, but problematic in this case: The Poisson distribution is only applicable as a limiting case of the binomial distribution with a very large series of independent Bernoulli experiments with a very low probability ($n \to \infty$, $p \to 0$). A rule of the thumb states that $n > 30$ and $p < .05$ as a precondition to approximate a binomial distribution (e.g. [23]). This is hardly fulfilled for the revised data sets where the probability of detecting a defect varies between .12 and .58 and the number of usability defects between 9 and 145. For the general case, the binomial distribution is the appropriate model of how often a defect is detected and how many defects each session reveals.

When discussing the Poisson model, Nielsen and Landauer argue that this model still holds under heterogeneity because of the Poisson distribution's additivity property, where the sum of two Poisson distributed random variables is again Poisson distributed. This is a weak argument because it does not allow for the conclusion, that the CGF prediction stays unaffected as well. And, in fact, it is completely different for the binomial distribution. Instead, if there is heterogeneity in defects, S^D becomes a *mixed* binomial distribution, whereas S^I is a *sum* of binomial variables.

As an example consider the case depicted in figure 2 with two sets of $n_1 = n_2 = 20$ defects with detection probabilities $p_1 = .2$ and $p_2 = .3$. These $n = n_1 + n_2 = 40$ defects are evaluated

	S_1	S_2	S_3	S_4	S^D
D_1	1	0	0	1	2
D_2	0	1	0	0	1
D_3	0	1	1	0	2
D_4	0	1	1	0	2
D_5	0	0	0	0	0
S^I	1	3	2	1	

Figure 1. Example data matrix from a usability evaluation study with five defects, four sessions and homogeneous probability $p = .3$. A successful defect detection event is denoted as 1.

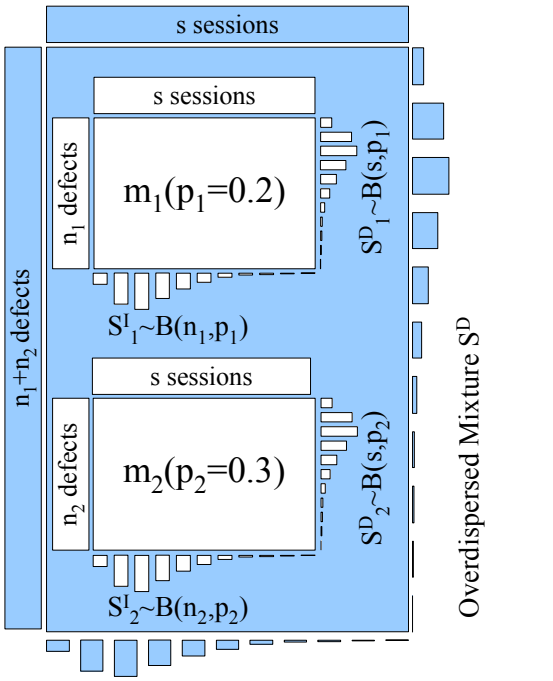

Figure 2. Sums and Mixtures of two subsets of defects with differing detectability

in $s = 30$ sessions. This results in two dichotomous response matrices $m_1[s \times n_1]$ and $m_2[s \times n_2]$, where the margin sums can be computed. For m_1 the number of defects found per session is binomial distributed with $S_1^I \sim B(n_1, p_1)$, analogously for m_2. Orthogonally, the number of times a defect detected in $m_1$is $S_1^D \sim B(s, p_1)$ and analogously in m_2. If we now combine the two matrices to $m[s \times n]$ with $n = n_1 + n_2$ and two new margins sums S^I and S^D, neither of these margin sums is binomial distributed. Whereas the average margin sum is still np, with $p = p_1/2 + p_2/2$, the variances from the the binomial distribution.

Instead, S^I (the number of defects found in each session) is now the *sum* of two binomial variables and thus underdistributed, which means that $Var(S^I) < np(1-p)$. Vice versa, S^D (the number of times a defect is detected) is a *mixed* binomial variable and thus overdistributed with $Var(S^D) > np(1-p)$. The same phenomena appear in the general case with more than two mixtures or sums. A general proof for sums of unequal binomial random variables (S^I in the example) being underdispersed towards the binomial distribution is given by Marshall and Olkin [13] (cited after [18]). The overdispersion of mixed binomial distributions (S^D in the example) is proven by Whitt ([25], cited after [18]).

4.2 A Simple Test on Overdispersion

Nielsen and Landauer restrict their model to the homogeneous case and assume "that all evaluations [...] find exactly the mean number of problems". Even in the homogenous case this is an oversimplification because the binomial process already imposes a variance of $np(1-p)$. The introduced properties of sums and mixtures of different binomial variables are the key to determine whether there is additional variance that does not stem from a homogeneous Bernoulli series. Establishing a statistical test specific for overdispersion towards the binomial distribution is straightforward by means of stochastic Monte-Carlo simulation: Testing the sessions for homogeneity from the matrix $m[s \times n]$ is done in the following steps:

1. Compute the observed variance $Var_{obs} = var(S^I)$

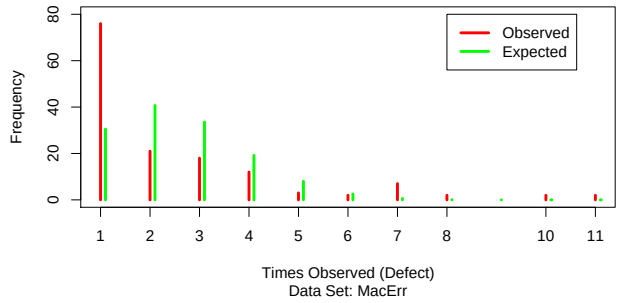

Figure 3. Overdispersion of defect margin sums in the MACERR data set

2. Estimate p from the matrix m as the relative amount of detection events
3. Generate s binomial random numbers $X_1 \ldots X_s \sim B(n,p)$
4. Compute the variance $V = Var(x_1 \ldots x_s)$
5. Repeat 3 and 4 a number of times (e.g. $r = 1000$), resulting in $v_{sim} = \{V_1 \ldots V_r\}$.
6. Count the elements of v_{sim} that are equal or larger than Var_{obs} and compute the relative frequency $\alpha = \frac{|V \in v_{sim}, V \geq Var_{obs}|}{r}$

The relative frequency α can be interpreted in the usual way under the null hypotheses: *The observed random variable var_{obs} has the variance to be expected under the binomial distribution*. Note, that in the case of both being heterogeneous – defects and sessions – there must appear a slight compensation of overdispersion by the underdispersive effect of the orthogonal binomial sums. But, as Rivest showed, the amount of underdispersion usually is much weaker than overdispersion [18]. In any case, there is no risk of overconfidence.

Figure 3 illustrates how this test works. It shows the observed frequency of how often a defect is detected versus the frequency expected by the binomial model (from the MACERR data set). It appears that there are too many defects only detected once and too many detected more than six times. This results in greater variance which is revealed by the overdispersion test.

4.3 Data Analysis

The Monte-Carlo test on overdispersion with respect to the binomial distribution was implemented as a program in R [17]. The program takes the dichotomous response matrix (like depicted in figure 1) of a usability evaluation as input and outputs the observed variance in the matrix Var_{obs}, the theoretical variance Var_{theo}, the mean of the variances generated with the Monte-Carlo experiment and the relative frequency α. If the absolute frequency of $V \geq Var_{obs}$ is zero, a maximum value for α is printed, based on the number of Monte-Carlo runs.

The five data sets introduced above have undergone the overdispersion test with $r = 10000$ Monte-Carlo runs in both directions: defects found per session and number of times a defect was found. A significant α ($\alpha \leq .05$) can be interpreted that there exists heterogeneity in the skills of inspectors, respectively the sensibility of participants in a usability testing study.

The results are shown in table 2. In four cases the observed variance was larger than both, theoretical and simulated variance. In three studies this appears as highly significant. In one case (UPI07) there still is a tendency. But, note that the observed variance is still twice as large as the theoretical and this was the smallest sample. Only in the WC01 study the observed variance is below the theoretical value. As explained above this may be an effect of underdispersion due to sums of binomial random variables. For verification an underdispersion test was conducted, by simply changing the last step in the procedure to $\alpha = \frac{|V \in v_{sim}, V \leq Var_{obs}|}{r}$. With a result of $\alpha = .37$ underdispersion could not be verified for the data set.

Table 2. Test on heterogeneity of session sensitivity

Data Set	Var_{obs}	Var_{theo}	Var_{sim}	$P(Var_{obs} \geq Var_{sim})$
MACERR	97.26	19.76	19.79	$< 0.0001^{***}$
MANTEL	13.25	7.05	7.04	$< 0.001^{***}$
SAVINGS	18.45	8.71	8.73	$< 0.001^{***}$
UPI07	13.39	7.35	7.40	$.06^{+}$
WC01	3.17	3.93	3.90	.64

Table 3. Test on heterogeneity of defect detectability

Data Set	Var_{obs}	Var_{theo}	Var_{sim}	$P(Var_{obs} \geq Var_{sim})$
MACERR	4.93	2.04	2.04	$< 0.0001^{***}$
MANTEL	546.15	17.86	17.86	$< 0.0001^{***}$
SAVINGS	51.77	6.73	6.68	$< 0.0001^{***}$
UPI07	3.80	2.10	2.11	0.0015^{***}
WC01	19.63	2.94	2.96	$< 0.0001^{***}$

Table 3 shows the results of the overdispersion tests regarding defect heterogeneity. In all five studies the observed variance of margin sums is highly significant beyond the expected variance under the binomial model.

4.4 Discussion

The overdispersion test is based on a simple procedure, is easy to conduct and interpret, but still a powerful tool to reveal heterogeneity – at least in mid to large size evaluation studies.

When conducting usability evaluations, one definitely has to account for heterogeneous detectability of defects. This appeared in all five data sets, regardless of the evaluation method employed or other contextual factors.

In most cases one also has to expect session heterogeneity. An exception is the WC01 study. This is interesting for two reasons: First, this was a falsification study, where the aim is to verify previously proposed defects. This most likely results in a very systematic design of testing tasks, which may in some way equalize the participants sensitivity. Second, the original authors argued that the high variance in process outcome (with subsets of three participants) is due to variance in users [26]. They were wrong as the overdispersion test shows.

5. DETERMINING HETEROGENEITY

In the previous section the impact of heterogeneity on the margin sums of the response matrix were introduced with a discrete mixture of binomial variables. The overdispersion test can be used to easily detect heterogeneity. But, a further aim is to determine the amount of heterogeneity in the sample. For example, one may want to analyse, whether a certain method has the wanted effect of equalizing the process – by making some intractable defects easier or help novices to catch up with experts. Another purpose in evaluation method research is to compare different studies. Clearly, the results from two evaluation studies (or conditions in a strictly experimental study) can at best be interpreted when there are approximately equal conditions – determining the variance of impact factors is one such criterion.

5.1 Fitting the beta-binomial distribution

In general, estimating the variance of the heterogeneity factor requires to assume a certain distribution of p. As p is a probability, a distribution must range in $[0,1]$. In such cases often the beta distribution is employed: It has exactly the required range and comes with two parameters allowing it to take a variety of shapes with deliberate mean and variance.

If we assume $p \sim Beta(a,b)$ and undertake a series of n Bernoulli experiments with p, than the sum of results is beta-binomial distributed $S^I \sim BetaBin(a,b,n)$ with mean and variance

$$\mu = \frac{na}{a+b} \tag{4}$$

$$\sigma^2 = \frac{nab(n+a+b)}{(a+b)^2(1+a+b)} \tag{5}$$

Estimating the parameters a and b from the margin sum is probably not an everyday statistical technique, but a few programs exist with an implementation. In the following the VGAM package [27] for the statistical computing environment R [17] serves for estimating the parameters. This package allows estimating the parameters for a large variety of distributions with the maximum likelihood (ML) method. The ML method identifies the set of parameters $x_1 \ldots x_n$ for a certain distribution that are most likely given the observed data D. For some types of distributions there exists a symbolical solution (e.g. μ and σ^2 of the normal distribution), but in most cases numerical algorithms have to be used to find the maximum of the likelihood function $L(x_1 \ldots x_k|D)$.

There are two restrictions of the beta-binomial model: First, it relies on a single margin sum vector and thus is only capable of capturing one heterogeneity factor at the time. Second, the beta-binomial model is not appropriate in the case of underdispersion. As was explained in section 4.1, slight underdispersion may appear from the sum of binomial random variables, but this is usually overcompensated by the overdispersive mixture. Consequently, if both factors are mixtures a slight underestimation of the factor's variance may appear. There are critical cases, where underdispersion may happen: First, one factor is purely binomial distributed and the other is a mixture. Underdispersion may also arise, when there is stochastic dependence between defects or sessions. This may happen in usability testing studies, if a severe defect causes a test person to give up early and in effect "shields" other defects later in the task flow. Stochastic dependencies between sessions may appear, if a previously identified defect is not recorded on later occurances; obviously this can be avoided by holding the sessions strictly independent. Fortunately, this is common practice in industrial and research applications.

In presence of underdispersion it is not even possible to estimate the parameters of the underlying beta distribution as this would require the variance of the prior beta distribution to be negative, which cannot happen with any real-valued distribution. Accordingly, the ML estimation will stop at unreasonable high values for a and b and throw an error. It is therefore recommended to first check for over- and underdispersion with the statistical test introduced above.

5.2 Model selection criteria

The previous section 4.2 introduced a test that employs a simple frequentist approach to test for a margin sum having the variance expected under the binomial model. But, it is a stronger claim that the beta-binomial model is a better approximation of the evaluation process than the binomial model. In order to decide between competing models regarding their match to the data, an appropriate selection criterion is needed. Because model selection is a statistical approach rarely found in the HCI literature [4], this will briefly being introduced in the following.

Two straightforward selection criteria are the residual deviance (e.g. residual sum of squares, RSS) after applying the model or the value of the maximized likelihood function. The better of two models would have a smaller residual deviance and the larger likelihood. However this may ignore an important directive for scientific reasoning – Occam's Razor, which demands that the more parsimonous of two theories must be preferred if it has the same explanation power. In statistical reasoning parsimony applies as the numbers of parameters of two competing models. Obviously, the more parameters a model has, the more versatile it is in taking the shape of the observed data. Consequently, a proper criterion for model selection balances goodness-of-fit and parsimony of models.

Table 4. Estimated beta-binomial parameters for session margin sums

Data Set	a	b	Var
MACERR	5.25	26.57	.004
MANTEL	12.57	20.73	.006
SAVINGS	11.76	31.20	.005
UPI07	15.61	36.42	.004
WC01	nc	nc	nc

Several so called *information criteria* put a penalty on the number of parameters and thus allow for proper model selection respecting the directive of parsimony. One of the widest known is the Akaike Information Criterion (AIC), which has a deep foundation in mathematical information theory [3]. It is amazingly simple to apply after an ML estimation of a model's parameters (or the special case of the least squares method with normally distributed errors). Usually, one adds a further correction term for the sample sizes to the *AIC*. The corrected AIC_c computes as follows with the maximized likelihood function L_{max}, k model parameters and n observations:

$$AIC_c = -2\log L_{max}(x_1 \ldots x_k) + 2k + \frac{2k(k+1)}{n-k-1} \tag{6}$$

As the AIC_c grows with the number of parameters the model with the lowest value fits the data best with respect to parsimony. The *AIC* does not assume that one of the model candidates is the true model (which is assumed by the Bayes Information Criterion, see [3] for a thorough discussion). This can in most cases be regarded as a realistic and favorable approach. In the case of modelling evaluation process data there is considerable interest in finding a better alternative to the binomial model, but choosing the beta distribution is more or less a pragmatic than a well founded decision. Another issue with interpreting the *AIC* is that values of $\log L_{max}$ are usually very large, making the difference between two *AIC*s appear quite small. This may tempt the naive conclusion, that the models do not differ much in explanation power. In fact, this is not a problem as the relative difference does not affect the selection of the best model [3].

5.3 Data Analysis

It was shown so far that in virtually all situations heterogeneity due to defect detectability arises. In most cases one has also to account for session heterogeneity. Consequently, it is to expect that in most cases the beta-binomial model will fit the data better than the binomial model. Once the two parameters a and b have been estimated, the variance of the underlying beta distribution follows equation 5. This may serve as a measure of how large the heterogeneity is.

Beta-binomial parameters are being estimated via the ML method (provided by the VGAM package for R [27]). This procedure also yields the maximized likelihood function, which enables calculating the AIC_c. Consequently, a likewise estimation of the binomial parameter p in order to select the better model with the smaller AIC_c complements the parameter estimation. Usually, the estimated p is exactly the probability of detection events as in table 1 and is not reported again.

As the WC01 has an observed variance lower than the binomial for sessions, this data set is not analysed, because this would result in unreasonable values (virtually $b \to \infty$) which may even produce program errors. As table 4 shows, the estimation proce-

Table 5. Model selection for session margin sums

Data Set	binomial		beta-binomial	
	$\log L$	AIC_c	$\log L$	AIC_c
WC01	-131.32	265.04	nc	nc
MACERR	-966.17	1934.65	-948.20	*1901.41*
MANTEL	-1511.40	3024.85	-1502.13	*3008.42*
SAVINGS	-875.60	1753.33	-870.07	*1744.52*
UPI07	-213.80	*430.11*	-213.04	431.80

Table 6. Estimated beta-binomial parameters for defect margin sums

Data Set	a	b	Var
MACERR	2.37	12.00	.009
MANTEL	.79	1.21	.080
SAVINGS	1.18	3.15	.037
UPI07	3.62	8.41	.016
WC01	.62	.65	.110

Table 7. Model selection for defect margin sums

Data Set	binomial		beta-binomial	
	$\log L$	AIC_c	$\log L$	AIC_c
MACERR	-966.17	1934.37	-937.42	*1878.93*
MANTEL	-1511.40	3024.93	-1083.99	*2172.43*
SAVINGS	-875.60	1753.30	-770.11	*1544.52*
UPI07	-213.80	429.73	-210.11	*424.59*
WC01	-131.32	264.92	-95.31	*195.55*

dure yields reasonable values for a and b and the variance in the remaining four data sets ranges from .004 to .006.

For model selection the AIC_c is computed like introduced above. Table 5 shows that, as expected, the beta-binomial model is usually to be preferred to the binomial model. Only for the UPI07 data set and, most likely, WC01 the binomial model fits better, which is consistent to the results from the overdispersion test.

Table 6 shows the parameters of the beta-binomial model on defects. In all cases the estimation produced reasonable values without errors. The variance ranges from .009 to .110.

As can be obtained the AIC_c values in table 7, for all data sets one has to prefer the beta-binomial model to the binomial model.

5.4 Discussion

The results from estimating the beta-binomial parameters clearly confirm the previous results from the overdispersion test. The values for beta-binomial variance suggest that session heterogeneity is quite comparable between the studies (except the WC01 data set, of course), whereas there appears a larger range for defects.

In most cases, the beta-binomial model is preferred due to a smaller value of AIC_c. In two cases, however, the binomial model fits better: session heterogeneity in the data sets WC01 and UPI07 are better explained with a binomial model. Again, this is expected as no significant overdispersion could previously be revealed. But, note that for the UPI07 the two $\log L$ values are virtually equal for both cases. Thus, it is no harm to apply the beta-binomial model right from the start as it will always fit equally good or better than the binomial model. It is, though, required to check for the rare case of underdispersion first, as this will give useless results, if any.

6. PROCESS PREDICTION UNDER HETEROGENEITY

The main argument of Nielsen and Landauer is to estimate the process outcome given the process size n and probability p with the CGF introduced by Virzi. Their aim is to predict the process from the data of the process itself and they suggest a curve fitting procedure using the number of previously undetected defects in each step. Based on the additive property of Poisson distributed variables it was believed that the CGF approach is robust to heterogeneity in the process. As the next section will show, this is not the case.

6.1 Biases in estimators for p

The curve fitting procedure of estimating p from previous process data has later been replicated by Lewis [10] and found to be less efficient than his approach of directly estimating p from the number of defects detected so far n_d and the average number of detection events per session $\overline{e_d}$.

$$\hat{p} = \overline{e_d}/n_d \tag{7}$$

This idea originated in an earlier publication [9], but later Lewis undertook a major revision [10]. With small samples (in the early phase of a usability evaluation process) the naive estimator of p as the average proportion of defects found by each person is biased towards overestimation. This is caused by still undetected defects not been taken into account. In his revision Lewis compared several correction terms in application to real data sets. The final suggestion was an equally weighted combination of a simplified Good-Turing (GT) adjustment and a normalization procedure (NORM) proposed by Hertzum and Jacobsen [8] (see equation (8)). The GT adjusts for still unseen defects by taking into account the number of defects detected exactly once n_1. NORM accounts for a possible overestimation due to the small sample size in the beginning of the process. The so adjusted estimator for p writes as

$$\hat{p}_{GT-Norm} = \frac{(\hat{p}-\frac{1}{n})(1-\frac{1}{n})}{2} + \frac{\hat{p}}{2(1+n_1/n_d)} \tag{8}$$

It was shown with several real data sets that $\hat{p}_{GT-Norm}$ yields the best estimation of p for small process sizes. For the moment, this seems to be the choice for practitioners to control their processes. However, from a more theoretical point of view it does not fully convince. The particular combination of both adjustment terms arose empirically, but is not theoretically justified by a consistent model of the process, in other words: it is a more or less deliberate choice. This is problematic, because the properties of the estimator $\hat{p}_{GT-Norm}$ are mostly unknown. Thus, it is not clear, how robust the estimator is to violations (esp. heterogeneity) and how it behaves with more extreme parameters (e.g. very high number of defects or very low p). In fact, Lewis already observed an increasing underestimation for larger process sizes ($n > 6$). Industrial evaluation studies are regularly conducted with several tens of sessions, so this is clearly an annoying effect.

6.2 Fitting the cumulative beta-geometric function

A principal problem of all estimators for p is, that they do not take the heterogeneity of the process into account. In general, it seems to be quietly assumed, that the CGF is only sensitive to the expected value of p, but is not biased by variance between defects. This is not the case; instead the function of process size on process outcome is different if defect heterogeneity is present: The process outcome approaches the asymptote at 1 more slowly. This effect is depicted in figure 4 for three mixed distributions with the same average p but different discrete distributions underlying p.

Eventually, this is why Lewis' GT adjusted estimator failed to predict the process properly . If defect heterogeneity is indeed the source of bias in the CGF, then the cumulative beta-geometric function (CBGF) should fit the data better than the CGF. The procedure to test this is straightforward from what was introduced so far: p for the CGF and a and b for the CBGF are estimated

Table 8. Comparison of cumulative beta-geometric vs. cumulative geometric model fit: Residual Sum of Squares

Data Set	RSS_{CGF}	RSS_{CBGF}
MacErr	*.023*	.081
Mantel	.416	*.077*
Savings	.330	*.003*
Upi07	*.010*	.013
WC01	.173	*.063*

from the margin sum S^D. The empirical curve of process outcome is generated by a Monte-Carlo procedure. Both theoretical functions – CGF and CBGF – are then plotted against the empirical function and the deviance computes as residual sums of squares (RSS). Estimating the beta-geometric parameters of Monte-Carlo sampled data with the VGAM package usually produced program errors instead of usable values. Therefore, the following analysis will not include a model selection step by *AIC*, but compares the RSS alone, which is not perfect with respect to model parsimony. In any case, process prediction is largely a practical problem, where model fit (i.e. minimizing the deviance) counts more than scientific truth.

6.3 Data Analysis

The previous analysis dealt with showing the existence and determining the size of heterogeneity from the response matrix. It is still an open question how heterogeneity in the data sets impacts the process outcome, respectively how it differs from the outcome predicted by the CGF. It has already been demonstrated that with defect heterogeneity the curve flattens more rapidly with larger process sizes. To show this effect on the data sets, three functions will be drawn and compared:

1. The CGF with p estimated from the relative frequency of detection in the response matrix. Note, that this is a post-hoc analysis on the whole response matrix, so there is no need for a small sample adjustment like GT-NORM here.
2. The cumulative beta-geometric function with a and b estimated with the beta-binomial procedure (see table 6)
3. The average "true" process outcome estimated with Monte-Carlo sampling

Table 8 shows the CBGF to fit better for the data sets MANTEL, SAVINGS and WC01. In the remaining two studies (MACERR, UPI07) the CGF has the lower deviation from the observed data.

Next, the graphs for the two process models are plotted and visually compared. Figure 5 shows those three models where the CBGF fitted better. A first observation is that the CGF predicts a steeper progress for small process sizes. It generally appears closer to a rectangular shape. In all three cases the CBGF curve is much closer to the observed process outcome for the first third

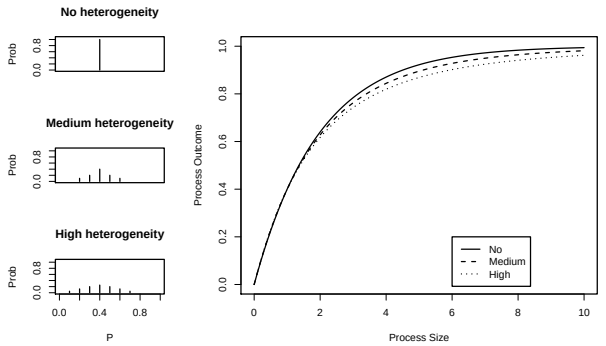

Figure 4. Impact of defect heterogeneity on process outcome

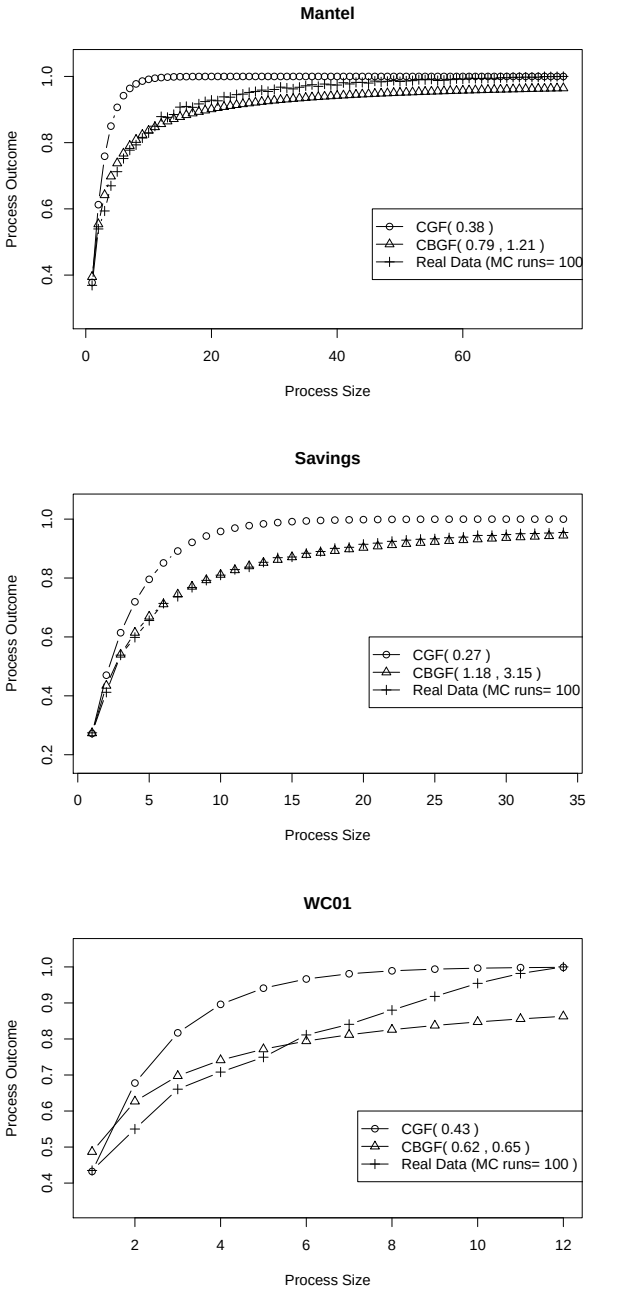

Figure 5. Graphical comparison of process outcome models for studies with $RSS_{geom} > RSS_{betageom}$

of process size, i.e. the steep part. In the MANTEL and WC01 studies the "asymptotical" last part of the process is closer to the CGF curve. Solely for the SAVINGS study a close-to-perfect match of the beta-geometric model is given.

Figure 6 shows the same comparison for the two studies where the CGF provided a slightly better fit. As expected, all three curves are quite close to each other. Especially, there is much less difference in the "steep" part of the process. Interestingly, both models understimate the process outcome for the last few sessions of the studies, with a slightly better performance of the CGF.

6.4 Discussion

The comparison of process outcome models gives a mixed picture. In three cases the CBGF fits the data better than the geo-

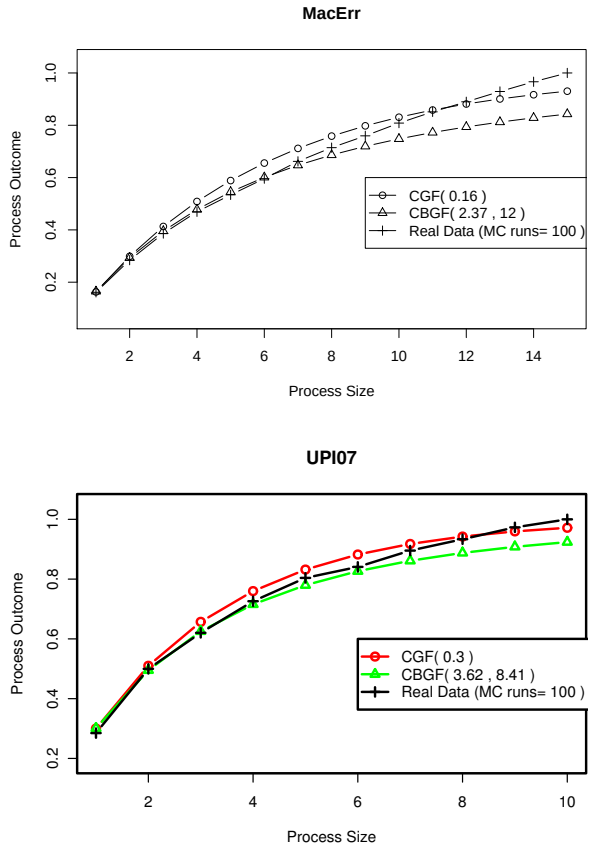

Figure 6. Graphical comparison of process outcome models for studies with $RSS_{geom} < RSS_{betageom}$

metric model. Two further observations can be made: First, for MACERR and UPI07 the CGF fits better, but the differences in RSS are comparably small. Second, these both studies showed the smallest variance in defects detectability (cf. table 6). This is confirmed by the graphical comparison: When variance of detectability is low, the process predictions are quite similar. When variance gets larger the CBGF is in favor, but mainly for the early process.

A final clarification of these mixed findings is out of reach at the moment, but some speculations can be given: The deviance in the "asymptotic" part of the process may be due to a limitation of the Monte-Carlo procedure. In fact, there is no longer fine-grained sampling, when the maximum process size is reached. For example, there are only two samples left for $s-1$. This makes the smallest possible steps larger than at the beginning of the process. By the way, this is a similar problem, which made Lewis introduce the NORM term for small process sizes. Also, the models did not include the variance of sessions. It may be, that this causes a further bias on the process outcome model, which impacts with larger process sizes.

Be aware, that the post-hoc approach presented here is not applicable for controlling an industrial process. But, in hope that future research comes up with online prediction models accounting for heterogeneity, some practical advice can already be given: In general, if finding defects is mission-critical it is safe to account for defect heterogeneity. In the worst case this leads to a slight underestimation of current process outcome, but there is no harm from an optimistic bias. In all less ambitious studies, where the process goal is around 80% or smaller, it is highly recommended to apply a heterogeneous prediction model as this performs at least equal and better in most cases.

A special caveat on using prediction models is revealed by a further look at the data set UPI07. The authors reported, that the experiment was held under tight time constraints with a limited set of user tasks [20]. They have subsequently conducted a usability testing study in order to validate the defects via falsification testing. In this study they found a larger number of defects not detected by the ten inspection sessions. But, the incomplete data set neatly complies to the CGF prediction model and asymptotically reaches the upper bound. This is well in line with the findings of Lindgaard and Chattratichart who re-analysed the CUE-4 data sets [14] and found that the effectiveness of a usability study was largely affected by task desgn and task coverage [12]. It follows, that any prediction model will not uncover the fact that some defects have not been found due to flawed study design. In other words: Usability practitioners are not freed from their liability to carefully compile relevant user tasks and assign sufficient ressources to their studies.

7. GENERAL DISCUSSION

Several researchers have already guessed it and the results reported here confirm it: Heterogeneity regularly appears in both factors – sessions and defects – of usability evaluation studies. Defects differ in their detectability in all five studies analysed here. In most cases one also has to expect sessions to differ in the detection capability, regardless of usability tests or expert evaluations are being conducted.

What has largely been overlooked in research is that heterogeneity has considerable impact on the prediction of process outcome. Defect heterogeneity causes a lower process outcome than to be expected under the CGF model. For industrial usability studies the risk arises to stop the process to early. As an example, have a look at prediction of the SAVINGS data set, where a total of 44 defects was observed (see middle graph in figure 5). Assumed, that it was decided to stop the process when 80% of defects are detected. The CGF predicts this being the case with the "magic" number of five evaluators. But, only 67% are detected at this stage, which means that only 30 instead of 35 defects are revealed. In fact, the double process size of 10 is required to reach the 80% goal. The bias of the CGF becomes largest with a process size of 9. Here, the CGF predicts 94%, whereas the observed rate is only 78% with an absolut difference of around 7 defects. Obviously, underestimations in this magnitude are very harmful for large, mission-critical studies. In comparison the CBGF accounts for defect heterogeneity and predicts this data set very well with a maximum deviation of only one defect.

Another issue rarely being adressed in the current literature is the variance of outcome at different points in the process. This was not the major topic here, but, it is known from an earlier simulation experiment that session heterogeneity causes additional variance of outcome [22]. Consequently, when an evaluation study has a strict goal an extra safety margin is required.

Determining the size of this safety margin is beyond the scope of this work. But, the methodological approach introduced here is the way to go: Statistical model selection provides powerful means to determine the underlying stochastic process (or at least the best known approximation). An appropriate statistical model may serve both: point estimators and variance for process outcome. For simple models (which are unlikely to be good approximations) confidence intervals can easily be obtained from general textbooks on statistics. For more complex models Monte-Carlo simulations provide a flexible alternative.

The beta-geometric model suggested here is an advance regarding better point estimators. But, in order to also determine the confidence interval for the estimators a model is required that includes both impact factors – session and defect heterogeneity – at the same time. This is the case with the Rasch model (see section 2). But this approach requires large sample sizes for pa-

rameter estimation and is thus not applicable for extrapolating the process from early sessions.

Another good starting point are the so-called capture-recapture models, which have their origins in biostatistics (e.g. [6]). Capture-recapture models address a problem analogous to the defect detection process: The number of animals living in an area has to be estimated without counting them all. There already exists a variety of statistical models, some of them allow for time effects (comparable to session heterogeneity) and animal heterogeneity (comparable to defect heterogeneity). In fact, these have successfully been applied to software inspection processes in Software Engineering (e.g. [2]). Work on evaluating and adapting capture-recapture models for usability evaluation processes is currently in progress.

Still, a way to go is to use the beta-geometric model for the planning of studies (i.e. at time of project negotiation). A usability service company may have tens or hundreds of data sets from past projects. These can be fed into the beta-binomial estimation procedure to derive the heterogeneity measures. Stored into a database these can act as a priori estimators for to-be-planned studies. Probably, the value will increase further when data sets are classified according to some relevant factors. Common statistical techniques like ANOVA or factor analysis may reveal these factors which, and at the same time may give further insights into the origins of heterogeneity.

8. CONCLUSION

Heterogeneity in usability evaluations is a fact. This puts forth several open research questions:

- How large is heterogeneity in both factors under various conditions?
- What causes heterogeneity?
- Is it a continuous variable or are there distinct classes of usability inspectors (test users) and defects?
- What kind of model can replace the CGF in order to extrapolate the process?

Heterogeneity puts harmful bias on the well known CGF predictor. As a result, usability practitioners take the risk of stopping the process far too early. An approved solution for reliable process prediction is currently work-in-progress. Until then, advice to practitioners running usability studies can be given as follows:

- Run the test on binomial overdispersion on your past data sets to convince yourself of heterogeneity.
- Pay special attention to defect heterogeneity (which is likely to occur), as this leads to harmful overestimation of outcome.
- For small studies or at the beginning of a larger study ($n < 6$) apply the CGF model with the GT-NORM estimator suggested by Lewis [11].
- When process size increases ($n > 10$), estimate the beta-binomial parameters and feed them to the CBGF. Repeat this when new data points arrive and control your process towards the targeted goal.
- Make sure to always have a generous safety margin when usability is mission-critical. Expect up to 17% underestimation of outcome with the CGF. Have a look at [21, 7] to get an idea of the random variation at different process sizes.
- Keep in mind that predicted outcome is only reliable for well-designed studies. Pay attention to complete coverage and appropriate tasks.

Also, practicioners are encouraged to store their project data and begin to use it for planning of studies. Possibly, there will once appear a mature approach of *experienced-based sample size prediction* which makes use of past data sets and statistical models to give a best guess on required process size.

Some programs for analysing evaluation process data are available online [19] or on request. The author is willing to further assist and cooperate.

9. ACKNOWLEDGEMENTS

Work on this paper was made possible by a stipend of the Passau Graduate School of Business and Economics and generous support of Chair Prof. Dr. Franz Lehner, Passau University.

Thanks to all authors who have published their complete data sets [10, 16, 26]. Find another one at [19].

10. REFERENCES

[1] B. W. Boehm and V. R. Basili. Software defect reduction top 10 list. *IEEE Computer*, 34(1):135–137, 2001.

[2] L. C. Briand, K. El Emam, and O. Freimut, B. G.and Laitenberger. A comprehensive evaluation of capture-recapture models for estimating software defect content. *IEEE Transactions on Software Engineering*, 26(6):518–540, 2000.

[3] K. P. Burnham and D. R. Anderson. Multimodel inference. understanding AIC and BIC in model selection. *Sociological Methods & Research*, 33(2):261–304, 2004.

[4] P. Cairns. HCI...not as it should be: Inferential statistics in HCI research. In L. J. Ball, M. A. Sasse, C. Sas, T. C. Ormerod, A. Dix, P. Bagnall, and T. McEwan, editors, *Proceedings of the HCI 2007*, volume 1 of *People and Computers*, pages 195–201. British Computing Society, 2007.

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

[6] A. Chao. An overview of closed capture-recapture models. *Journal of Agricultural, Biological, and Environmental Statistics*, 6(2):158–175, 2001.

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

[8] M. Hertzum and N. E. Jacobsen. The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction*, 13(4):421–443, 2001.

[9] J. R. Lewis. Sample sizes for usability studies: Additional considerations. *Human Factors*, 36:368–378, 1994.

[10] J. R. Lewis. Evaluation of procedures for adjusting problem-discovery rates estimated from small samples. *International Journal of Human-Computer Interaction*, 13(4):445–479, 2001.

[11] J. R. Lewis. Sample sizes for usability tests: Mostly math, not magic. *Interactions*, 13(6):29–33, 2006.

[12] G. Lindgaard and J. 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.

[13] A. W. Marshall and I. Olkin. *Inequalities: Theory of Majorization and Its Applications*. Academic Press, New York, 1979.

[14] R. Molich and R. Jeffries. Comparative expert review. In *Proceedings of the CHI 2003, Extended Abstracts*, pages 1060 – 1061. ACM Press, 2003.

[15] J. Nielsen and T. 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.

[16] J. Nielsen and R. Molich. Heuristic evaluation of user interfaces. In *Proceedings of the CHI 1990*, 1990.

[17] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2006. ISBN 3-900051-07-0.

[18] L.-P. Rivest. Why a time effect often has a limited impact on capture-recapture estimates in closed populations. *The Canadian Journal of Statistics*, 35(4), 2007. in print.

[19] M. Schmettow. Heterogenity in the usability evaluation process - accompaning website. Website, January 2008. http://schmettow.info/Heterogeneity.

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

[21] M. Schmettow and W. Vietze. Introducing item response theory for measuring usability inspection processes. submitted to CHI2008, September 2007.

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

[23] S. S. Shapiro and A. J. Gross. *Statistical Modeling Techniques*. Marcel Decker, New York, 1981.

[24] R. A. Virzi. Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, 34(4):457–468, 1992.

[25] W. Whitt. Uniform conditional variability ordering of probability distributions. *Journal of Applied Probability*, 22:619–633, 1985.

[26] A. Woolrych and G. Cockton. Why and when five test users aren't enough. In J. Vanderdonckt, A. Blandford, and A. Derycke, editors, *Proceedings of IHM-HCI 2001 Conference*, volume 2, pages 105–108. Cepadeus, Toulouse, France, 2001.

[27] T. W. Yee. *VGAM: Vector Generalized Linear and Additive Models*, 2007. R package version 0.7-5.