Estimating the number of usability problems affecting medical devices: modelling the discovery matrix

Vincent Vandewalle
Inria Centre de recherche Lille Nord Europe
Alexandre Caron (alexandre.caron2@univ-lille.fr)
Univ. Lille, CHU Lille, EA2694, F-59000 Lille, France https://orcid.org/0000-0002-9872-0633
Coralie Delettrez
CHU LILLE
Sylvia Pelayo
Centre d'Investigation Clinique - Innovation Technologique Lille
Alain Duhamel
Universite de Lille
Benoit Dervaux
Universite de Lille

Research article

Keywords: usability study, medical device, missing data, Bayesian statistics, maximum likelihood

Posted Date: November 11th, 2019

DOI: https://doi.org/10.21203/rs.2.16958/v1

Version of Record: A version of this preprint was published on September 18th, 2020. See the published version at https://doi.org/10.1186/s12874-020-01091-y.

Title Page

Title of the manuscript: Estimating the number of usability problems affecting medical devices: modelling the discovery matrix

Author listing:

Vincent Vandewalle[a,b*], PhD, **Alexandre Caron**[a*], MD, MSc, **Coralie Delettrez**[c], MSc, **Sylvia Pelayo**[a,d], PhD, **Alain Duhamel**[a], PhD, **Benoit Dervaux**[a], PhD.

*Equal contributions

Affiliations:

- [a]Univ. Lille, CHU Lille, EA 2694 - Santé publique : épidémiologie et qualité des soins, F-59000 Lille, France,
- [b]Inria, F-59000 Lille, France,
- [c]CHU Lille, Direction de la Recherche et de l'Innovation, F-59000 Lille, France
- [d]CIC-IT/Evalab 1403, CHU Lille, F-59000 Lille, France.

Corresponding Author: Alexandre Caron (alexandre.caron2@univ-lille.fr)

Word Count: 6411

Keywords: usability study, medical device, missing data, Bayesian statistics, maximum likelihood

Declaration

Ethics approval and consent to participate: Not applicable.

Consent for publication: Not applicable.

Availability of data and materials: The data generated and analysed during the current study are included in this published article and its supplementary information files.

Competing interests: The authors declare that they have no competing interests.

Funding source: This research was funded by the Swiss National Science Foundation (grant number: SNSF-164279) and the French Agence Nationale de la Recherche (grant number: ANR-15-CE36-0007). The funding bodies had no role in the design of the study and collections, analysis, and interpretation of data, and in writing the manuscript.

Contributors' Statement: AC, BD and VV conceptualized and designed the study. AC, CD and VV carried out the analysis. AC, AD, BD, CD, SP and VV contributed to the interpretation of the results. AC and VV drafted the initial manuscript. AD, BD, CD and SP critically reviewed and revised the manuscript. All authors approved the final manuscript as submitted and have agreed both to be personally accountable for the author's own contributions and to ensure that questions related to the accuracy or integrity of any part of the work, even ones in which the author was not personally involved, are appropriately investigated, resolved, and the resolution documented in the literature.

Abstract

Background. Usability studies of medical devices are mandatory for market access. The studies' goal is to identify and eliminate usability problems that could cause harm the user or limit the device's effectiveness. In practice, human factor engineers study participants under actual conditions of use and list the problems encountered. This results in a binary discovery matrix in which each row corresponds to a participant, and each column corresponds to a usability problem. One of the main challenges in usability studies is estimating the total number of problems, in order to assess the completeness of the discovery process. Today's margin-based methods fit the column sums to a binomial model of problem detection. However, the discovery matrix actually observed is truncated because of undiscovered problems, which corresponds to fitting the marginal sums without the zeros. Margin-based methods fail to overcome the bias related to truncation of the matrix. The objective of the present study was to develop and test a matrix-based method for estimating the total number of usability problems.

Methods. The matrix-based method models the likelihood of the discovery matrix, and allows one to account for all the available information. It also circumvents a drawback of margin-based methods by simultaneously estimating two unknown parameters: the probability of problem detection and the total number of problems. Furthermore, the matrix-based method takes account of a heterogeneous probability of detection, which better reflects a real-life setting. As suggested in the usability literature, a logit-normal prior for the probability of detection is selected.

Results. We assessed the matrix-based method's performance in a range of settings reflecting real-life usability studies and with both homogeneous and heterogeneous probabilities of problem detection. In our simulations, the matrix-based method

improved the estimation of the number of problems (in terms of bias, consistency, and coverage probability of the confidence interval) in a wide range of settings. We also applied our method to real data from a usability study of infusion pumps.

Conclusions. Our method should be applied by regulators and device manufacturers to estimate the number of usability problems using the set of statistical routines provided.

Main manuscript text

I. Background

A. Introduction

The usability study is a cornerstone of medical device development, and proof of usability is mandatory for market access in both the European Union and the United States [1]. The overall objective of a usability assessment is to ensure that a medical device is designed and optimized for use by the intended users in the environment in which the device is likely to be used [2]. The goal is to identify and then eliminate problems (called "use errors") that could cause harm the user or impair medical treatment (e.g. an inappropriate number of inhalations, finger injection with an adrenaline pen, etc.) [3]. The detection of usability problems must be as comprehensive as possible because medical devices are safety-critical systems. However, the total number of usability problems is never known in advance. The main challenge during the usability study is thus to estimate this number, in order to assess the completeness of the problem discovery process.

In practice, participants are placed under actual conditions of use (real or simulated), and usability problems are observed and listed by human factor engineers. The experimental conditions are defined in a risk analysis that gathers together possible usability problems. Throughout the usability study, problems are discovered and added to a discovery matrix - a binary matrix with the participants as the rows and the problems as the columns. The current approach involves estimating the total number of problems as the usability study progresses, starting from the first sessions. The number is estimated iteratively as the sample size increases, until the objective of completeness has been achieved [4].

From a statistical perspective, the current estimation procedure is based on a model of how the usability problems are detected; this is considered to be a binomial process. The literature suggests that the total number of usability problems can be estimated from the discovery matrix's problem margin (the sum of the columns) [5–9]. However, this estimation is complicated by (i) the small sample size usually encountered in usability studies of medical devices [10] and (ii) as-yet unobserved problems that truncate the margin and bias estimates [11–13].

The objective of the present study was to develop a matrix-based estimation of the number of usability problems affecting a medical device. This new method is based on the likelihood of the discovery matrix (rather than the matrix's margins alone), so as to avoid the loss of associated information.

B. Data collected during the usability study: the discovery matrix

The human factor engineer collects the results of the usability study in a problem-discovery matrix $\mathbb{d}$. Each row corresponds to a participant, and each column corresponds to a usability problem. The result is 1 if the participant discovered the problem and 0 if not. Considering that after the inclusion of n participants, j problems have been discovered, a $n \times j$ matrix is built. By way of an example, the discovery matrix obtained after $n = 8$ participants (in rows) might be the one presented below:

$$\mathbb{d} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

In this example, $j = 10$ different problems (in columns) have been detected so far. The first participant discovered only one problem (column 1), whereas the second discovered two new problems (columns 2 and 3), etc.

At this stage, some problems might not have been detected, and the total number of usability problems (m) is unknown. It should be noted that by definition, $m \geq j$ and $m - j$ problems remain undetected. Indeed, $\mathbb{d}$ comes from a complete but unobserved matrix of dimensions $n \times m$. This matrix is denoted as $\mathbb{x}$. Thus, the "observed" matrix $\mathbb{d}$ is a truncated version of the "complete" matrix $\mathbb{x}$; it lacks the columns corresponding to the as-yet undetected problems. Hereafter, we use the following notation: $\mathbb{x} = (x_{il})_{1 \leq i \leq n, 1 \leq l \leq m}$ where $x_{il} = 1$ if the participant i experiences the problem l, and $x_{il} = 0$ otherwise.

$$\mathbb{x} = \begin{pmatrix} x_{11} & \cdots & x_{1l} & \cdots & x_{1m} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & \cdots & x_{il} & \cdots & x_{im} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nl} & \cdots & x_{nm} \end{pmatrix}$$

The human factor engineer's goal is to estimate the total number of problems m from the discovery matrix $\mathbb{d}$ and thus deduce $m - j$ - the number of problems that have not been detected. The new method presented below addresses this goal.

C. Conventional estimation of m using a margin-based probabilistic model

In this section, we describe the margin-based methods currently employed to estimate the number of usability problems. As mentioned above, m is currently estimated by fitting a probabilistic (binomial) model to the discovery matrix's problems margin. More specifically, the probability with which a given usability problem is discovered by a

participant is modelled by a Bernoulli trial with a probability of success (i.e. detection) p. For a given problem, the Bernoulli trial is considered to apply independently to each of the n participants in the usability study. Thus, the problem margin sums can be considered as an independent, identically distributed sequence of Bernoulli trials, in which the number of times a given usability problem (a random variable X) has been observed after n participants follows a binomial distribution, $X \sim \mathcal{B}in(n, p)$. Considering the binomial distribution of the margin sums, the proportion of problems that has been discovered at least once after n participants is given by the cumulative function of the shifted geometric distribution [4, 14, 15]:

$$P(X > 0) = 1 - (1 - p)^n \quad (1)$$

The total number of problems m is then deduced from the following relationship:

$$j = (1 - (1 - p)^n) \times m \quad (2)$$

The discovery progress is thus assessed in two steps: the probability of detection p is first estimated and then plugged into Equation (2) to estimate the number of problems m. A wide range of literature methods are available for estimating the probability of problem detection. The simplest way involves computing the naive estimate (denoted as $\hat{p}$) using the observed discovery matrix $\mathbb{d}$:

$$\hat{p} = \frac{\sum_{i=1}^{n} \sum_{l=1}^{j} x_{il}}{n * j} \quad (3)$$

As mentioned above, the naïve estimate is systematically biased - especially for small samples. Indeed, unobserved problems result in zero columns that shrink the probability space and lead to overestimation of p, particularly at the beginning of the process when $j \ll m$. Consequently, m is systematically underestimated, which generates safety

153 concerns in the medical device field. In response, several strategies have been employed
154 to overcome the truncated matrix problem.

155 In 2001, Hertzum and Jacobsen [7] suggested normalizing the value of $\hat{p}$. This procedure
156 considers that the lower boundary of the probability of detection estimated with n
157 participants is $1/n$. For example, in a sample of 5 participants, $\hat{p} \in [0.2\ ;\ 1]$. Conversely,
158 the normalized estimator $\hat{p}_{Norm} \in [0; 1]$, and is computed as follows:

159 $$\hat{p}_{Norm} = \frac{\hat{p} - \frac{1}{n}}{1 - \frac{1}{n}} \quad (4)$$

160 However, the normalized approach suffers from a major limitation when estimating the
161 total number of problems with Equation (4). In fact, if each participant has discovered
162 only one problem and if each problem was discovered only once, $\hat{p} = \frac{1}{n}$, $\hat{p}_{Norm} = 0$, and
163 the estimated number of problems $\widehat{m}$ is infinite. We will not discuss this estimation
164 method further.

165 Turing and Good developed a discounting method for estimating the probability of
166 unseen species on the basis of observed data [16]. Lewis suggested that the Good-Turing
167 (GT) adjustment could be used to reduce the magnitude of the overestimation of p by
168 increasing the probability space and thus accounting for unobserved usability problems
169 [6]. The GT adjustment is computed as the proportion of singletons relative to the total
170 number of events (i.e. the proportion of problems discovered only once, $x_{il} = 1$), and is
171 incorporated in the estimation as follows:

172 $$\hat{p}_{GT} = \frac{\hat{p}}{1 + GT} \quad (5)$$

However, Lewis observed that use of the GT estimator overestimated p. He empirically assessed the best adjustment for a small sample size by carrying out Monte Carlo simulations on a range of usability study databases involving web or software user interfaces with known true values. Based on these simulations, Lewis concluded that the best method was to average the GT adjustment and a "double deflation" term:

$$\hat{p}_{Lewis} = \frac{1}{2}\left[\frac{\hat{p}}{1 + GT_{adj}}\right] + \frac{1}{2}\left[\left(\hat{p} - \frac{1}{n}\right) \times \left(1 - \frac{1}{n}\right)\right] \quad (6)$$

Nevertheless, the degree of adjustment of the probability space for unobserved problems is essentially empirical. The residual bias is not known to trend towards over- or underestimation.

In 2008, Schmettow considered the problem margin sums in a zero-truncation framework [17]. Indeed, the distribution of the problems so far observed follows a binomial distribution with only a positive integer as support (i.e. a positive or conditional distribution). The distribution is truncated because the random variable cannot be equal to zero. The probability is then estimated using standard mathematical techniques, such as the maximum likelihood or moment estimator [18–20]. The probability mass function is:

$$P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k} \quad (7)$$

and zero truncation is achieved as follows:

$$P(X = k)_{zt} = \begin{cases} 0 & if\ k = 0 \\ \dfrac{P(X = k)}{1 - P(X = 0)} & if\ k > 0 \end{cases} \quad (8)$$

The probability of problem discovery is then estimated by using maximum likelihood techniques to fit the marginal sums to the zero-truncated binomial distribution. It should

be noted that the expected probability of unobserved problems, $\Pr(X = 0)$, is deduced from the non-truncated function [17].

D. Methods taking account of a heterogeneous problem detection probability

All the methods presented above assume that the probability of detection is the same for all usability problems (i.e., the same p). However, this assumption seems unrealistic and very unlikely to hold in real-life usability studies. Schmettow showed that overdispersion was frequent in the problem margin sums, reflecting heterogeneity in the probability of detection [21]. Furthermore, erroneously ignoring the presence of heterogeneity by using a single, average value of p leads to overestimation of the completeness of the discovery process (Jensen's inequality) [22]. Schmettow tackled this problem by developing a model that incorporated heterogeneity. The probability of detection was considered to be a random variable, which enabled each problem to have its own probability of detection. Schmettow used the logit-normal distribution as a prior for the probability of detection. Formally, the logit of the probability of detection follows a normal distribution $\mathcal{N}(\mu, \sigma)$. In this model, the problem margin sums follows a logit-normal binomial distribution and the probability mass function is:

$$P(X = k) = \binom{n}{k} \frac{1}{\sqrt{2\pi}\sigma} \int_0^1 (1-p)^{n-k-1} p^{k-1} \exp\left(-\frac{(logit(p) - \mu)^2}{2\sigma^2}\right) dp \qquad (9)$$

Using the zero truncation technique presented in equation (8), Schmettow developed the logit-normal binomial zero truncated (LNBzt) model and applied it to the usability of medical infusion pumps [23]. To the best of our knowledge, this model is the only one that accounts for both heterogeneity and unobserved problems.

E. Statistical limitations of margin-based methods

The primary limitation of the margin-based methods presented above is that they estimate the probability of detection only. The number of problems m is deduced but not estimated *per se*. It would be possible to estimate both m and p by summarizing the discovery matrix on the basis of the participants' margin. In such a case, each sum follows a binomial $\mathcal{B}in(m, p)$, thus enabling estimation of both the number of attempts and the probability of success in a binomial setting. However, DasGupta and Rubin established that there were no unbiased estimates for essentially any functions of either the number of attempts or the probability of success [24]. This problem was initially considered by Fisher and Haldane for estimating species abundance [25, 26]. It has also been considered by Olkin, Petkau, and Zidek, who developed both a moment and a maximum likelihood estimator, and by Carroll and Lombard, who proposed an estimator in a Bayesian setting (leading to a beta-binomial distribution) [27, 28]. Hall also considered this problem in an asymptotic framework [29].

The second limitation of margin-based methods is information loss, relative to the initially available data. For example, j and the number of singletons were the only data used in the GT estimates. In the same way, the zero-truncated method considered only the column sums for the problems and omitted the pattern of detection (i.e., the users).

Here, we tackled the problem by directly modelling the distribution of the discovery matrix and thus accounted for all the available information. We addressed the problem of estimating both p and m by considering a maximum likelihood estimator and Bayesian inference, in much the same way as Carroll and Lombard [27]. In the Methods section, we present the statistical basis of our method, i.e. the likelihood of the discovery matrix. In the Results section, we shall describe two simulations comparing our matrix-based

method with margin-based methods in the context of homogeneous and then heterogeneous problem detection probabilities. We also applied our method to real usability testing data for medical infusion pumps [23]. Lastly, we discuss the implications for estimating the number of problems in usability studies.

II. Methods

We first specify the statistical basis underpinning the matrix-based method in general and the principle of column permutation in particular. We next present our estimation of the total number of usability problems m in the context of a homogeneous probability of detection. In the third part of the Methods, we explain how the matrix-based model can be adapted for use with a heterogeneous probability of detection. The last part is dedicated to the methods used to assess the matrix-based model's performance.

A. The matrix-based method

We first present the matrix-based method and, for sake of clarity, assume that the probability of problem detection is homogeneous. Consider the complete discovery matrix $\mathbb{x}$. The probability of $\mathbb{x}$ can be written as follows:

$$P(\mathbb{x}|p,m) = p^{\mathbb{x}_{\bullet\bullet}}(1-p)^{nm-\mathbb{x}_{\bullet\bullet}} \quad (10)$$

where $\mathbb{x}_{\bullet\bullet} = \sum_{i=1}^{n}\sum_{l=1}^{m} x_{il}$ is the total number of problems observed by n individuals.

An example of a possible matrix $\mathbb{x}$ obtained from two participants during a usability study of a medical device with $m = 3$ problems is given below (with users in rows and problems in columns):

$$\mathbb{x} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix} \quad (11)$$

As seen above, the complete discovery matrix $\mathbb{x}$ is never observed, and the discovery matrix $\mathbb{d}$ is the only one available. It is similar to the matrix $\mathbb{x}$, except that unobserved problems are missing. Considering the above example, neither of the users observed the second problem, and the resulting observed discovery matrix $\mathbb{d}$ would be:

$$\mathbb{d} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \quad (12)$$

It should be noted that if the total number of problems m is known, then the complete matrix $\mathbb{x}$ could be reconstituted (with permutation), based on the matrix $\mathbb{d}$. For instance, if we take the matrix $\mathbb{x}$ and consider (wrongly, in this case) that the number of problems $m = 5$, then the reconstituted complete matrix denoted by $\hat{\mathbb{x}}^m$ would be obtained by padding the matrix $\mathbb{d}$ with columns of zeros (corresponding to as-yet unobserved problems):

$$\hat{\mathbb{x}}^{m=5} = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \end{pmatrix} \quad (13)$$

Thus, noting that $\mathbb{x}_{\bullet\bullet} = \mathbb{d}_{\bullet\bullet}$, it is possible to compute the likelihood of the complete matrix $\hat{\mathbb{x}}^m$ on the basis of the discovery matrix $\mathbb{d}$. This likelihood is given by the following formula:

$$P(\hat{\mathbb{x}}^m | p, m) = p^{\mathbb{x}_{\bullet\bullet}} (1-p)^{nm - \mathbb{x}_{\bullet\bullet}} \quad (14)$$

Thus, a maximum likelihood estimation of (p, m) based on $\hat{\mathbb{x}}^m$ would consist in maximizing $p(\hat{\mathbb{x}}^m | p, m)$ with respect to m and p. This process leads to $\hat{m} = j$, where j is the number of problems observed so far, and $p = \frac{\mathbb{x}_{\bullet\bullet}}{nj}$. However, as noted above, these parameter estimates are biased because $p(\hat{\mathbb{x}}^m | p, m)$ is not the likelihood of the observed data. In fact, the main problem is that the definition of observed data ($\hat{\mathbb{x}}^m$) depends on

the value m, which is unknown. We tackled this issue by modeling the distribution of the observed discovery matrix $p(\mathbb{d}|p,m)$.

It should be noted that the matrix $\mathbb{d}$ is defined in a lexicographic order, which simply means that the problems are ordered in the order of detection. For instance, the six possible complete matrices $\mathbb{x}$ leading to the previous matrix $\mathbb{d}$ if $m = 3$ are presented in Table 1.

Table 1: Six possible complete matrices $\hat{\mathbb{x}}^{m=3}$ leading to the observed discovery matrix $\mathbb{d} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$

Possibility 1	Possibility 2	Possibility 3
$\hat{\mathbb{x}}_1^{m=3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	$\hat{\mathbb{x}}_2^{m=3} = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$	$\hat{\mathbb{x}}_3^{m=3} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix}$
Possibility 4	**Possibility 5**	**Possibility 6**
$\hat{\mathbb{x}}_4^{m=3} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$	$\hat{\mathbb{x}}_5^{m=3} = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$	$\hat{\mathbb{x}}_6^{m=3} = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \end{pmatrix}$

In fact, if we could consider the label (the name of the usability problem) associated with each column, only one matrix $\mathbb{x}$ could lead to the matrix $\mathbb{d}$. However, since we have no means of finding the names of the columns in the initial matrix $\mathbb{x}$, we will consider that the matrix $\mathbb{d}$ has unnamed columns. Removing these column names allows us to consider the matrix $\mathbb{d}$ for the observed data (for which the definition does not vary as a function of the model's definition of the model – in contrast to $\hat{\mathbb{x}}^m$).

Since the probability of each matrix in Table 1 is $p^2(1-p)^4$, it follows that:

$$P(\mathbb{d}|m=3,p) = 6 \times p^2(1-p)^4 = A_3^2 \times p^2(1-p)^4 \quad (15)$$

More generally, the number of matrices $\mathbb{x}$ associated with an observed discovery matrix $\mathbb{d}$ is:

$$\frac{m!}{(m-j)!\,j_1!\ldots j_r!} = \frac{1}{j_1!\ldots j_r!} \times A_m^j \qquad (16)$$

where r is the number of different columns of $\mathbb{d}$, and j_h $(1 \leq h \leq r)$ is the number of repetitions of the column of type h. Of course, $j = j_1 + \cdots + j_r$. Here, we recognize a familiar equation: that associated with the number of anagrams of a word in which each type of column corresponds to a different letter, including the null column (repeated $m - j$ times).

Finally, we obtain the likelihood of $\mathbb{d}$ as follows:

$$P(\mathbb{d}|p,m) = \frac{1}{j_1!\ldots j_r!} \times A_m^j \times p^{\mathbb{x}_{\bullet\bullet}}(1-p)^{nm-\mathbb{x}_{\bullet\bullet}} \qquad (17)$$

In practice, the computation of $\frac{1}{j_1!\ldots j_r!}$ has no impact on the estimation, since it is the same for all values of m and p.

B. Estimation of m when the probability of detection is homogeneous

In the following section, we present two estimates of the total number of problems: the maximum likelihood and the Bayesian estimates.

Starting from matrix $\mathbb{d}$, the maximum likelihood estimator of m and p can be obtained by maximizing $p(\mathbb{d}|p,m)$ over (p,m):

$$(\hat{p},\hat{m}) = \arg \max_{m \geq j, p \in [0;1]} P(\mathbb{d}|p,m) \qquad (18)$$

Substituting p by $\frac{\mathbb{x}_{\bullet\bullet}}{nm}$ (the maximum likelihood estimator of p when m is fixed), the profiled likelihood function $\mathcal{L}$ that must be maximized is:

$$\mathcal{L}(m) = \frac{1}{j_1!\, j_2! \ldots j_r!} \times A_m^j \times \left(\frac{\mathbb{x}_{\bullet\bullet}}{nm}\right)^{\mathbb{x}_{\bullet\bullet}} \times \left(1 - \frac{\mathbb{x}_{\bullet\bullet}}{nm}\right)^{\mathrm{nm}-\mathbb{x}_{\bullet\bullet}} \quad (19)$$

For numerical convenience, we consider the logarithm of the likelihood ℓ:

$$\ell(m) = C + \log\left(A_m^j\right) + \mathbb{x}_{\bullet\bullet} \times \log\left(\frac{\mathbb{x}_{\bullet\bullet}}{nm}\right) + (\mathrm{nm} - \mathbb{x}_{\bullet\bullet}) \times \log\left(1 - \frac{\mathbb{x}_{\bullet\bullet}}{nm}\right) \quad (20)$$

The maximum does not have a closed form, and the values of m only need to be evaluated on a discrete grid. Once $\hat{m}$ has been found, $\hat{p}$ can be directly deduced by computing $\hat{p} = \frac{\mathbb{x}_{\bullet\bullet}}{n\hat{m}}$. For instance, we used the discovery matrix $\mathbb{d}$ presented on page 6 to plot the values of $\ell(m)$ according to m. We see that the estimated value of m is $\hat{m} = 12$.

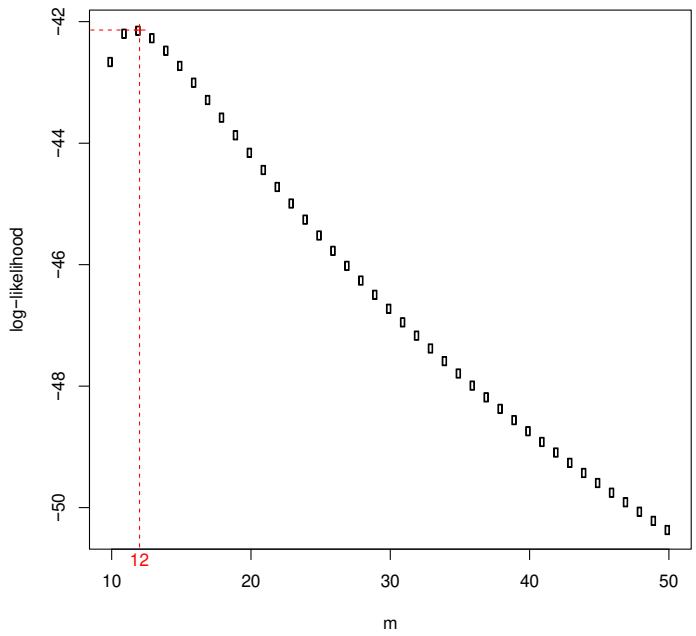

Figure 1: Log-likelihood according to the total number of problems (m), based on the observed discovery matrix (see $\mathbb{d}$ in page 6).. The maximum log-log likelihood is reached for m=12. Since the number of observed problems is j=10, the maximum likelihood estimate predicts that 2 usability problems remain undetected.

In theory, the maximum likelihood estimator has good properties, such as asymptotic convergence on the true value of the parameters [30]. However, if and only if $\mathbb{x}_{\bullet\bullet} = j$, the function $\ell(m)$ only increases - leading to an infinite value of $\hat{m}$ (see the proof in the Appendix). The situation where $\mathbb{x}_{\bullet\bullet} = j$ is particularly likely for small sample sizes. We

circumvented this possibility by replacing $\mathbb{x}_{\bullet\bullet}$ with $\mathbb{x}_{\bullet\bullet} + 1$, which produces a regularized solution associated with a finite value of $\hat{m}$.

Having built a probabilistic model of data generation, we can generate confidence intervals by applying a parametric bootstrap approach, i.e., by simulating new discovery matrices $\mathbb{d}$ based on $\hat{p}$ and $\hat{m}$ and by re-estimating p and m [31].

An alternative to the maximum likelihood estimation is the use of Bayesian inference, where unknown parameters p and m are considered to be random [32]. Here, we assume independent prior distributions for p and m, i.e., $p(m, p) = p(m)p(p)$. We also assume that the number of problems is distributed uniformly over the range of integers from 1 to M:

$$P(m) = \frac{1}{M}, \qquad \forall m \in \{1, \dots, M\} \tag{21}$$

We suppose that the prior distribution of p is a beta distribution with parameters α and β. We note that $p \sim \mathcal{B}(\alpha, \beta)$

$$P(p) = \frac{1}{\mathcal{B}(\alpha, \beta)} p^{\alpha-1} \times (1-p)^{\beta-1} \times \mathbb{1}_{[0,1]}(p) \tag{22}$$

where $B(.,.)$ is the beta function $\mathcal{B}(\alpha, \beta) = \int_0^1 \mathrm{p}^{\alpha-1} \times (1-p)^{\beta-1} dp$.

These two prior distributions require us to define the hyperparameters M, α, and β. It should be noted that the prior distribution of m is weakly informative. Taking $\alpha = \beta = 1$, we would obtain a uniform distribution over [0,1], which is also weakly informative. In the simulations, we chose $\alpha = \beta = 1$ and $M = 500$ because it is rare that $m > 150$ in typical usability studies. It should also be noted that prior distributions could also be based on the state of the art in usability tests.

Within a Bayesian framework, we are interested in $P(m, p\,|\mathbb{d})$, i.e. the probability of the parameters given the data. However, we also want to compute the marginal posterior distributions $P(m|\mathbb{d})$ and $P(p\,|\mathbb{d})$. It should be noted that above all, the cornerstone of the Bayesian inference in our setting is the probabilistic model of $\mathbb{d}$, i.e*.,* $P(\mathbb{d}|m, p)$, and so any approach based on $P(\hat{\mathbb{x}}^m|p, m)$ would fail.

We now specify the computation of $P(m|\mathbb{d})$:

$$P(m|\mathbb{d}) \propto P(m, \mathbb{d}) = \int_0^1 f(m, \mathbb{d}, p)df = \int_0^1 P(\mathbb{d}|m, p)\, P(m)\, f(p)dp \qquad (23)$$

and thus obtain $\forall m \in \{j, \dots, M\}$

$$\int_0^1 P(\mathbb{d}|m, p)\, P(m)\, f(p)dp = \frac{1}{j_1!\, j_2! \dots j_r!} \times A_m^j * \int_0^1 \frac{1}{\mathcal{B}(\alpha, \beta)} p^{\mathbb{x}_{\bullet\bullet}+\alpha-1} \times (1-p)^{nm-\mathbb{x}_{\bullet\bullet}+\beta-1} dp \quad (24)$$

and lastly,

$$P(m, \mathbb{d}) = \frac{1}{j_1!\, j_2! \dots j_r!} \times A_m^j \times \frac{\mathcal{B}(\mathbb{x}_{\bullet\bullet} + \alpha - 1, nm - \mathbb{x}_{\bullet\bullet} + \beta - 1)}{\mathcal{B}(\alpha, \beta)} \qquad (25)$$

Thus, removing the terms that do not depend on m yields $\forall m \in \{j, \dots, M\}$

$$P(m|\mathbb{d}) \propto A_m^j \times \mathcal{B}(\mathbb{x}_{\bullet\bullet} + \alpha, nm - \mathbb{x}_{\bullet\bullet} + \beta) \qquad (26)$$

Which provides the distribution of the number of problems, given the observed discovery matrix. Thus, m can be estimated by taking the *a posteriori* maximum $\widehat{m}_{MAP} = \arg\max_{m \geq j} [\, P(m|\mathbb{d})]$. The posterior expectation of m could also be considered but would be more sensitive to the choice of M - the border of the prior distribution of m. Thus, we included the *a posteriori* maximum in the simulations only. Using the discovery matrix $\mathbb{d}$

on page 6, the posterior distribution of the number of problems (given the discovery matrix) is shown in Figure 2.

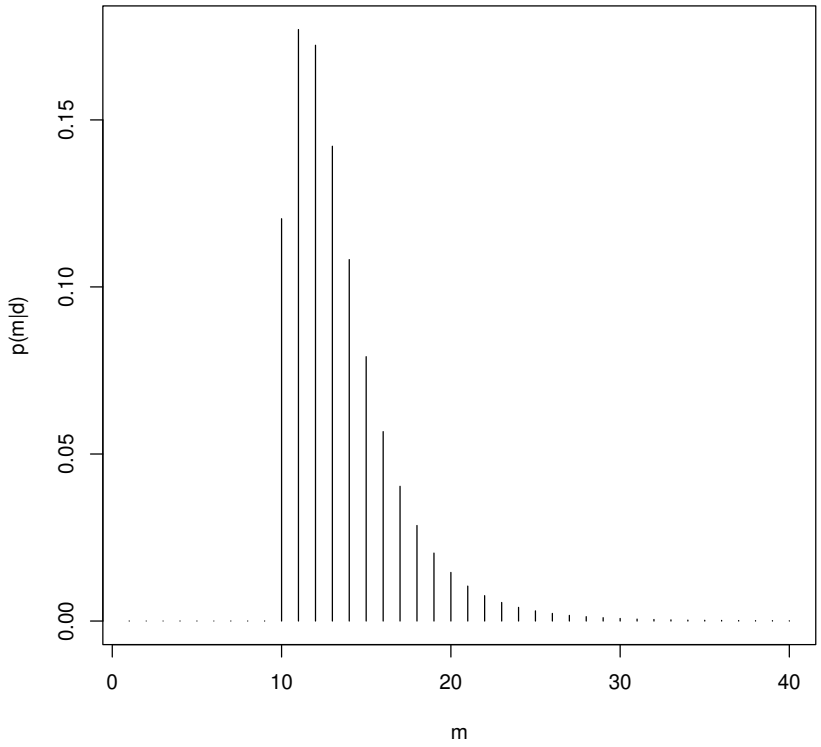

Figure 2: Posterior distribution of m, given the observed discovery matrix (see $\mathbb{d}$ in page 6). The distribution mode is at m=11. Since the number of observed problems is j=10, the Bayesian estimate predicts that 1 usability problem has yet to be detected.

The posterior distribution of p, i.e., $P(p\ |\mathbb{d})$, can also be obtained, as follows:

$$P(p\ |\mathbb{d}) = \sum_{m} P(p, m|\mathbb{d}) = \sum_{m} P(p|m, \mathbb{d}) \times P(m|\mathbb{d}) \qquad (27)$$

Moreover, due to the conjugacy of the prior in the previous equation, we obtain $p|m, \mathbb{d} \sim \mathcal{B}(\mathbb{x}_{\bullet\bullet} + \alpha, nm - \mathbb{x}_{\bullet\bullet} + \beta)$. Thus, the distribution of p given $\mathbb{d}$ is a mixture of beta distributions, where the weightings are given by $P(m|\mathbb{d})$. For the sake of simplicity in the simulations, p was estimated as $\frac{\mathbb{x}_{\bullet\bullet}+\alpha-1}{n\widehat{m}_{MAP}+\alpha+\beta-2}$, the mode of the posterior distribution $P(p|\widehat{m}_{MAP}, \mathbb{d})$. For the discovery matrix $\mathbb{d}$ presented on page 6, the posterior distribution, $P(p\ |\mathbb{d})$, is given in Figure 3.

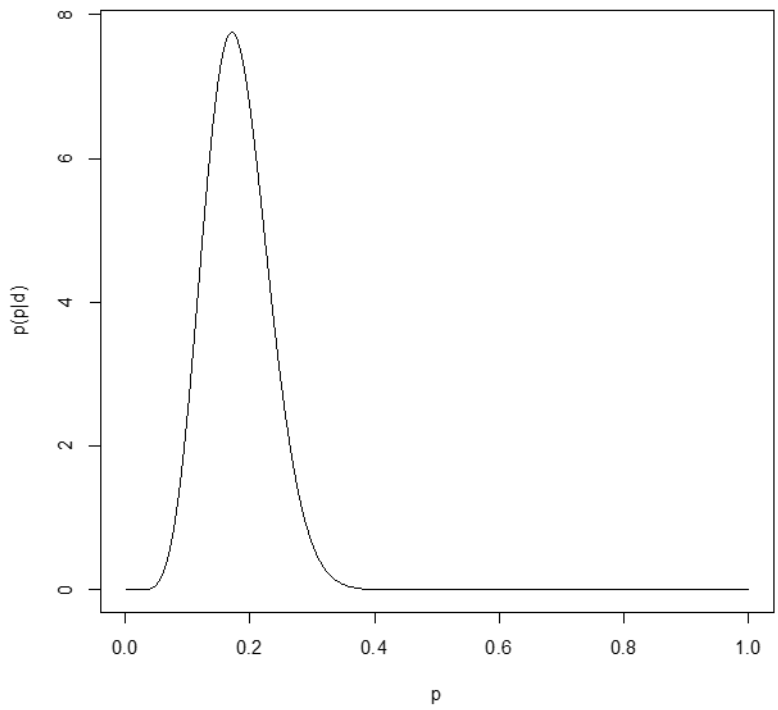

Figure 3: Posterior distribution of p, given the observed discovery matrix (see 𝕕 in page 6).

C. Estimation of m when the probability of detection is heterogeneous

Next, we considered a heterogeneous probability of detection; each problem l has its own probability of detection p_l. In line with Schmettow's method and for easier comparisons, we assume that the probabilities of detection are independent and follow a logit-normal distribution. We are interested in the likelihood of the complete discovery matrix $\mathbb{x}$ given μ, σ, and m:

$$P(\mathbb{x}|\mu,\sigma,m) = \int_0^1 \dots \int_0^1 P(\mathbb{x}|p_1,\dots,p_m,m) f(p_1,\dots,p_m|\mu,\sigma) dp_1 \dots dp_m \qquad (28)$$

by assuming that each column of the matrix is independent:

$$P(\mathbb{x}|p_1,\dots,p_m,m) = \prod_{l=1}^{m} p_l^{\mathbb{x}_{\bullet l}} (1-p_l)^{n-\mathbb{x}_{\bullet l}} \qquad (29)$$

Moreover, the prior distribution of the probability of detection can be written as follows:

$$f(p_1, \ldots, p_m|\mu,\sigma) = \prod_{l=1}^{m} P(p_l|\mu,\sigma) = \prod_{l=1}^{m}\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\text{logit}(p_l)-\mu)^2}{2\sigma^2}\right)\frac{1}{p_l(1-p_l)}\right) \quad (30)$$

Thus, we have:

$$P(\mathbb{x}|\mu,\sigma,m) = \prod_{l=1}^{m}\int_0^1\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\text{logit}(p_l)-\mu)^2}{2\sigma^2}\right)p_l^{\mathbb{x}_{\bullet l}-1}(1-p_l)^{n-\mathbb{x}_{\bullet l}-1}\right)dp_l \quad (31)$$

If we define $I_n(\mathbb{x}_{\bullet l},\mu,\sigma)$ as:

$$I_n(\mathbb{x}_{\bullet l},\mu,\sigma) = \int_0^1\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(\text{logit}(p_l)-\mu)^2}{2\sigma^2}\right)p_l^{\mathbb{x}_{\bullet l}-1}(1-p_l)^{n-\mathbb{x}_{\bullet l}-1}\right)dp_l \quad (32)$$

we can rewrite the likelihood of the complete discovery matrix $\mathbb{x}$ as follows:

$$P(\mathbb{x}|\mu,\sigma,m) = \prod_{l=1}^{m} I_n(\mathbb{x}_{\bullet l},\mu,\sigma) \quad (33)$$

By reusing the result of Equation (16), the probability of the observed discovery matrix $\mathbb{d}$ becomes:

$$P(\mathbb{d}|\mu,\sigma,m) = \frac{1}{j_1!\ldots j_r!}\times A_m^j \times \prod_{l=1}^{m} I_n(\mathbb{x}_{\bullet l},\mu,\sigma) \quad (34)$$

After substitution of the latter by $\mathbb{d}_{\bullet l}$ (where, by definition, $\mathbb{d}_{\bullet l} = 0$ for $l > j$), we can determine the probability of $\mathbb{d}$ given μ, σ, and m:

$$P(\mathbb{d}|\mu,\sigma,m) = \frac{1}{j_1!\ldots j_r!}\times A_m^j \times \prod_{l=1}^{j} I_n(\mathbb{d}_{\bullet l},\mu,\sigma)\times I_n(0,\mu,\sigma)^{m-j} \quad (35)$$

For numerical convenience, we will again consider the logarithm of the likelihood:

$$\log\big(P(\mathbb{d}|\mu,\sigma,m)\big) = C + \log\big(A_m^j\big) + \sum_{l=1}^{j} \log\big(I_n(\mathbb{d}_{\bullet l},\mu,\sigma)\big) + (m-j)\log\big(I_n(0,\mu,\sigma)\big) \quad (36)$$

The integral does not have a closed form but can be computed by numerical integration, and we can consider the estimation of μ, σ, and m by the maximum likelihood. If we wanted to consider a full Bayesian setting, we would have to assume a prior distribution for $p(\mu,\sigma|m)$; however, the choice of this prior would be complex.

D. Assessment of the performance of the matrix-based method

We performed two simulations to assess the performance of matrix-based vs. margin-based estimates of the total number of usability problems. In the first simulation, we assessed the performance of six methods (naïve, GT, Lewis, zero truncation, maximum likelihood, and Bayesian) in the context of a homogeneous probability of detection. In a second simulation, we compared the matrix-based method with the LNBzt method in the context of a heterogeneous probability of detection. Lastly, we applied the matrix-based method to the usability testing data provided as supplementary material in Schmettow's report on medical infusion pumps ("infpump.Rdata") [23].

Each simulation consisted in generating an observed discovery matrix $\mathbb{d}$ from the usability study of a hypothetical medical device with a known total number of usability problems m and a sample size n. The probability of detection was constant (p) or normally distributed ($\mathcal{N}(\mu,\sigma)$) on a logit scale, depending on the context. The combinations of parameters used in the simulations are specified in Table 2. The values were chosen to reflect a wide range of parameters encountered in usability studies of medical devices. In the case of a heterogeneous probability of detection, we started with a higher n because of the greater statistical complexity of the heterogeneous models,

which require more data to provide reliable estimates and an acceptable confidence interval.

Table 2: Combinations of parameters for the simulation studies with homogeneous and heterogeneous probabilities of detection.

Parameter	Homogeneous	Heterogeneous
Total number of usability problems	$m = 10,40,70,100$	$m = 20,50,100$
Sample size	$n = 5,10,15,20,25,50$	$n = 20,40,60$
Probability of problem detection	$p = 0.05, 0.1, 0.15, 0.25, 0.50$	$\mu = logit(0.1), logit(0.2)$ $\sigma = 0.2,0.5$
Number of combinations tested	**120**	**36**

In each homogeneous setting (i.e. each combination of m, p and n), we simulated $S = 10^5$ complete discovery matrices, $\mathbb{x}_{m,p,n,i}$, $i \in \{1,2, \dots, S\}$. In each heterogeneous setting (i.e. each combination of m, μ, σ and n), we simulated $S = 10^4$ complete discovery matrices. The matrices $\mathbb{d}$ were obtained by truncation of the zero columns (problems not yet discovered). In each setting, we averaged the estimates of m over the S simulations and computed the 95% fluctuation interval, together with the 0.025 and 0.975 quantiles. We also calculated the prediction's root mean squared error (RMSE) as the square root of the mean squared difference between the predicted and observed values of m:

$$RMSE(m,p,n) = \sqrt{\frac{1}{S}\sum_{i=1}^{S}(m - \hat{m}_i)^2} \qquad (37)$$

When the sample is small, little information is available; a tight credible interval might reflect overconfidence rather than a good estimation. Thus, to gauge the level of confidence that human factor engineers can place in each method, we computed the confidence intervals' coverage probability. In each setting, the coverage probability is the proportion of 95% confidence intervals of the simulated $\hat{m}_i$ that include the true value of m. The confidence intervals for $\hat{m}_i$ were computed using 1000 parametric bootstrap repetitions with the parameters $(\hat{m}_i, \hat{p}_i, n)$ or $(\hat{m}_i, \hat{\mu}_i, n)$.

All the analyses were carried out on a personal computer running R software (version 3.5.2). The coverage probability was computed using a server equipped with 12-core Intel® Xeon® E5-2650 v4 processors. The code is provided as supplementary material [see Additional file 1].

III. Results

A. Homogeneous probability of problem detection

The results of the simulation with a homogeneous probability of detection are presented for the four margin-based methods (naïve, GT, Lewis, and zero truncation) and the two matrix-based methods (maximum likelihood and Bayesian). The prediction error of m as a function of the sample size n and the RMSE are presented in Figure 4 and

Table 3, respectively. Three distinct areas are of major interest: (i) the first column, corresponding to problems with a very low probability of detection ($p = 0.05$), (ii) the first row, corresponding to medical devices with a small number of problems ($m = 10$), and (iii) the rest of the observations in the bottom right corner, with a probability of detection $p \geq 10\%$ and a number of problems $m \geq 40$. The behavior of the estimators for each area is described below.

When the probability of detection is low (the p $= 0.05$ column in Figure 4 and

Table 3), the margin-based methods strongly underestimate m in all settings. This is particularly true for the naïve and the zero-truncated methods. The GT method compensates for this underestimation but is still biased in situations with fewer than 15 participants. Conversely, it shows the best RMSE - regardless of the number of problems. The Lewis method is conservative, except when $n = 5$ participants. Unlike the five other methods, the Lewis estimator is inconsistent because the RMSE does not decrease with

sample size. In contrast, the matrix-based methods performed well when the probability of detection was low. The maximum likelihood estimate was unbiased for $n \geq 15$ participants but its behavior was erratic for smaller sample sizes. The Bayesian method was less biased, except when the total number of undetected problems was small ($m = 10$). Notably, the confidence interval of the margin-based methods narrows with both m and n, whereas the matrix-based confidence intervals narrow with increasing sample size. This overconfidence illustrates the loss of information, and leads one to falsely consider that the higher the number of problems, the more information one has.

When the medical device had fewer problems ($m = 10$ rows in Figure 4 and Table 3), the naïve and the zero-truncated estimators showed poor behavior with persistent underestimation. In this context, the GT estimator was less biased and exhibited a very low RMSE. Again, the Lewis estimator constantly overestimated the true m. The matrix-based estimators also show good convergence properties but were only reliable when $n \geq 15$ participants. For a large sample size ($n \geq 50$), all estimators were consistent and exhibited a low RMSE.

In the other settings (p $\geq$ 10% and m $\geq$ 40, except for the first row and first column in Figure 4 and Table 3), the matrix-based methods showed the best properties, with unbiased estimates for as few as 5 participants, a low RMSE, and consistency as the sample size increased. The margin-based methods were either biased downward (naïve and zero-truncated) or showed a non-monotonous change with sample size (GT and Lewis); both situations led to a high RMSE. Again, as pointed out by Lewis, the average of the GT adjustment and the double deflation is a conservative estimator of m and is upwardly biased - leading to the worst RMSE.

503 The coverage probabilities are shown in

504 Table 4. The coverage probability observed in the first row and first column is of
505 particular interest because this is where the uncertainty is greatest, and the estimates are
506 most biased. As expected, the Lewis and the GT methods provided a coverage probability
507 that was high when m was low but that fell dramatically as m increased. Conversely, the
508 coverage probability of both matrix-based methods increased with m. In all other
509 settings, the matrix-based methods provided the best coverages and the margin-based
510 methods showed overconfidence - especially for a high value of m.

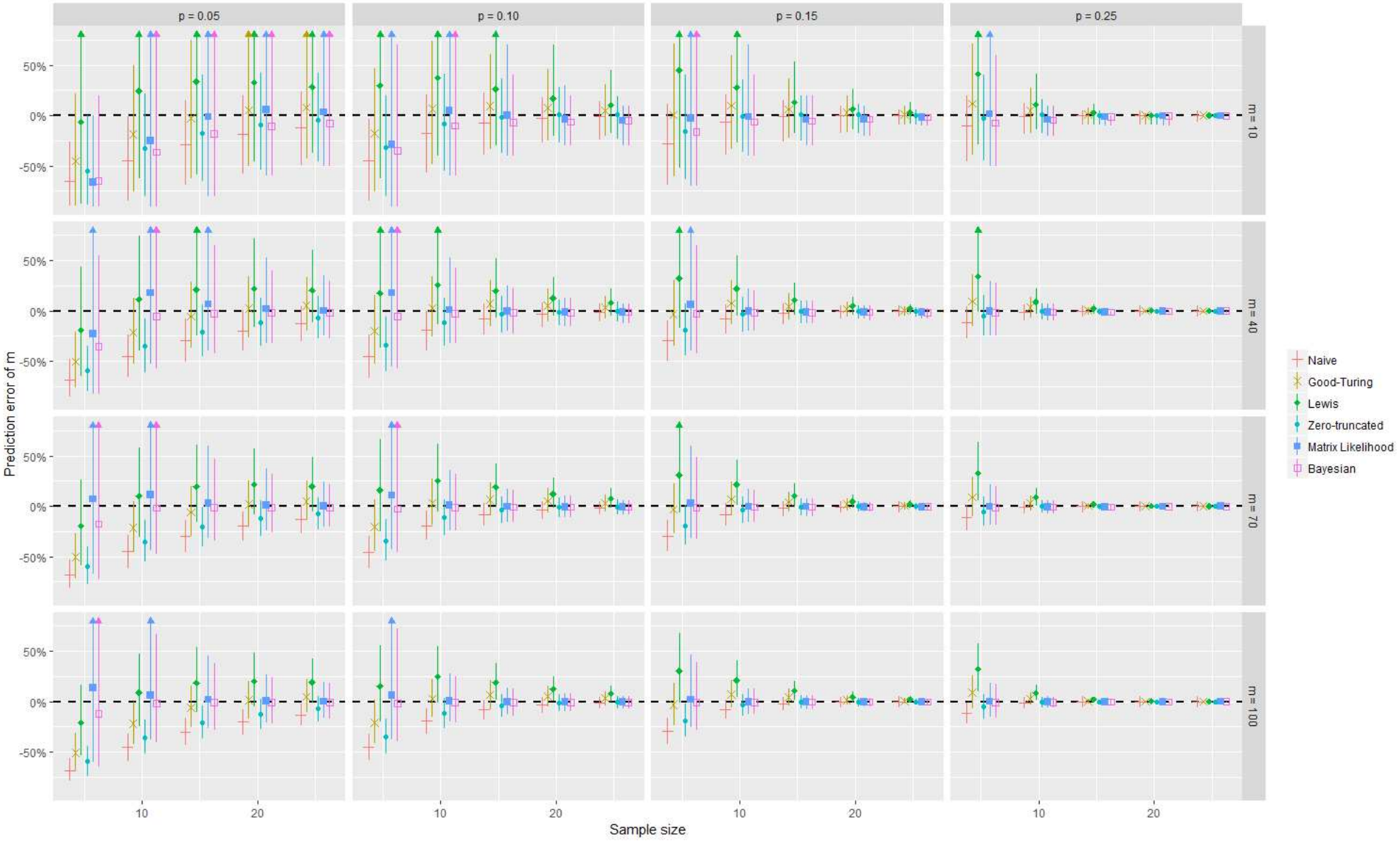

Figure 4: Prediction error for m (mean and 95% fluctuation interval, in %) as a function of the sample size (n), for 6 estimators. The results are presented for various probabilities of detection (p, columns) and various numbers of usability problems (m, rows). The dashed line represents the "true" m. The upper boundaries of credible intervals that exceed 100% are indicated by ▵.

Table 3: The RMSE for the estimation of m when the probability of problem detection is homogeneous.

		Probability of detection p=0.05						Probability of detection p=0.10						Probability of detection p=0.15						Probability of detection p=0.25						Probability of detection p=0.50					
	n	5	10	15	20	25	50	5	10	15	20	25	50	5	10	15	20	25	50	5	10	15	20	25	50	5	10	15	20	25	50
m = 10																															
Naïve		7	5	4	3	2	1	5	3	2	1	1	0	4	2	1	1	0	0	2	1	0	0	0	0	1	0	0	0	0	0
Good-Turing		5	4	3	3	3	1	4	3	3	2	1	0	3	3	2	1	1	0	3	1	0	0	0	0	1	0	0	0	0	0
Lewis		5	7	7	6	5	2	7	6	4	3	2	0	7	5	2	1	1	0	6	2	1	0	0	0	2	0	0	0	0	0
Zero truncated		6	4	3	3	2	1	4	3	2	1	1	0	3	2	1	1	0	0	2	1	0	0	0	0	1	0	0	0	0	0
Max likelihood		7	6	6	5	4	1	5	5	3	2	1	0	5	3	1	1	0	0	4	1	0	0	0	0	1	0	0	0	0	0
Bayesian		7	6	5	4	3	1	5	4	2	2	1	0	4	2	1	1	0	0	3	1	0	0	0	0	1	0	0	0	0	0
m = 40																															
Naïve		27	19	13	9	6	2	19	9	4	3	2	0	12	4	2	1	1	0	6	2	1	0	0	0	1	0	0	0	0	0
Good-Turing		21	11	7	6	6	3	11	6	5	4	3	0	7	5	3	2	1	0	6	2	1	0	0	0	2	0	0	0	0	0
Lewis		13	13	13	13	11	4	14	14	10	7	4	1	17	11	5	3	1	0	16	4	1	0	0	0	5	0	0	0	0	0
Zero truncated		24	15	10	7	5	2	15	7	4	3	2	0	9	4	2	1	1	0	4	2	1	0	0	0	1	0	0	0	0	0
Max likelihood		20	28	15	9	6	2	28	9	4	3	2	1	16	4	2	1	1	0	6	2	1	0	0	0	1	0	0	0	0	0
Bayesian		21	18	11	8	6	2	18	8	4	3	2	1	11	4	2	1	1	0	5	2	1	0	0	0	1	0	0	0	0	0
m = 70																															
Naïve		48	32	22	15	10	3	32	15	7	4	3	1	21	7	3	2	1	0	9	2	1	0	0	0	2	0	0	0	0	0
Good-Turing		36	18	10	8	8	4	17	8	8	6	4	1	9	8	5	2	1	0	9	4	1	0	0	0	3	0	0	0	0	0
Lewis		21	17	19	19	17	6	20	21	16	10	7	1	26	17	8	4	2	0	25	7	2	1	0	0	8	0	0	0	0	0
Zero truncated		42	26	16	11	7	3	25	10	6	4	3	1	15	5	3	2	1	0	6	2	1	0	0	0	2	0	0	0	0	0
Max likelihood		41	40	16	11	8	3	41	11	6	4	3	1	18	6	3	2	1	0	7	2	1	1	0	0	2	0	0	0	0	0
Bayesian		35	26	14	10	7	3	26	10	6	4	3	1	15	5	3	2	1	0	7	2	1	1	0	0	2	0	0	0	0	0
m = 100																															
Naïve		68	46	31	21	14	3	46	20	10	5	3	1	30	9	4	2	1	0	13	3	1	1	0	0	2	0	0	0	0	0
Good-Turing		51	24	12	10	10	5	23	10	10	7	5	1	11	10	6	3	2	0	12	4	1	1	0	0	3	0	0	0	0	0
Lewis		27	20	24	25	23	8	25	29	21	14	9	1	35	23	11	5	3	0	35	10	2	1	0	0	10	0	0	0	0	0
Zero truncated		60	37	23	15	10	3	35	14	7	4	3	1	21	7	3	2	1	0	8	3	1	1	0	0	2	0	0	0	0	0
Max likelihood		73	37	19	12	9	3	44	12	7	4	3	1	19	7	4	2	1	0	8	3	1	1	0	0	2	0	0	0	0	0
Bayesian		51	29	17	12	9	3	30	12	7	4	3	1	17	7	4	2	1	0	8	3	1	1	0	0	2	0	0	0	0	0

Table 4: Coverage probability (in %) of the parametric bootstrap 95% confidence interval of $\hat{m}$ with each combination (m, p, n).

		Discovery rate p=0.05						Discovery rate p=0.10						Discovery rate p=0.15						Discovery rate p=0.25						Discovery rate p=0.50					
	n	5	10	15	20	25	50	5	10	15	20	25	50	5	10	15	20	25	50	5	10	15	20	25	50	5	10	15	20	25	50
***m* = 10**																															
Naïve		2	14	51	63	71	93	15	55	80	88	91	100	39	80	91	96	99	100	71	93	99	100	100	100	97	100	100	100	100	100
Good-Turing		43	80	94	97	97	99	83	97	99	99	99	100	94	99	99	99	100	100	99	99	100	100	100	100	99	100	100	100	100	100
Lewis		97	98	99	99	99	100	99	99	99	100	100	100	99	98	100	100	100	100	95	100	100	100	100	100	99	100	100	100	100	100
Zero truncated		6	36	68	84	91	96	37	85	94	95	96	100	68	94	96	98	99	100	90	97	99	100	100	100	98	100	100	100	100	100
Max likelihood		76	95	96	93	90	84	95	92	85	85	80	95	96	87	83	81	86	100	88	83	88	97	99	100	83	99	100	100	100	100
Bayesian		14	42	64	74	79	77	47	76	77	75	77	95	69	77	77	76	85	100	77	76	88	97	99	100	82	99	100	100	100	100
***m* = 40**																															
Naïve		0	0	2	12	34	87	0	13	54	77	88	98	1	56	85	92	91	100	35	89	92	99	100	100	90	100	100	100	100	100
Good-Turing		3	59	92	98	98	98	62	98	96	96	98	98	95	96	96	99	98	100	95	98	98	99	100	100	99	100	100	100	100	100
Lewis		87	99	91	76	63	69	97	63	43	45	61	99	71	31	42	77	99	100	14	36	99	100	100	100	6	100	100	100	100	100
Zero truncated		0	26	66	81	88	94	30	82	91	94	95	98	71	92	94	95	94	100	90	95	94	99	100	100	94	100	100	100	100	100
Max likelihood		100	96	94	95	94	90	96	94	93	92	90	82	95	93	90	85	73	99	93	88	71	88	97	100	82	96	100	100	100	100
Bayesian		55	86	91	91	92	88	87	91	91	91	89	82	90	91	89	85	71	99	92	87	69	88	97	100	81	96	100	100	100	100
***m* = 70**																															
Naïve		0	0	0	2	11	84	0	2	34	70	85	95	0	35	82	91	93	100	14	87	91	98	100	100	90	100	100	100	100	100
Good-Turing		0	36	89	98	96	92	41	98	88	84	91	96	95	85	84	98	99	100	84	88	98	98	100	100	89	100	100	100	100	100
Lewis		79	98	77	48	29	30	94	29	11	12	22	100	41	5	9	36	89	100	1	5	82	100	100	100	0	100	100	100	100	100
Zero truncated		0	11	48	71	82	94	15	73	89	93	95	95	57	89	94	95	95	100	87	95	93	98	100	100	95	100	100	100	100	100
Max likelihood		99	95	95	95	94	92	95	95	94	93	92	71	95	94	93	89	82	98	94	91	77	80	95	100	88	94	100	100	100	100
Bayesian		78	92	93	93	92	91	91	93	93	92	91	70	93	93	92	88	82	98	93	90	76	80	95	100	88	94	100	100	100	100
***m* = 100**																															
Naïve		0	0	0	0	4	80	0	0	19	60	82	91	0	21	75	90	93	100	5	85	94	96	100	100	90	100	100	100	100	100
Good-Turing		0	20	86	98	94	83	25	98	79	71	81	95	94	73	71	94	99	100	71	76	99	96	100	100	74	100	100	100	100	100
Lewis		72	97	63	28	11	11	91	12	3	3	6	100	20	1	1	13	63	100	0	1	49	100	100	100	0	100	100	100	100	100
Zero truncated		0	4	34	61	77	95	6	64	87	92	94	92	44	87	95	95	95	100	85	95	95	96	100	100	95	100	100	100	100	100
Max likelihood		98	95	95	95	95	93	95	95	95	94	93	63	95	95	94	91	86	97	95	92	84	72	93	100	89	90	100	100	100	100
Bayesian		85	93	94	94	94	93	93	94	93	93	92	63	93	94	93	91	85	97	94	91	84	72	93	100	89	90	100	100	100	100

B. Heterogeneous probability of problem detection

The results of the simulation in the context of a heterogeneous probability of detection are presented below for the LNBzt margin-based method and the maximum likelihood matrix-based method. The prediction error of m as a function of the sample size n is presented in Figure 5, and the RMSE is presented in Table 5.

The Matrix likelihood method showed less bias overall; the bias ranged from-5.3% to +0.1% for the 36 combinations simulated. In comparison, the bias associated with the LNBzt method ranged from 0 to +17.3%. The LNBzt also strongly overestimated the total number of problems at low probabilities of detection ($\mu = logit(0.1)$). A greater degree of heterogeneity in the probability of detection ($\sigma = 0.5$) mitigated this bias. Conversely, the maximum likelihood method was associated with greater bias when σ increased. Both methods showed consistency, although the RMSE was high for low probabilities of detection and $n = 20$ - especially with LNBzt. The coverage of the confidence interval was high (Table 6), and was greater for the LNBzt methods when the total number of problems was high. However, an important limitation of the LNBzt method is the large number of unsuccessful fittings – causing the estimation to fail in 33% of cases. Although changing the start values for the optimization function might reduce the fitting failure rate, this is a major drawback of the LNBzt method.

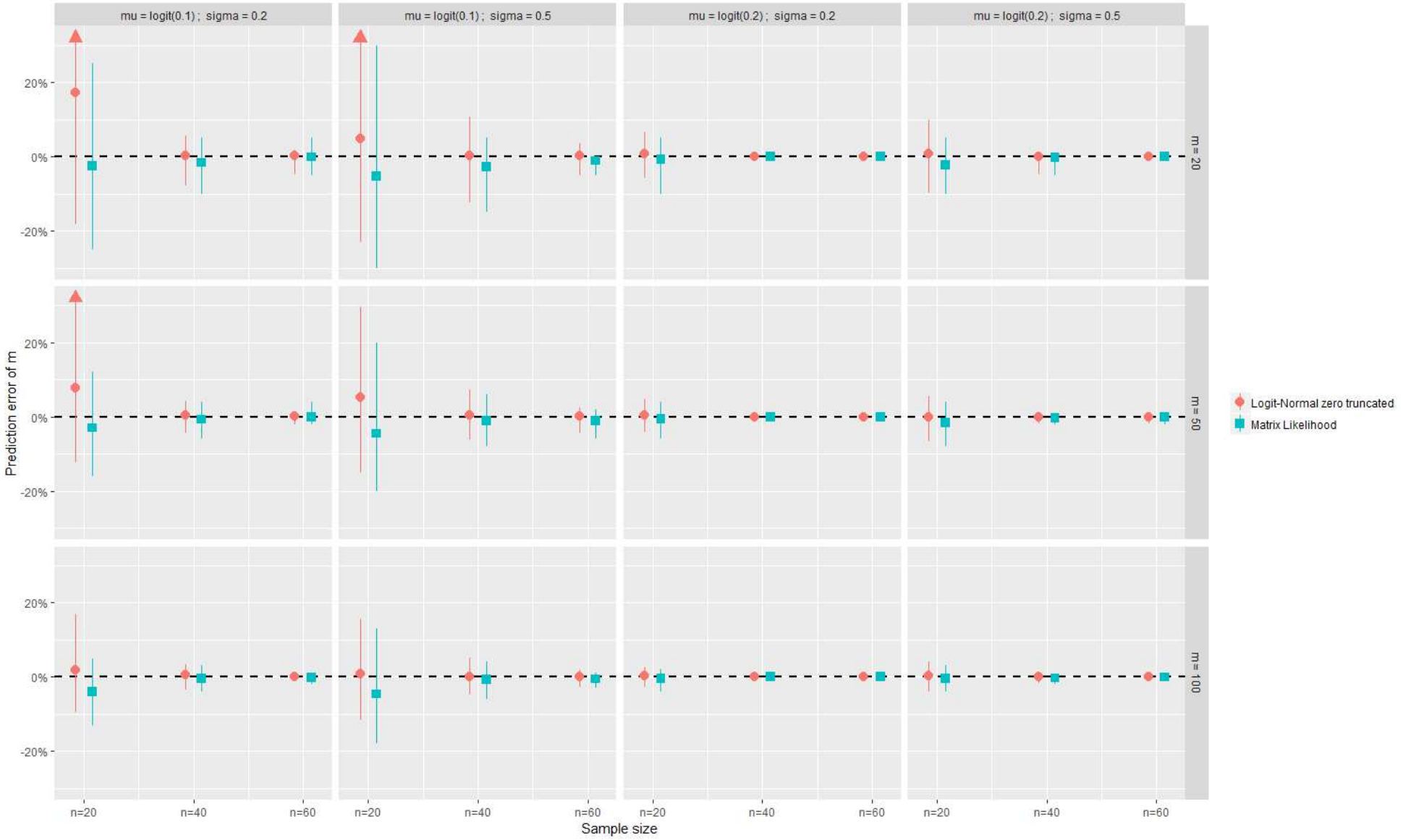

Figure 5: Prediction error of m (mean and 95% fluctuation interval, in %) as a function of the sample size (n) in the context of a heterogeneous probability of problem detection. The results are presented for various probabilities of problem detection ((μ, σ), columns) and various numbers of usability problems (m, rows). The dashed line represents the "true" m. The upper boundaries of the credible intervals that exceed 100% are indicated by ∆.

Table 5: The RMSE for the estimation of m when the probability of problem detection is heterogeneous.

		Probability of detection mu=0.1; sigma=0.2			***Probability of detection mu=0.1; sigma=0.5***			***Probability of detection mu=0.2; sigma=0.2***			***Probability of detection mu=0.2; sigma=0.5***		
	n	20	40	60	20	40	60	20	40	60	20	40	60
m= 20													
Logit-normal binomial zero truncated		30.5	0.8	0.3	7.1	1.2	0.6	0.7	0.2	0.1	1.1	0.3	0.2
Matrix likelihood		3.2	0.8	0.4	4.4	1.2	0.6	0.7	0.1	0	1.1	0.3	0.1
m= 50													
Logit-normal binomial zero truncated		19.4	1.2	0.5	19.4	1.8	0.9	1.2	0.2	0.1	1.6	0.4	0.2
Matrix likelihood		4.5	1.3	0.6	5.5	1.9	1	1.2	0.2	0.1	1.8	0.5	0.2
m= 100													
Logit-normal binomial zero truncated		7.4	1.8	0.7	7.2	2.5	1.2	1.5	0.2	0.1	2.2	0.6	0.3
Matrix likelihood		6.5	1.9	0.8	9.6	2.5	1.4	1.6	0.3	0.3	2.2	0.7	0.3

Table 6: Coverage probability (in %) of the parametric bootstrap 95% confidence interval of $\hat{m}$ with each combination $(m, (\mu, \sigma), n)$.

		Discovery rate mu=0.1; sigma=0.2			***Discovery rate mu=0.1; sigma=0.5***			***Discovery rate mu=0.2; sigma=0.2***			***Discovery rate mu=0.2; sigma=0.5***		
	n	20	40	60	20	40	60	20	40	60	20	40	60
m= 20													
Logit-normal binomial zero truncated		99%	100%	100%	97%	97%	98%	99%	100%	100%	98%	100%	100%
Matrix likelihood		99%	99%	99%	96%	98%	99%	99%	100%	100%	99%	99%	100%
m= 50													
Logit-normal binomial zero truncated		97%	98%	99%	97%	97%	95%	98%	100%	100%	95%	100%	100%
Matrix likelihood		94%	95%	96%	86%	91%	91%	96%	99%	100%	90%	92%	98%
m= 100													
Logit-normal binomial zero truncated		98%	99%	99%	97%	96%	95%	99%	100%	100%	96%	97%	100%
Matrix likelihood		77%	94%	94%	72%	91%	85%	93%	97%	95%	90%	85%	95%

C. Lessons learned from the simulation studies

When the probability of detection was homogeneous, the matrix-based methods (Bayesian and maximum likelihood) showed the best properties over a large range of settings (bottom right corner). When the probability of detection and/or the number of problems was low, all the estimators were challenged because few problems were observed in the discovery matrix (the first column and first row in Figure 4 and

Table 3). In such cases, the GT method performed slightly better than the matrix-based methods, and all the methods converged rapidly on the expected number and the best RMSE. With a small m and small p, the other methods either systematically underestimated the true value (Naïve and Zero-truncated) or overestimated the true value (Lewis) - especially in the most critical "top left" setting ($m = 10; p = 0.05$). With regard to coverage of the confidence interval, the Lewis and the GT methods provide a high coverage probability when the number of problems was low. In all other settings, the matrix-based methods were more reliable. For margin-based methods, information loss led to overconfidence for high values of m.

When the probability of detection was heterogeneous, the matrix-based method showed less bias and a lower RMSE - especially for low values of m. At the same time, the precision of LNBzt method increased as the probability of detection became more heterogeneous. This might be related to a better handling of the truncation of the discovery matrix by this matrix-based method. Conversely, the precision decreased as the heterogeneity increased.

D. Application to real data from a usability study of medical infusion pumps

The methods were empirically tested on the data provided by Schmettow [23]. The infusion pumps were in early-stage development, and an additional re-design phase (for

fixing the usability problems discovered) was planned; this explains why a large number of unique problems (107) were detected by the 34 participants in the usability study. The data were highly heterogeneous, as evidenced by the overdispersion of the marginal sums. By applying the LNBzt model to the data, Schmettow concluded that there were 15 undiscovered problems (i.e. $m = 122$).

Schmettow's "infpump.Rdata" dataset contained the discovery matrix for 20 randomly selected participants. At this stage, 90 problems had been detected. We performed an "as-if" analysis, and applied all the methods to these data. As expected, the six methods that did not take account of heterogeneity found that the discovery process was complete, i.e. there were no problems left to discover. Conversely, the methods that took account of heterogeneity predicted that various problems had yet to be detected. The LNBzt method and the maximum likelihood method predicted $m = 106$ (95%CI: 90 to 123) and $m = 104$ (95%CI: 90 to 121), respectively. With a high number of problems, the two methods produce similar estimates and confidence intervals. The true number of problems with the pump was not known because a re-design cycle was carried out after 34 participants had tested the device. However, the breadth of the confidence interval emphasizes the uncertainty of the two estimates.

IV. Discussion

We decided to model the whole discovery matrix, and thus took account of all the available information. The estimation problem was considered simultaneously in terms of the probability of problem detection and the number of problems. Since the experimental conditions in real-life usability studies are unknown and no one method outperforms the others, it is hard to determine which method is the most reliable.

Most of the currently available methods assume that the probability of detection is the same for all problems. This assumption is likely to be wrong, since real data show that the probability of detection varies [23]. Furthermore, ignoring heterogeneity is known to strongly bias the results [22, 33]. We therefore extended our method to account for heterogeneity in the probability of problem discovery p by using a logit-normal prior to model this uncertainty. The choice of this distribution was convenient in that it allowed us to compare our method with the only published model that accounts for heterogeneity. However, there are no data for confirming the validity of this choice. In fact, the broad confidence intervals observed with the infusion pump data could be explained by the choice of an incorrect prior for p. Nevertheless, this limitation could be easily overcome by replacing the logit-normal by another distribution (such as beta or gamma) if it proves to be more appropriate. This choice could be made using model choice criteria (e.g. the Akaike information criterion or the Bayesian information criterion). However, it should be borne in mind that for a small sample size, fitting for both incompleteness and heterogeneity is complex and inevitably leads to a high degree of uncertainty.

There are two key moments in medical device development for assessing the best method. Early in the development cycle, the device is not mature; usability studies are referred to as "formative" because many usability problems are being discovered and corrected in an iterative design improvement process. Just before market access, usability studies are referred to as "validation" studies or "summative" studies; they are performed on the final version of the device to ensure that no critical usability problems remain [1, 2].

The number of participants in the validation study is an important parameter for both the regulatory authorities and the device manufacturer. Indeed, a sufficient sample size will (i) guarantee the medical device's compliance with the safety standards required for

market authorization, and (ii) avoid a "black swan" effect that would strongly affect the manufacturer's credibility and profitability [34]. The validation study focuses on the detection of infrequent usability problems, corresponding to the "top left corner" of our results. The US Food and Drug Administration requires a minimum of 15 participants [1]. This minimum is based on a naïve estimate, which has been proven to dramatically underestimate the true number of usability problems for this number of participants [10]. Indeed, the coverage probability observed in our simulations for $p = 0.05$ is below 50%. Furthermore, this threshold does not consider heterogeneity in the probability of problem detection. Our findings suggest that a minimum of $n = 20$ participants would markedly increase the reliability of estimating m. Once this threshold in the number of participants is reached, the need to continue the usability study should be assessed by applying matrix-based methods; the latter have excellent statistical properties with as few as 20 participants and in both homogeneous and heterogeneous contexts.

Since the validation study only concerned problems that are probably less frequent, one could question the need to use methods that account for a heterogeneous probability of problem detection. In fact, problems are expected to be "homogeneously rare". Hence, in the absence of heterogeneity and when the sample size is small ($n \leq 10$), the Lewis method (which is conservative for decision-making, with a coverage probability >90% in most cases) could be of great value because it is easy to implement and minimizes the underestimation of m. To the best of our knowledge, however, this assumption of homogeneity for rare problems has not been studied. Furthermore, the human factor engineers' choice of experimental conditions is likely to introduce a degree of heterogeneity by facilitating the detection of problems described in the literature (i.e. in the risk analysis). Thus, a validation study should include a minimum of 15 to 20 participants if it is to produce a relevant estimate with the matrix-based method.

The choice is more obvious for “formative” studies. In our simulations, the “formative” study corresponds to a setting in which usability problems are frequent and numerous. Schmettow’s usability study of a medical infusion pump is also an example of a formative assessment because it was followed by a redesign. Here, we proved that matrix-based methods are more reliable and have low bias and high consistency. Although the LNBzt method could be an interesting option in the context of a heterogeneous probability of detection, it suffers from a high number of failures and thus is difficult to use in practice. As in the case of the infusion pump, a reliable estimate from a small number of participants is an economic advantage for the manufacturer, who can shorten redesign cycles, accelerate device development, and hasten market access. The matrix-based methods meet this requirement, since relatively few participants are needed to guarantee good statistical properties. Lastly, prior knowledge from an earlier stage in device development or a formative usability assessment could be embedded in the Bayesian method, in order to increase its accuracy for a small sample size and reduce the overall sample size.

V. Conclusions

Our method should be applied by regulators and device manufacturers to estimate the number of usability problems using the set of statistical routines provided.

VI. List of abbreviations

LNBzt: logit normal binomial zero truncated

GT: Good-Turing

VII. References

1. CDRH. Applying Human Factors and Usability Engineering to Medical Devices - Guidance for Industry and Food and Drug Administration Staff. 2017.

2. MHRA. Human Factors and Usability Engineering – Guidance for Medical Devices Including Drug-device Combination Products. 2017;:47.

3. FDA. Medical Device Recall Report FY2003 to FY2012. Cent Devices Radiol Health. 2012.

4. Lewis JR. Sample Sizes for Usability Studies: Additional Considerations. Hum Factors J Hum Factors Ergon Soc. 1994;36:368–78.

5. Kanis H. Estimating the number of usability problems. Appl Ergon. 2011;42:337–47.

6. Lewis JR. Evaluation of Procedures for Adjusting Problem-Discovery Rates Estimated From Small Samples. Int J Hum-Comput Interact. 2001;13:445–79.

7. Hertzum M, Jacobsen NE. The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. Int J Human–Computer Interact. 2001;13:421–43.

8. Schmettow M. Sample size in usability studies. Commun ACM. 2012;55:64–70.

9. Borsci S, Londei A, Federici S. The Bootstrap Discovery Behaviour (BDB): a new outlook on usability evaluation. Cogn Process. 2011;12:23–31.

10. Faulkner L. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. Behav Res Methods Instrum Comput. 2003;35:379–383.

11. Lewis JR. Overestimation of p in problem discovery usability studies: How serious is the problem. Tech. Rep; 2000.

12. Sauro J, Lewis JR. Quantifying the User Experience: Practical Statistics for User Research. Morgan Kaufmann; 2016.

13. Thomas DG, Gart JJ. Small Sample Performance of Some Estimators of the Truncated Binomial Distribution. J Am Stat Assoc. 1971;66:169–77.

14. Virzi RA. Refining the Test Phase of Usability Evaluation: How Many Subjects Is Enough? Hum Factors J Hum Factors Ergon Soc. 1992;34:457–68.

15. Nielsen J, Landauer TK. A mathematical model of the finding of usability problems. In: Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems. ACM; 1993. p. 206–213. http://dl.acm.org/citation.cfm?id=169166. Accessed 28 Jul 2016.

16. Good IJ. The population frequencies of spieces and the estimation of population parameters. Biometrika. 1953;40:237–64.

17. Schmettow M. Controlling the usability evaluation process under varying defect visibility. In: Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology. British Computer Society; 2009. p. 188–197. http://dl.acm.org/citation.cfm?id=1671034. Accessed 29 Jun 2016.

18. Finney DJ. The Truncated Binomial Distribution. Ann Eugen. 1947;14:319–28.

19. Rider PR. Truncated Binomial and Negative Binomial Distributions. J Am Stat Assoc. 1955;50:877–83.

20. Shah SM. The Asymptotic Variances of Method of Moments Estimates of the Parameters of the Truncated Binomial and Negative Binomial Distributions. J Am Stat Assoc. 1961;56:990–4.

21. Schmettow M. Heterogeneity in the usability evaluation process. In: Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction-Volume 1. British Computer Society; 2008. p. 89–98. http://dl.acm.org/citation.cfm?id=1531527. Accessed 29 Jun 2016.

22. Caulton DA. Relaxing the homogeneity assumption in usability testing. Behav Inf Technol. 2001;20:1–7.

23. Schmettow M, Vos W, Schraagen JM. With how many users should you test a medical infusion pump? Sampling strategies for usability tests on high-risk systems. J Biomed Inform. 2013;46:626–41.

24. DasGupta A, Rubin H. Estimation of binomial parameters when both n, p are unknown. J Stat Plan Inference. 2005;130:391–404.

25. Fisher RA. The Negative Binomial Distribution. Ann Eugen. 1941;11:182–7.

26. Haldane JBS. The Fitting of Binomial Distributions. Ann Eugen. 1941;11:179–81.

27. Carroll RJ, Lombard F. A Note on N Estimators for the Binomial Distribution. J Am Stat Assoc. 1985;80:423–6.

28. Olkin I, Petkau AJ, Zidek JV. A Comparison of n Estimators for the Binomial Distribution. J Am Stat Assoc. 1981;76:637–42.

29. Hall P. On the Erratic Behavior of Estimators of N in the Binomial N, p Distribution. J Am Stat Assoc. 1994;89:344–52.

30. Wald A. Note on the Consistency of the Maximum Likelihood Estimate. Ann Math Stat. 1949;20:595–601.

31. Efron B. The Jackknife, the Bootstrap, and Other Resampling Plans. SIAM; 1982.

32. Robert C. The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation. Springer Science & Business Media; 2007.

33. Woolrych A, Cockton G. Why and when five test users aren't enough. In: Proceedings of IHM-HCI 2001 conference. Citeseer; 2001. p. 105–108.

http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.85.7896&rep=rep1&type=pdf. Accessed 1 Aug 2016.

34. Bias RG, Mayhew DJ, editors. Cost-Justifying Usability, Second Edition: An Update for the Internet Age, Second Edition. 2 edition. Amsterdam ; Boston: Morgan Kaufmann; 2005.

VIII. Appendix. Proof of the necessary and sufficient conditions for the degeneracy of the maximum likelihood estimator

The likelihood $\ell(m)$ can be written

$$\ell(m) = \sum_{k=0}^{j-1} \ln(m-k) + \mathbb{x}_{\bullet\bullet} \ln\left(\frac{\mathbb{x}_{\bullet\bullet}}{nm}\right) + (nm - \mathbb{x}_{\bullet\bullet}) \ln\left(1 - \frac{\mathbb{x}_{\bullet\bullet}}{nm}\right).$$

We will first prove that $j = x_{\bullet\bullet}$ is a necessary condition for the degeneracy of the maximum likelihood estimator (MLE), and we will then show that it is also a sufficient condition.

The condition that is necessary for degeneracy of the likelihood

We will show that a necessary condition for the likelihood to have a maximum of $+\infty$, (subsequently referred to as a degenerate MLE) is $j = x_{\bullet\bullet}$, i.e. each user has discovered at most one problem and that all the discovered problems are different.

First note that the function $\ell(m)$ can be derived with regard to m for $m \geq j$ (subsequently denoted as $\ell'(m)$). The necessary condition for a degenerate likelihood is that $\ell'(m) > 0$ for $m \in v(+\infty)$.

$\ell'(m)$ is expressed as follows:

$$\ell'(m) = \sum_{k=0}^{j-1} \frac{1}{m-k} + n \ln\left(1 - \frac{\mathbb{x}_{\bullet\bullet}}{nm}\right).$$

By putting $u = \frac{1}{m}$, we obtain:

$$\ell'\left(\frac{1}{u}\right) = \sum_{k=0}^{j-1} \frac{u}{1-ku} + n \ln\left(1 - \frac{\mathbb{x}_{\bullet\bullet}}{n} u\right).$$

Then, by making a second-order Taylor expansion for $u \in v(0^+)$ and using basic algebra, we obtain:

$$\ell'\left(\frac{1}{u}\right) = (j - \mathbb{x}_{\bullet\bullet})u + \left(\frac{(j-1)j}{2} - \frac{\mathbb{x}_{\bullet\bullet}^2}{2n}\right) u^2 + o(u^2),$$

and then:

$$\ell'(m) = (j - \mathbb{x}_{\bullet\bullet}) \times \frac{1}{m} + \left(\frac{(j-1)j}{2} - \frac{\mathbb{x}_{\bullet\bullet}^2}{2n}\right) \times \frac{1}{m^2} + o\left(\frac{1}{m^2}\right).$$

Thus, a necessary condition for $\ell'(m) > 0$ when $m \in v(+\infty)$ is that $j \geq \mathbb{x}_{\bullet\bullet}$, which is only possible when $j = \mathbb{x}_{\bullet\bullet}$ since $j \leq \mathbb{x}_{\bullet\bullet}$ by definition.

Moreover, as soon as $n \geq 3$, $\left(\frac{(j-1)j}{2} - \frac{\mathbb{x}_{\bullet\bullet}^2}{2n}\right) > 0$, which may suggest that this condition is sufficient.

The condition that is sufficient for degeneracy of the MLE

Above, we showed that $j = \mathbb{x}_{\bullet\bullet}$ was necessary for a degenerate MLE. However, we now have to prove that the MLE is effectively degenerate if this is the case. The aim will be simply to prove that in this case, $\ell(m)$ is an increasing monotonic function.

Firstly, replacement of $\mathbb{x}_{\bullet\bullet}$ by j yields:

$$\ell(m) = \sum_{k=0}^{j-1} \ln(m-k) + j\ln\left(\frac{j}{nm}\right) + (nm - j)\ln\left(1 - \frac{j}{nm}\right).$$

We now show that $\forall m \geq j, \ell'(m) > 0$ and thus that the MLE is degenerate.

Firstly, note that

$$\ell'(m) = \sum_{k=0}^{j-1} \frac{1}{m-k} + n\ln\left(1 - \frac{j}{nm}\right)$$

Next, note that $j \leq n$ and $f(n) = n\ln\left(1 - \frac{j}{nm}\right)$ is an increasing monotonic function of n, $f(n) \geq f(j)$. Consequently:

$$\ell'(m) \geq \sum_{k=0}^{j-1} \frac{1}{m-k} + j\ln\left(1 - \frac{1}{m}\right)$$

Let $b(m)$ denote this lower boundary, $b(m) = \sum_{k=0}^{j-1} \frac{1}{m-k} + j\ln\left(1 - \frac{1}{m}\right)$. We can then show that $\forall m \geq j, b(m) > 0$.

By computing $b(m+1) - b(m)$ and using basic algebra, we obtain

$$b(m+1) - b(m) = -j\left[\frac{1}{(m+1)(m-j+1)} - \ln\left(\frac{m^2}{m^2-1}\right)\right]$$

since

$$-\ln\left(\frac{m^2}{m^2-1}\right) = \int_{m^2-1}^{m^2} -\frac{1}{t}\,dt > \int_{m^2-1}^{m^2} -\frac{1}{m^2-1}\,dt = -\frac{1}{m^2-1} = -\frac{1}{(m+1)(m-1)},$$

yielding

$$\frac{1}{(m+1)(m-j+1)} - \ln\left(\frac{m^2}{m^2-1}\right) > \frac{1}{(m+1)(m-j+1)} - \frac{1}{(m+1)(m-1)} \geq 0,$$

as soon as $j \geq 2$. Consequently $b(m+1) - b(m) > 0$. Moreover, since $b(m) > 0$ for $m \in v(+\infty)$ (as shown when proving the necessary condition), we conclude that $b(m) > 0$ for all $m \geq j$ and thus that $\forall m \geq j, \ell'(m) > 0$.

Consequently $j = \mathbb{x}..$ is a necessary and sufficient condition for degeneracy of the likelihood.

Figures

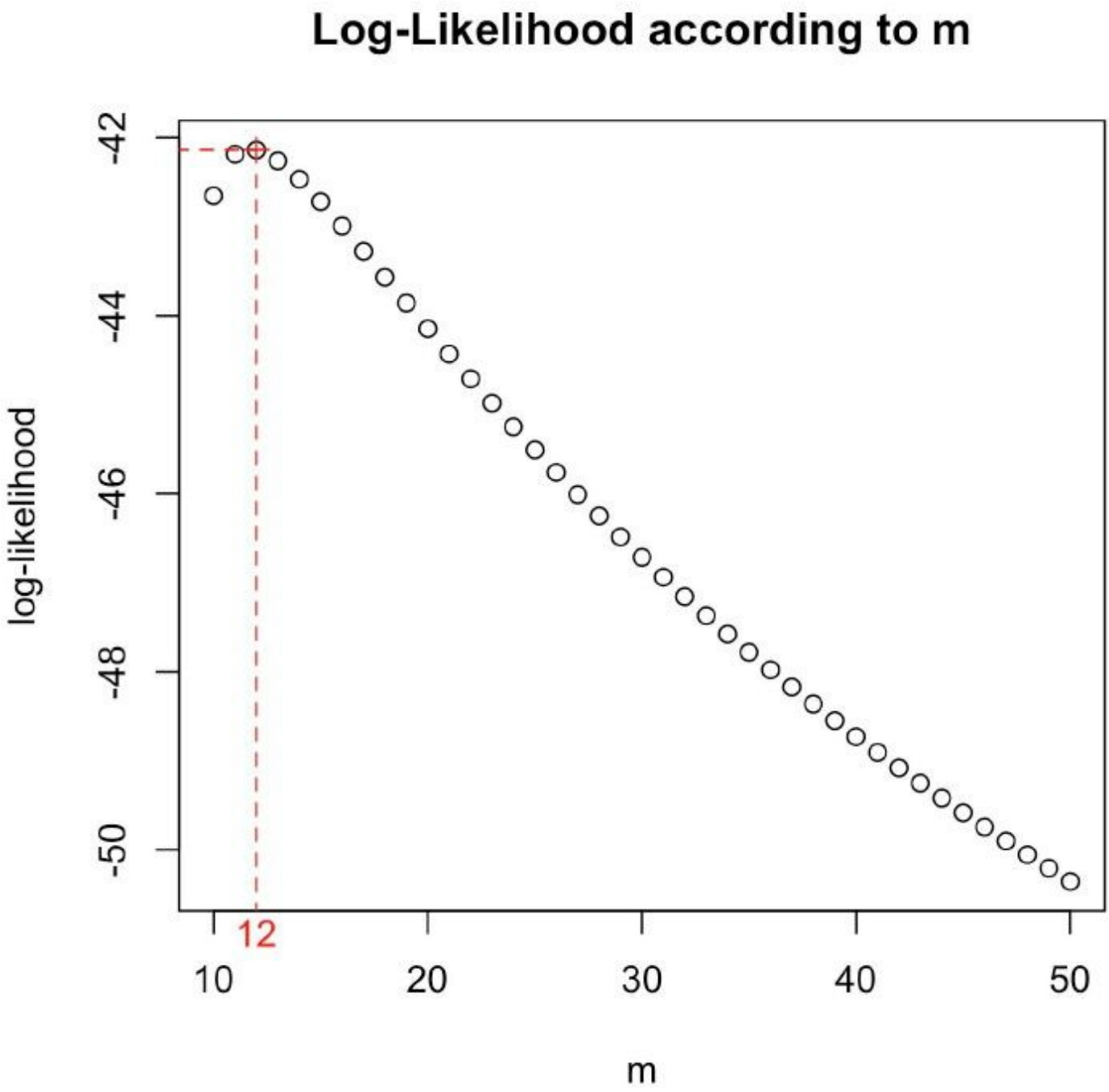

Figure 1

Log-likelihood according to the total number of problems (m), based on the observed discovery matrix (see d in page 6).. The maximum log-log likelihood is reached for m=12. Since the number of observed problems is j=10, the maximum likelihood estimate predicts that 2 usability problems remain undetected.

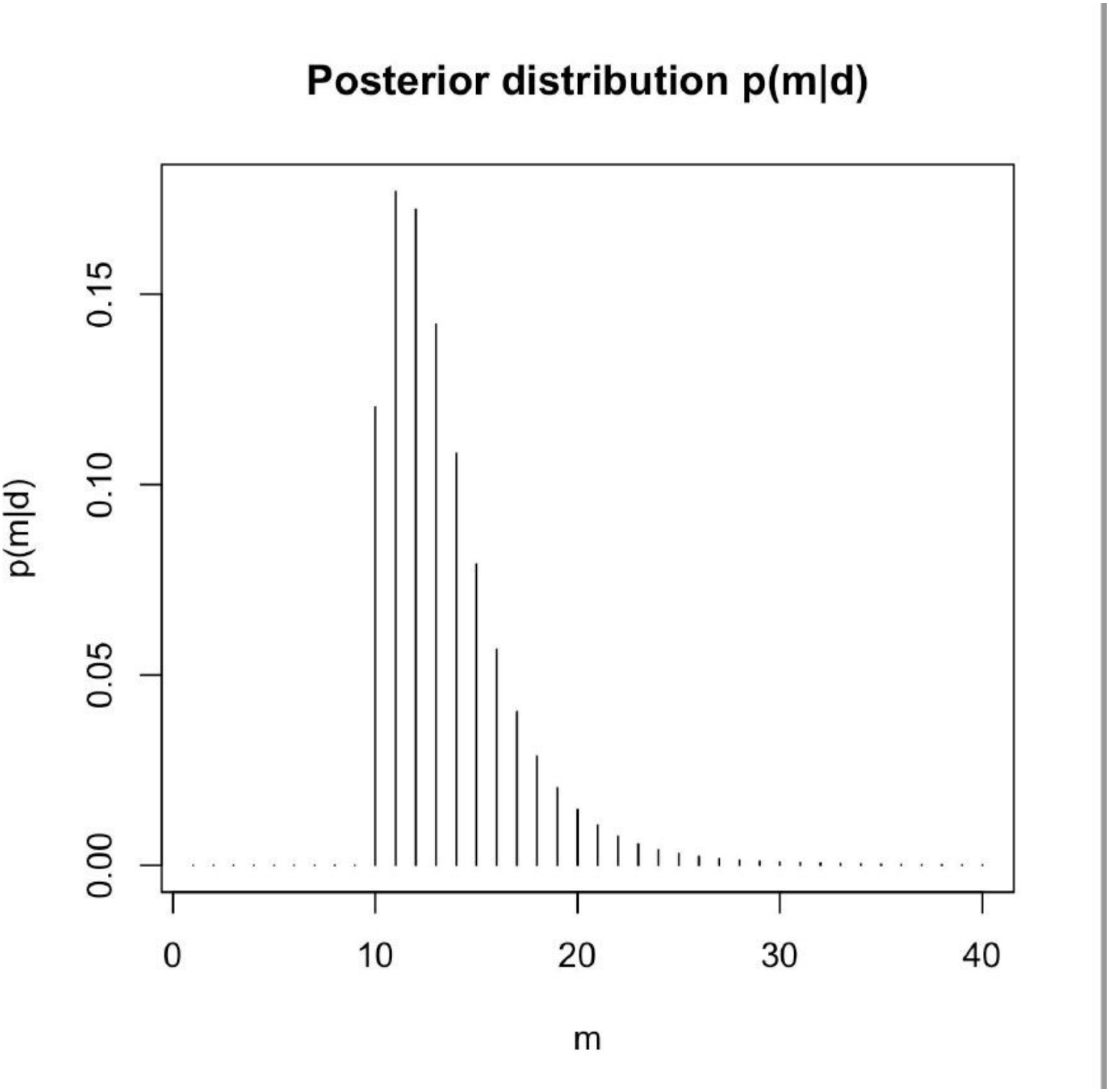

Figure 2

Posterior distribution of m, given the observed discovery matrix (see d in page 6). The distribution mode is at m=11. Since the number of observed problems is j=10, the Bayesian estimate predicts that 1 usability problem has yet to be detected.

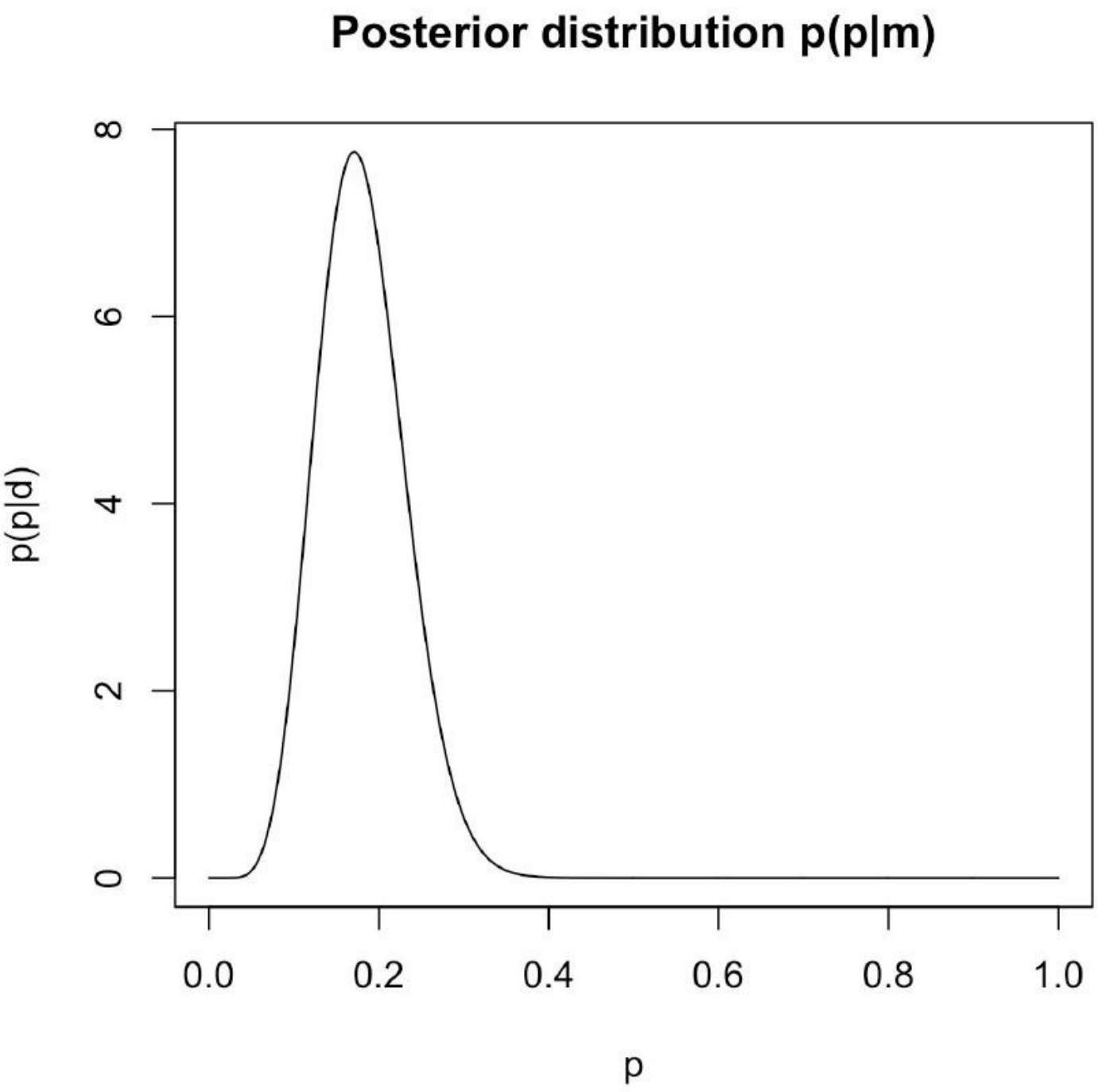

Figure 3

Posterior distribution of p, given the observed discovery matrix (see d in page 6).

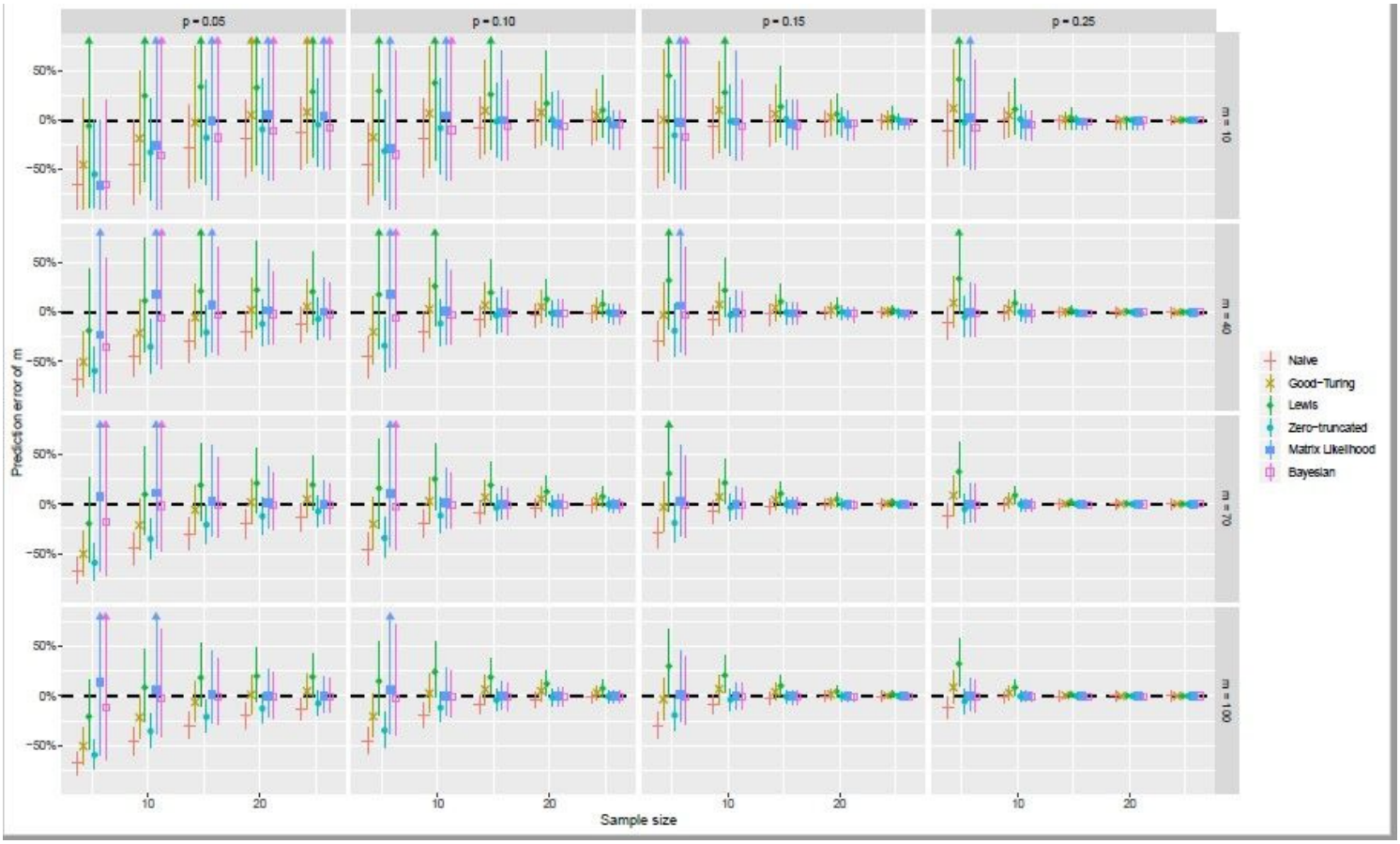

Figure 4

Prediction error for m (mean and 95% fluctuation interval, in %) as a function of the sample size (n), for 6 estimators. The results are presented for various probabilities of detection (p, columns) and various numbers of usability problems (m, rows). The dashed line represents the “true” m. The upper boundaries of credible intervals that exceed 100% are indicated by .

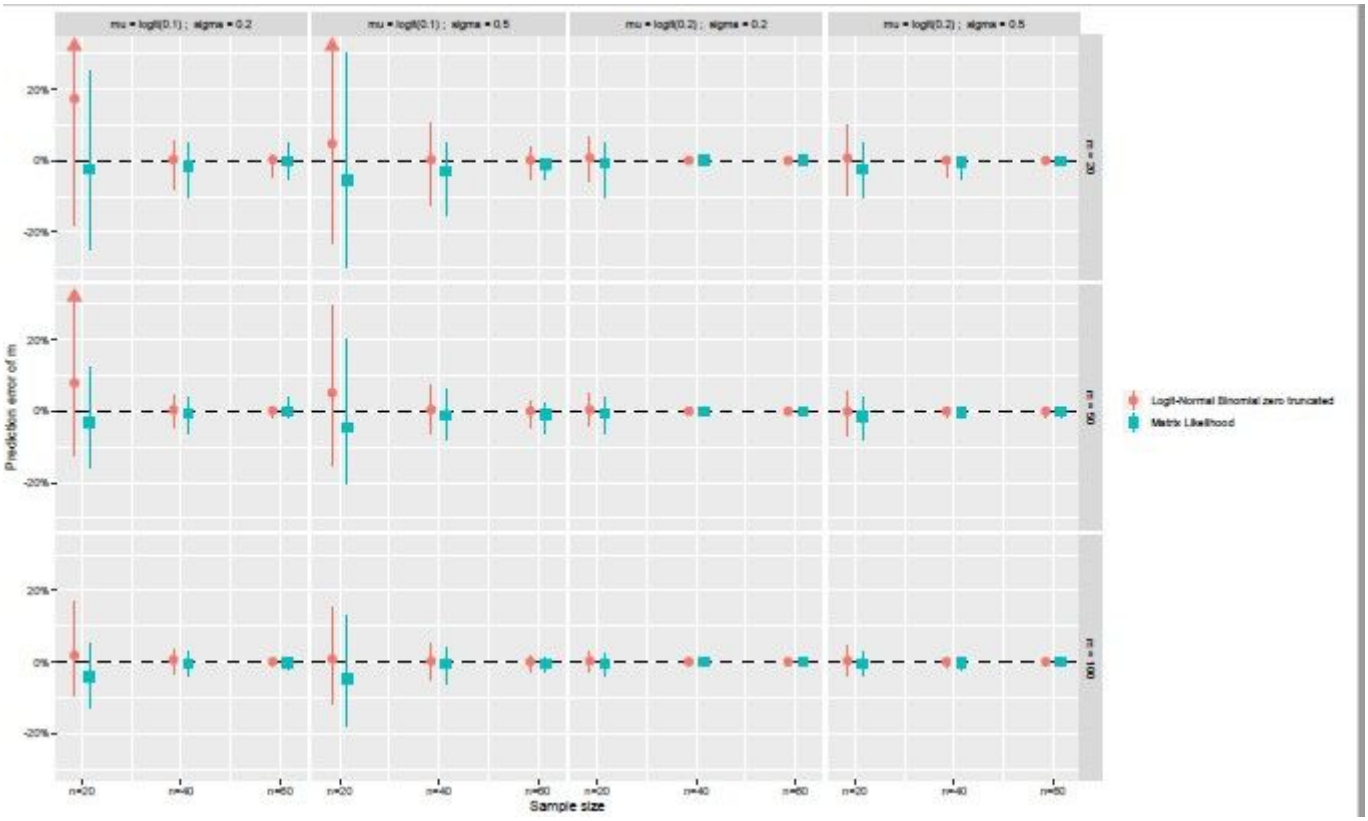

Figure 5

Prediction error of m (mean and 95% fluctuation interval, in %) as a function of the sample size (n) in the context of a heterogeneous probability of problem detection. The results are presented for various probabilities of problem detection ((μ,σ), columns) and various numbers of usability problems (m, rows). The dashed line represents the “true” m. The upper boundaries of the credible intervals that exceed 100% are indicated by .

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Additionalfile1.rar