Reviewing and Extending the Five–user Assumption: a Grounded Procedure for Interaction Evaluation

SIMONE BORSCI, Brunel University
ROBERT D. MACREDIE, Brunel University
JULIE BARNETT, Brunel University
JENNIFER MARTIN, University of Nottingham
JASNA KULJIS, Brunel University
TERRY YOUNG, Brunel University

The debate concerning how many participants represents a sufficient number for interaction testing is well-established and long-running, with prominent contributions arguing that five users provide a good benchmark when seeking to discover interaction problems. We argue that adoption of five users in this context is often done with little understanding of the basis for, or implications of, the decision. We present an analysis of relevant research to clarify the meaning of the five-user assumption and to examine the way in which the original research that suggested it has been applied. This includes its blind adoption and application in some studies, and complaints about its inadequacies in others. We argue that the five-user assumption is often misunderstood, not only in the field of Human-Computer Interaction, but also in fields such as medical device design, or in business and information applications. The analysis that we present allows us to define a systematic approach for monitoring the sample discovery likelihood, in formative and summative evaluations, and for gathering information in order to make critical decisions during the interaction testing, while respecting the aim of the evaluation and allotted budget. This approach – which we call the 'Grounded Procedure' – is introduced and its value argued.

Categories and Subject Descriptors: H.1.2 [**Models and Principles**]: User/Machine Systems; H.5.2 [**Information Interfaces and Presentation**]: Evaluation/methodology; Theory and methods.

General Terms: Measurement, Human Factors

Additional Key Words and Phrases: Five Users Debate, Number of Users, Return on Investment, Usability Testing, User Experience

Author's addresses: Simone Borsci, Brunel University, School of Information Systems Computing and Mathematics, Kingston Lane, Uxbridge, Middlesex UB8 3PH, UK. E-mail: simone.borsci@brunal.ac.uk. Robert Macredie, Brunel University, School of Information Systems Computing and Mathematics, Kingston Lane, Uxbridge, Middlesex UB8 3PH, UK. E-mail: robert.macredie@ brunel.ac.uk . Julie Barnett, Brunel University, School of Information Systems Computing and Mathematics, Kingston Lane, Uxbridge, Middlesex UB8 3PH, UK. E-mail: julie.barnett@brunel.ac.uk. Jennifer Martin, The University of Nottingham, University Park, Nottingham NG7 2RD, UK. E-mail: jennifer.martin@nottingham.ac.uk. Jasna Kuljis, Brunel University, School of Information Systems Computing and Mathematics, Kingston Lane, Uxbridge, Middlesex UB8 3PH, UK. E-mail: Jasna.kuljis@brunel.ac.uk. Terry Young, Brunel University, School of Information Systems Computing and Mathematics, Kingston Lane, Uxbridge, Middlesex UB8 3PH, UK. E-mail: terry.young@brunel.ac.uk.

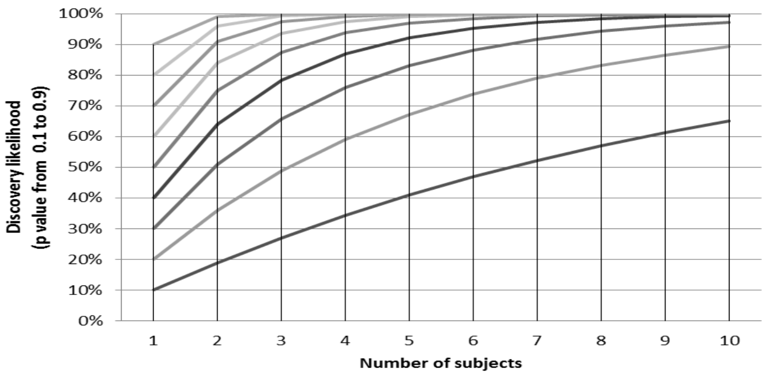
1. INTRODUCTION

The recruitment and selection of subjects (i.e., users and experts) for usability tests, together with the minimum number of subjects required to obtain a set of reliable data, is a hotly-debated topic in technology evaluation [Lewis 1994; Lewis 2006; Turner et al. 2006; Virzi 1992]. In the field of Human-Computer Interaction (HCI), such evaluation is well-defined and integrated into the design process, and is used to ascertain the interaction properties of a given technology at reasonable cost and effort. For the purposes of this paper, we call the number of subjects, *N*, and the total percentage of errors or problems identified by the cohort of subjects, *D*. The discovery likelihood, *p*, denotes the average percentage of errors discovered by an expert, or of problems identified by users. The underpinning equation in evaluating error is focused on how many errors or problems remain undiscovered after *N* subjects have evaluated the product [Nielsen and Landauer 1993] :

$$D=1-1-pN \qquad (1)$$

We term this the *Error Distribution Formula*. The challenge in all of this is that neither *p* nor *D* is known, although clearly given one, the other can be readily calculated. This leaves those wishing to evaluate the usability of a product or service the inverse problem of whether N subjects have identified a sufficient number of problems to ensure that a given threshold percentage, D_{th}, of the total errors or problems has been exceeded in the evaluation.

A straightforward approach to this has been to consider how many new errors or problems that each new subject identifies – there are individual differences amongst subjects, meaning that different subjects will not necessarily find the same errors. This approach to equation 1 is the Return on Investment (ROI) model proposed by Nielsen and Landauer [1993], which assumes stochastic independence of the subjects in their evaluation of the product. Figure 1 shows the increase in *D* with increasing subjects for different average values of *p* from 0.10 to 0.90 for the model.



Number of Subjects										
p value	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
0.10	10%	19%	27%	34%	41%	47%	52%	57%	61%	65%
0.20	20%	36%	49%	59%	67%	74%	79%	**83%**	**87%**	**89%**
0.30	30%	51%	66%	76%	**83%**	**88%**	**92%**	**94%**	**96%**	**97%**
0.40	40%	64%	78%	**87%**	**92%**	**95%**	**97%**	**98%**	**99%**	**99%**
0.50	50%	75%	**88%**	**94%**	**97%**	**98%**	**99%**	**99.9%**	**99.9%**	**99.9%**
0.60	60%	**84%**	**94%**	**97%**	**99%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**
0.70	70%	**91%**	**97%**	**99%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**
0.80	**80%**	**96%**	**99%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**
0.90	**90%**	**99%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**	**99.9%**

Fig. 1. The discovery likelihood of a hypothetical sample of 10 increasing the *p*-values from 0.10 to 0.90 (in this article we use the term *p*-value to refer to the value of *p*, the discovery likelihood, in estimation models).

The rationale of the ROI model is to infer the number of subjects required to exceed the threshold, D_{th}, based on the number of new problems or issues identified by each additional subject. Key to this is the analysis of the smaller increments in these problems/errors discovered by new subjects given a higher number of subjects who have already identified problems/errors. Thus, the experiment tends to an asymptotic saturation level where all of the errors have been found. In this context, and given an average *p*-value for each subject of around 0.30 (estimated as the optimal solution for website tests by a set of multiple empirical analyses [Nielsen and Landauer 1993]) identifying the first 80% of the problems/errors requires five subjects, while the next 19.5% requires a further 10. This represents a gain of less than a quarter, while trebling the evaluation cost/effort – a situation shown in Figure 2. As an important aside here, seeing 'gain' in this way treats all errors as equally critical, a situation which is often not the case. While it has historically been common practice for usability evaluations to focus on percentage of errors discovered as a key metric, in practice the nature of the errors is critical. It makes a significant practical difference if, when finding 80% of errors, the remaining 20% are critical problems rather than minor issues.

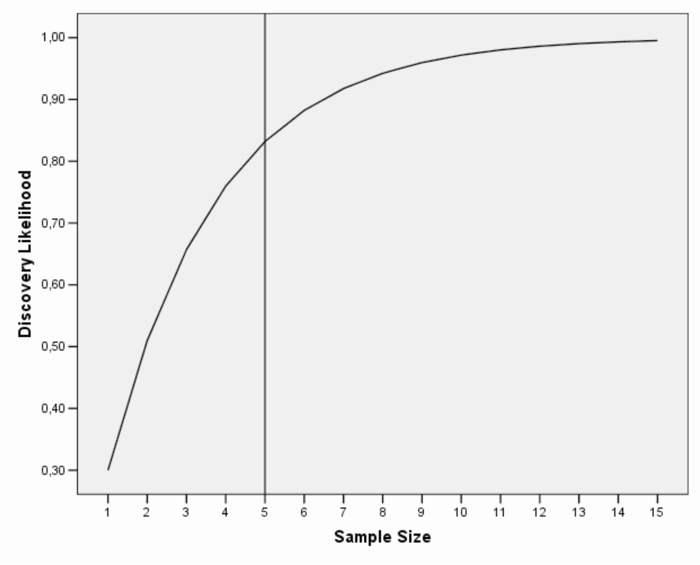


Fig. 2. The asymptotic behavior of discovery likelihood of a hypothetical sample with p=0.30.

Nielsen and Landauer [1993] applied the ROI model to analyzing data sets from verbal protocol techniques and expert-based evaluations and showed that a sample composed of a range of between three and five subjects with appropriate skills was generally enough to assess an interaction with a web interface and identify at least 80% of the interface problems. This result has been confirmed by several studies [Nielsen 1995; Nielsen 2000; Nielsen 2012; Nielsen and Landauer 1993; Virzi 1990; Virzi 1992] and is known as the 'five–user assumption' [Nielsen and Landauer 1993]. This model carries the latent assumption that *p* averages around 0.30 where an acceptable threshold, D_{th}, of 80% of interface problems are found. Although *D* is not highly-sensitive to variations in *p* so long as the average value for the cohort is used, there is an obvious complication. The perception of how close one is to identifying 100% of the issues or problems is sensitive to the order in which the 'new' subjects' findings are added to the results of the cohort, if the subjects' characteristic *p*-value varies. The approach to a false asymptote, for instance, could readily be presented, by listing the findings of higher *p*-value subjects before those of lower *p*-value subjects.

Nielsen (2000) invoked the five-user assumption in order to explain the ROI model, but from a historical point of view it has guided practitioners in scoping their evaluations. Today, the five-user assumption is widely used, and still recommended for many cases [see Nielsen 2012], and also regularly condemned, suggesting that there is a need to revisit its application to set it on a firmer basis. The aim of this paper is therefore to survey the approaches to the equation behind the five-user assumption, to examine its use in evaluation, and to explore both the limits and advantages of the different approaches. Following this, we propose a novel pragmatic, or grounded, procedure to support evaluators in managing and monitoring the number of problems discovered by a sample of users when they are testing a prototype during its development lifecycle (i.e., formative evaluation), or when they are testing a product at an advanced stage of its development or after it has already been released on to the market (i.e., summative evaluation) [Kirakowski 2005].

2. HISTORICAL BACKGROUND

Over the years, the HCI community has sought different ways to estimate the *p*-value, usually by re-examining the contribution made by each subject in a controlled trial, and then by extrapolating or reconfiguring the findings. The aim has been mainly to control the cost of evaluation by keeping the number of subjects to a minimum within the constraint of exceeding a notional threshold for the discovery of problems/errors. For this reason, the estimation of the *p*-value was initially widely seen as an issue relevant to the cost-benefit analysis of, for example, web interfaces [see Bias and Mayhew 2005], helping to make the return on investment case to justify the cost of usability assessment. While the reduction of assessment costs remains a key outcome associated with *p*-value estimation, the debate has become increasingly focused on the reliability of the data gathered when using a small sample of users [Borsci et al. 2012; Nielsen 2012; Sauro and Lewis 2012; Schmettow 2012]. The approach taken to *p*-value estimation is also relevant to both issues of cost and reliability, with different approaches having focused on such factors as the order of the subjects in the evaluation [Nielsen and Landauer 1993], the nature of the errors and problems identified by the sample [Turner, Lewis and Nielsen 2006], and the properties of the interface [Borsci et al. 2011].

From the 1960s to the 1980s two main barriers prevented developers from adopting a systematic approach to evaluation in the design cycle. The first was the idea that assessment was only a verification process and was separate from the design process. This led to the second issue, which was that developers considered the cost of evaluation to be additional to, and not a critical part of, the design process. By the 1980s, designers and researchers had begun to experiment with the concept of a 'simulated user', looking for the most effective and efficient principles that would guide interface-developers [Molich and Nielsen 1990; Shneiderman 1986]. Focused on the artifact being evaluated, this approach was developed by simulating the users' needs, and defined by Kurosu [2007] as the 'small usability approach', where evaluation was only a secondary step to verify the quality of the design and of the artifact's functionality.

By contrast, Norman [Norman 1983; Norman 1988; Norman and Draper 1986] proposed a new design philosophy at the end of the 1980s, User Centered Design (UCD), which integrated both design and evaluation by focusing on the properties of the interface needed to meet the users' needs. Designers eschewed this approach for a period, considering it an ideal rather than a pragmatic or even a necessary way forward. This behavior was justified on the grounds that developers were focused on controlling the costs of design, and were therefore looking for low-cost techniques instead of what might be seen as a 'grand scheme'.

This raised interesting questions about the cost of design in the context of whole-life costs, but it was not until 1998 that the International Organization for Standardization (ISO) explicitly endorsed the UCD process under ISO 9241-11[1998], after which designers were forced to adopt a new perspective. Kurosu [2007] identified this as the 'big usability approach', in which the evaluation is fully integrated into the product development cycle in the context of user needs. This ISO standard also presents a perspective from which a single technique is no longer sufficient to evaluate usability on its own, but where such techniques become part of a multidimensional construct. Following this approach, a practitioner applies different evaluation techniques and tools and involves the final users in interaction assessment. As ISO 9241-11 was widely taken up, it supported developers by creating a framework for evaluation while increasing costs by mandating the involvement of users.

The ROI approach to the Error Distribution Formula, at least until 2001, was seen as the only reliable way to comply with the standard while managing costs. In 2001, a series of studies started to challenge the ROI model and its five-user assumption, splitting the evaluation community into two broad camps: those who seem to accept and apply the model; and those who are, to varying degrees, critical of it even if they use it in their research.

2.1. Views on the five-user assumption

Spool and Schroeder [2001] provided one of the first studies post Nielsen [2000] that reflected on the five-user assumption, reporting an experiment in which they found that five users were far too few to reach the threshold discovery percentage suggested by Nielsen. They described an evaluation of four web sites by 49 subjects, reporting that to identify more than 85% of the problems required considerably more than five subjects. Until then, the five-user assumption had been generally accepted as a reliable guideline. Almost a decade later, Alshamari and Mayhew [2009] contrasted Nielsen's [2000] expectation that five users would unearth 80-85% of the issues with Lindgaard and Chattratichart's [2007] study that identified only 35%, showing on-going concern about the five-user assumption. Indeed, studies that identify a lower discovery rate than Nielsen's [2000] expectation are now relatively common.

Of those studies that demonstrate discovery rate issues, Faulkner's [2003] exploration of discovery rate appears relatively comprehensive, reporting an evaluation of a website interface by 60 subjects and then with sub-samples of five, 10 and 15 users, up to 55 subjects. Faulkner concluded that "the risk of relying on any one set of five users was that nearly half of the identified problems could have been missed; however, each addition of users markedly increased the odds of finding the problems" [Faulkner 2003]. It is difficult to determine how much weight should be given to the outcomes of this study since the primary data is not available for detailed analysis by other researchers. Further, the study did not make any connection between the average discovery likelihood and the likely percentage of discovered problems. Yet, despite these criticisms, Faulkner's study has been highly influential, especially with the US Food and Drug Administration (FDA) which recently included it in their draft guidance on medical device testing [FDA 2011] as an appendix entitled: "Considerations for Determining Sample Sizes for Human Factors Validation Testing". The FDA guidance recommends a sample of 15 subjects to find a minimum of 90% and an average of 97% of all problems [FDA 2011], and it is interesting to note the application of Faulkner's [2003] findings well beyond mainstream HCI.

The continued interest in, and diverging views on, the five-user assumption in relation to the ROI model can be seen through an analysis of research that has cited Nielsen and Landauer's [1993] original work. Fifty post-2001 citations of Nielsen and Landauer [1993], were identified using Google Scholar, with the citations appearing in peer-reviewed journal articles (56%), conference papers (28%), book sections (14%), and industrial or company reports (2%) (as of 1 March 2012, the latest citation being from 2009).

Figure 3 shows the number of 'adopters' and 'critics' in the citation sample. The adopters group contains 31 pieces of work that have simply adopted the five–user assumption in their studies. In many of these cases, there is an acknowledgement of the limitations associated with the five-user assumption, but it is adopted regardless. An example of this can be seen in Crystal and Greenberg [2005] who state that: "This sample size is not intended to yield definitive results, but models of usability testing [Nielsen and Landauer 1993] suggest that testing with five users is sufficient to uncover most usability problems". The 19 critical references are more explicit in raising the shortcoming of the five-user assumption and draw on sources that have raised serious concerns about its validity. An example of this is seen in Hong, Heer, Waterson, and Landay [2001], who state that: "Despite previous claims that about five participants are enough to find the majority of usability problems [Nielsen and Landauer 1993; Virzi 1992], a recent study by Spool and Schroeder [2001] suggests that this number may be nowhere near enough"

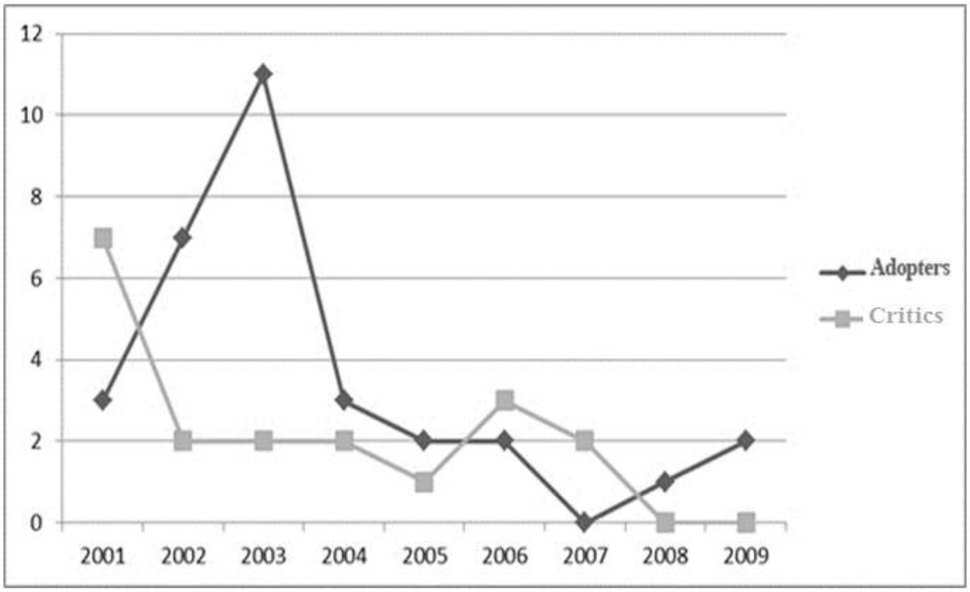


Fig. 3. Analysis of 50 2001 to 2009 citations of Nielsen and Landauer's [1993] work obtained using Google Scholar on 1 March 2012. The 'adopters' group comprises 31 references, while the 'critics' group comprises 19 references.

Of the 50 papers identified, 78% were HCI studies while the remaining 22% were related studies from application areas including healthcare, automotive engineering, and business and information management. Of the 11 papers from outside the HCI area, nine were 'adopters' and two were 'critics', demonstrating that the application of the ROI model extends beyond HCI and into justifying the appropriateness of a given sample size for user evaluation in other fields.

The continued interest in the ROI model and five-user assumption, whether seen from the perspective of its continued application in HCI evaluation studies or from the more critical perspective which argues its weaknesses, suggests that it would be valuable to offer improvements to the basis on which the sample size judgment is made when using the ROI model.

2.2. Alternative models for estimating p

Several methodological developments have been proposed to this end, largely focusing on better estimates of *p* and, by association, N. First, there is Good-Turing procedure, originally proposed in the HCI field by Lewis [2001] following the previous formalization of this model by Good in collaboration with Turing [Good 1953; Jelinek 1997; Manning and Schutze 1999]. In Lewis' development of the original model [Lewis 2001; Sauro and Lewis 2012] he sought to normalize the *p*-value estimated by Good-Turing model, including an adjustment in tune with Hertzum and Jacobsen's study [2003].As Turner, Lewis, and Nielsen [2006] suggest this produces a more conservative estimate of p, which is desirable since overestimating *p* leads to the belief that one is closer to D_{th} for a given N than is actually the case. However, this gain comes at the expense of increased computational complexity and less insight. The Good-Turing procedure formula is expressed as follows:

$$padj=12pest1+EN1N+pest-1n1-1n \qquad (2)$$

In equation (2), p_{est} is the initial estimate computed from the raw data of a usability study, $E(N_1)$ is the number of usability problems discovered only once in the evaluation across all users, N is the total number of problems, and n is the number of test participants.

Second, the Monte Carlo re-sampling method is a statistical simulation technique that has been used to simulate the impact of the subjects taking part in the evaluation in different orders [for a review, see: Fishman 1995]. Lewis [1994; Lewis 2000] applied this in conjunction with the Good-Turing procedure and showed that it delivers a more conservative and more reliable value of *p* than the classic ROI model.

Third, the Bootstrap Discovery Behavior model, proposed by Borsci, Londei, and Federici [2011; see also Borsci, Federici, Mele, Polimeno, & Londei, 2012] is another re-sampling method that builds on the Good-Turing and Monte Carlo methods. It adopts a bootstrapping approach [Efron 1979; Fox 2002] and modifies the ROI equation (1) as follows:

$$DL=Mta-1-pL+q \qquad (3)$$

In equation (3), Mt represents the total number of problems in the interface. The value a is the representativeness of the sample expressed as the maximum limit value of problems collected by 5000 possible bootstrap samples (with repetition). The value *p* represents the normalized mean of the number of problems found by each subsample, as the estimated probability of the detection of a generic problem by an evaluator in the chosen population. The q variable expresses the hypothetical condition L = 0 (an analysis without evaluators). In other words, since *D* does not vanish when L = 0, D(0) represents the number of evident problems that can be effortlessly detected by any subject, and q the possibility of detecting a certain number of problems that have already been identified (or are evident to identify) and were not addressed by the designer, as expressed in equation (3a):

$$D0=Mta-1-pq \qquad (3a)$$

The value q represents the properties of the interface from the evaluation perspective, with its extreme value being the 'zero condition' where no problems are found. The Bootstrap Discovery Behavior model (as expressed in equation (3)) enlarges the perspective of analysis by adding two new parameters not considered in equation (1): (i) all the possible discovery behaviors of participants (a); and (ii) a rule in order to select the representative data (q). In this sense, the Bootstrap Discovery Behavior model proposes a modification of the ROI model to include new factors within the sample discovery likelihood estimation.

The modifications to the original ROI model all require assumptions or estimations in order to make use of them for practical interface evaluation studies. As Sauro and Lewis [2012] note, it is possible in the modified models to apply the logit-normal-binomial (LBN) proposed by Schmettow [2009] to estimate the *p* value, and to use the zero-truncated LBN to estimate the number of remaining defects in the product. Although the LBN model has a number of potential applications in HCI, as Schmettow makes clear, the zero-truncated LBN "still makes assumptions and it is unclear how these are satisfied for typical data sets in the wild" [2012]. As such, there are always uncertainties and assumptions in the use of such models, making it critical that evaluators understand their basis and limitations rather than simply using the model and adopting what may have become an established assumption, as seems to be the case for some researchers with respect to the five–user assumption.

The next section will consider the ROI model in more detail, stressing its strengths and weaknesses and using them to frame the steps that an evaluator should take in order to make most effective use of the model in their specific evaluation context. This will lead to the presentation of an approach – the Grounded Procedure – which we argue can guide decision-making in relation to the most suitable user sample size given the evaluation aims and budget constraints.

3. THE CHARACTERISTICS OF THE ROI MODEL AND KEY DECISIONS ASSOCIATED WITH ITS EFFECTIVE USE

As noted in section 1, the ROI model is an application of the Error Distribution Formula (equation 1). This simple formula has several characteristic benefits. First, it is accessible, provides insight into error distribution which can guide evaluation, and is easy to apply. Second, it provides a way into a dialogue about the cost and effectiveness of an evaluation, as is the case through its application in the ROI model. Third, as an established approach, and through the ROI model, it exerts a standardizing influence on the industry, providing an accepted approach to evaluation that is widely used. Fourth, with the five-user assumption it provides a basis for evaluation that is useful in many instances and is better than doing nothing or being paralyzed by lack of knowledge. Fifth, it is relatively insensitive to small variations in parameter values and, providing the discussion is kept broad, yields a coherent, numerate basis for a discussion around the emerging range of users required in a given evaluation context (i.e., it will differentiate clearly between the need for 15, as opposed to 5 users, but not necessarily between 5 and 6).

There are, however, two broad problems with the model: (i) since the extent of the problem-space (or issue-space) is never known, one is always left to apply the model in an inverse fashion; and (ii) its simplicity means that the complexities of the real world may not be considered or identified by evaluators – which may be behind the 'blind adoption' issue in relation to the five-user assumption. Associated with this second point, a range of resultant drawbacks have been articulated in the literature. First, the ROI model is based on the idea that all the subjects exhibit the same probability of encountering usability problems, without reference to their varying skills [Caulton 2001]. Second, in many practical cases there is an additional problem in assuming that all subjects meet or approach the p=0.30 criterion [Lewis 1994; Schmettow 2008; Woolrych and Cockton 2001], with an allied issue being that the ROI model does not address the representativeness of the participants selected for the test. Third, the ROI model does not account for the evaluation methodology or the context of the system being evaluated – it considers all systems as having the same probability for being perceived as problematic by users, while in reality there is variability in the discoverability of resident errors in systems arising, at least to some extent, from differences in system complexity. In fact, there are several concepts bound up in the *p*-value that require unpacking – the ability of subjects to identify issues may also reflect the extent to which they represent the community of users and are representative of them. This, in turn, leads back to the nature of the evaluation being undertaken and draws into question whether a model developed for interface assessment can be applied directly to, for example, medical device evaluation with their specialist cohorts of users. Fourth, the ROI model takes a limited view of discoverability and the philosophical question as to whether errors that are never discovered by anyone are errors at all. Related to this is the 'zero condition' described by Nielsen and Mack [1994].

This discussion brings three factors to the fore. First, the Error Distribution Formula provides an elegant and informative way to assess the number of subjects needed in an evaluation. Second, its very elegance means that it is not difficult to find situations where the requirement for 80% of the issues to be identified by five users who share a value of *p* close to 0.30 do not come together conveniently to meet the five-user assumption when the *p*-value is estimated as an average [Schmettow 2008; Woolrych and Cockton 2001]. Some of the contexts in which the formula has been shown to 'fail' to help developers and evaluators are far removed from the original context. Finally, just as evaluation is a cost-effectiveness exercise, so, too, is the business of estimating the appropriate number of users in that assessment. Therefore, we recognize that there are scenarios in which it is safe to proceed with the five-user assumption. Typically these may be where the cost of errors or undiagnosed issues is low, where the assessors are known to be able to identify most of the issues or problems (i.e., they are characterized by a *p*-value close to, or exceeding, 0.30), or where there are no overriding constraints of safety or success and a decision must be made quickly.

On the other hand, where there are overriding reasons to characterize the evaluation very accurately, there are ways of improving upon the simple formula and customizing the findings to the context and skills of the evaluators. This may also be the case where evaluation frequently takes place, and always within a well-controlled environment, where the cost of the extra process of reviewing the sample size (N) may be set against savings in the evaluation that may be reaped over and over again.

In the latter cases, we note that it would be possible to produce very accurate, well-calibrated formulae, refined and characterized through frequent use, but the cost would be complex algorithms for estimating N and significant data-collection to inform and validate such algorithms.

Figure 4 seeks to bring these constraints together, noting that it may be a good decision to adopt the five-user assumption, but also illustrating where a more nuanced judgment around sample size is required. Within Figure 4 and the subsequent discussion, we assume that where N is mandated (such as by a standard or client), it will be higher than five. As Figure 4 shows, the evaluation decision process considers three main constraints:

—The costs of error identification against the available budget;

—The kind of product and the level of safety required for optimal interaction;

—The external issues that may require evaluation with more than five users.

When none of these constraints affect the evaluation, or when only the cost of error identification is considered to be an important issue (i.e., where only a low evaluation budget is available), the outcomes of the decision process (cases 1 and 4 in Figure 4) support the decision to test the product only with five users, or with two or three different groups composed of three to five users of each kind [Nielsen 2000].

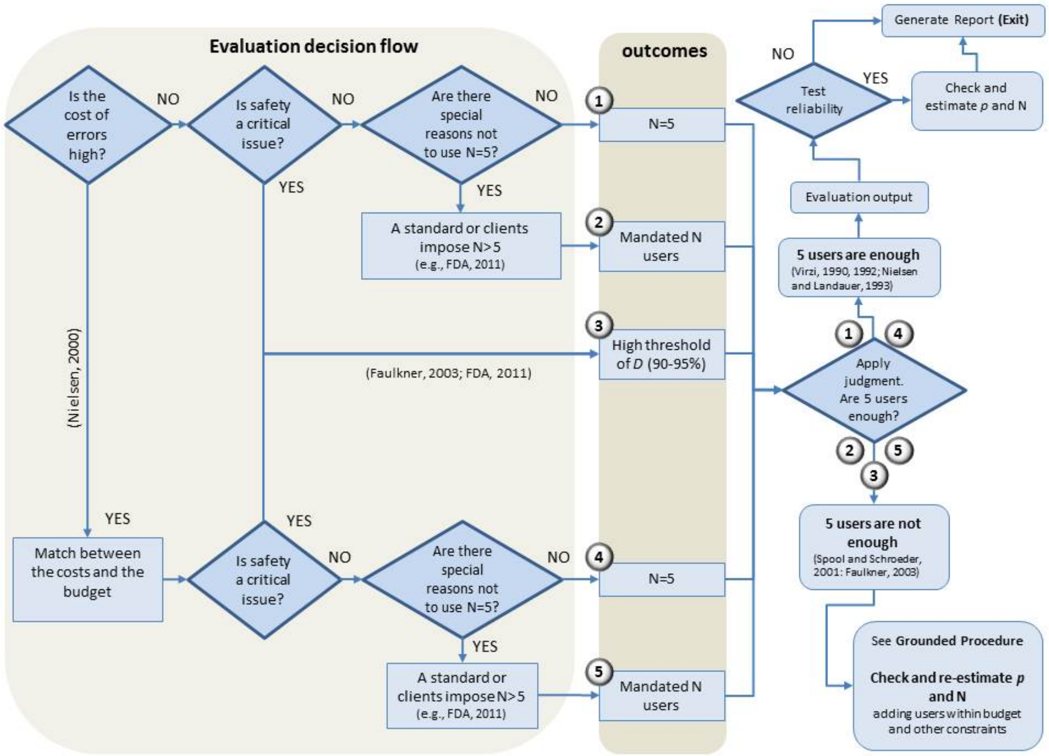

Evaluation decision flow
outcomes
Is the cost of errors high?
NO
Is safety a critical issue?
NO
Are there special reasons not to use N=5?
NO
1
N=5
YES
YES
A standard or clients impose N>5 (e.g., FDA, 2011)
2
Mandated N users
3
High threshold of D (90-95%)
(Faulkner, 2003; FDA, 2011)
(Nielsen, 2000)
YES
YES
Match between the costs and the budget
Is safety a critical issue?
NO
Are there special reasons not to use N=5?
NO
4
N=5
YES
A standard or clients impose N>5 (e.g., FDA, 2011)
5
Mandated N users
Generate Report (Exit)
NO
YES
Test reliability
Check and estimate p and N
Evaluation output
5 users are enough (Virzi, 1990, 1992; Nielsen and Landauer, 1993)
1
4
Apply judgment. Are 5 users enough?
2
5
3
5 users are not enough (Spool and Schroeder, 2001: Faulkner, 2003)
See Grounded Procedure
Check and re-estimate p and N
adding users within budget and other constraints

Fig. 4. The decision process and outcomes associated with using the five-user model, and the identification of cases where an alternative sample should be used.

In this case, a check test of the *p*-value of the sample aims only to identify how many problems have been discovered by the N users actually involved in the evaluation. The client, after the generation of the evaluation report, will have to decide whether to increase the budget in order to improve the reliability of the assessment by running a second evaluation process, adding more users in order to extend the reliability of the evaluation until a certain *p*-value threshold is reached.

In cases where the product being evaluated requires a specific level of safety, the evaluation process must aim to identify a high percentage of interaction problems – i.e., D=90-95% (case 3 in Figure 4). In this case the evaluator cannot assume that five users is a large enough sample, but s/he can use the first five users as a starting point for the assessment, by estimating the *p*-value using the different models introduced in section 2.2 in order to determine information about how many problems are discovered by the actual sample, and if or when the *D* threshold (D_{th}) is reached.

In cases where a N higher than five is imposed on the evaluator (cases 2 and 5 in Figure 4), the estimation of *p* is a necessary step for optimizing and controlling the use of the budget. In such cases D_{th} is not necessarily set higher than the 'standard' 80-85%, but until D_{th} is reached it will be necessary to add new users to the sample and, with each added user, for the evaluator to check the sample *p*-value and monitor the modification of the overall sample distribution after each single assessment.

In summary, as Lewis suggests, each usability testing process can substantially vary in the representativeness of the tested participants, tasks and environments (Lewis, 2010). It follows that the outcomes of a test are reliable only for the specific technology tested under a set of specific conditions (selected users, task and scenarios). For instance, a sample of five people may be enough to test a website in one scenario, but the same sample may not be enough to test the same website in another scenario. Moreover, a sample that discovers a high number of problems in a web site may demonstrate a low level of performance in terms of discoverability with another website or technology.

The theoretical assumption behind the estimation models is that it is not possible to determine empirically all of the existing problems in an interface or in family of products because, for any given interface, the total number of usability problems is unknown. However, if this unknown number is equated to 100% of the possible problems, a practitioner may apply formula (1) to estimate the *p*-values and, by using the different estimation models, can determine a workable estimate of the percentage of the problems that have been identified by a test, and what percentage is likely to remain undiscovered at any point. As we will discuss in the next section, in order to generate the evaluation report when the decision process outcomes require a sample greater than five (i.e., cases 2, 3 and 5 in Figure 4), the evaluator has to follow a specific set of steps that we call the Grounded Procedure. We contend that this procedure is also useful when the five-user assumption is accepted (cases 1 and 4) if the evaluators have, for specific reasons, to test the reliability of their assessment.

4. A GROUNDED PROCEDURE: MONITORING AND ESTIMATING THE SAMPLE DISCOVERY LIKELIHOOD

The Grounded Procedure (GP)[1] resonates with Lewis' contention that: "Practitioners can obtain accurate sample size estimates for problem-discovery goals ranging from 70% to 95% by making an initial estimate of the required sample size after running two participants, then adjusting the estimate after obtaining data from another two (total of four) participants" [2001].

We propose that in order to apply the ROI model in a useful way, practitioners should follow a specific procedure for assessing the evaluation process and then determining from the emerging findings how many more subjects are likely to be needed to meet their evaluation aims. In such cases, the practitioner would need to decide at what point to monitor the changes in sample behavior and p-value, and the levels of heterogeneity of the cohort, to give them the best time-effort trade-off. While, in extremis, this would be done after each additional user, there are clear practical difficulties in scheduling one user at a time. As such, it may be more practical to undertake the monitoring as part of a time-based schedule (such as after a quarter or half day) deciding as a result whether to stop the testing or continue to add new users. Whatever the specific choices made by the practitioner, the GP offers a dynamic process for collecting information about the sample discovery likelihood and taking subsequent decisions about the assessment. This procedure is based on three main assumptions:

[1] GP, though not based on Grounded Theory [Glaser and Strauss 1967] shares with it the idea that practitioners can make inferences and take decisions (in this case about the evaluation) only from the data at hand, that emerge from the observation of the evaluation cohort behavior.

(1) An evaluation (whether formative or summative) is a counterbalanced process in which the evaluators, in light of the aims and available budget, reduce the variability of the possible interaction behavior (i.e., the divergent user experiences), typifying the kinds of user that have to be involved in the assessment (i.e., through selection criteria), the tasks and the goals of the interaction, and the environment of use.
(2) These reductions in the variability, along with the available budget and the evaluation aim (e.g., to identify more than the 85% of the problems), lead practitioners to select specific evaluation techniques to be used and the form of the evaluation process (i.e., summative or formative) so affecting the resulting data that is gathered [see for a complete review: Tullis and Albert 2008].
(3) By monitoring the sample discovery likelihood after the first three or four users, practitioners, as Lewis (2001) suggests, can obtain reliable information about the gathered data in order to determine whether the problems discovered by the sample have a certain level of representativeness (i.e., reliability and quality).

We propose that practitioners start by assuming a specific *p*-value standard (e.g., 0.30 if the aim is to reach the 80-85% of the problems), and use this value as a comparator against which the behavior of the real population of subjects can be assessed. In light of this, practitioners have to compare the *p*-value of their actual tested sample to the standard in order to make the following two main judgments, leading to the associated decisions and actions:

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

This illustrates that the GP consists of an information-seeking process that aims to obtain reliable evidence for deciding whether practitioners have to extend their evaluation by adding users or whether they can stop the evaluation because they have sufficient information. The GP consists of three main steps:

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

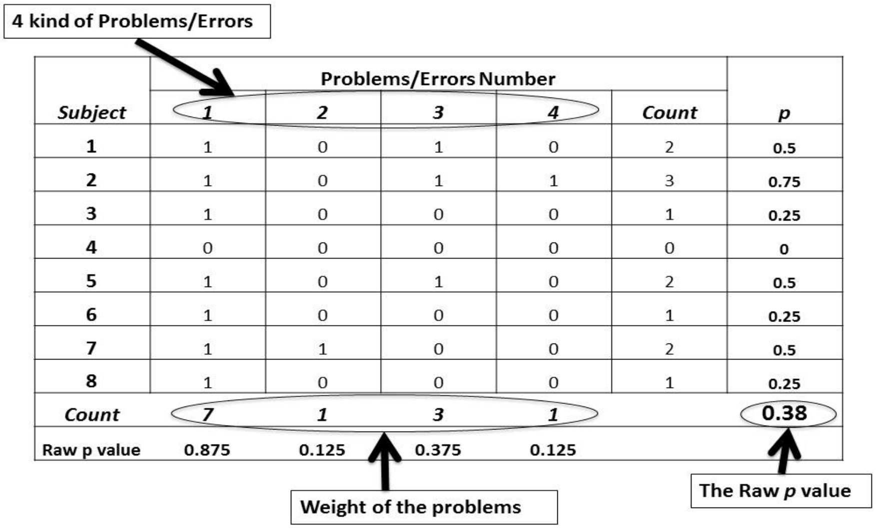
Each of these steps will be discussed in turn, drawing on an example scenario throughout, to provide more detail.

4.1. Monitoring the table of problems/errors

When practitioners run an evaluation using a sample of experts (i.e., an expert-based test), these subjects, simulating the final user group's interaction with the product and following an explicit or implicit user model, identify a certain number of errors related to the technological functioning that could affect, and cause problems in, the final users' experiences. When practitioners assess the interaction with a sample of final users (i.e., a user-based test), these subjects identify interaction problems that may arise from errors in the technological functioning or a mismatch between the designer's and the user's mental models of the product [see for a complete framework on mental models: Norman 1983; Norman 1988]. Whether experts or target users are engaged in the evaluation, the practitioner collects a series of subjects' behaviors each instance of which can be represented in a binary way: i) 0 = Problem/error not found; ii) 1= Problem/error found.

As a consequence, a large group of subjects has a higher probability of identifying a larger number of problems than a small group because a large group has greater scope for divergent behavior than does a small one. In this sense, the aim of any estimation model is not to identify how many users are needed for an evaluation, but to identify the smallest group with the greatest quality of discoverability behavior (that is, *p*-value). In this context, the quality of the behavior is seen as the ability of a small sample to best represent the behavior of a larger sample. We can thus define the representativeness of a sample as the degree to which the problems/errors identified by the sample accurately and precisely represent the interaction problems that can be identified by all possible users/experts of a specific product.

Figure 5 shows an example taken from Turner et al. [2006]; the number of problems/errors collected by the first eight users in this example is four. From this table practitioners can calculate the weight of each identified problem/error and the raw *p*-value, which in this example, following the ROI model, is equal to 0.38.



4 kind of Problems/Errors

	Problems/Errors Number					
Subject	***1***	***2***	***3***	***4***	***Count***	***p***
1	1	0	1	0	2	**0.5**
2	1	0	1	1	3	**0.75**
3	1	0	0	0	1	**0.25**
4	0	0	0	0	0	**0**
5	1	0	1	0	2	**0.5**
6	1	0	0	0	1	**0.25**
7	1	1	0	0	2	**0.5**
8	1	0	0	0	1	**0.25**
Count	***7***	***1***	***3***	***1***		**0.38**
Raw p value	**0.875**	**0.125**	**0.375**	**0.125**		

Weight of the problems

The Raw *p* value

Fig. 5. Example of discovery likelihood of eight subjects. From this table, a practitioner may analyze the behavior of each subject, controlling for each problem how many subjects have identified it. The weight of each problem is calculated as the sum of users that have detected it, while the count of the problems identified by each subject is used for calculating the raw *p*-value, and the means of calculating each individual's *p*-value.

By organizing the data in this way the practitioners can make an important judgment about the sample's behavior in discovering problems. The sample may demonstrate homogeneous behavior, which means that the sample has a coherent rate of discovering problems. There are two cases of homogeneous behavior. The first is where all of the users have found all of the problems/errors (i.e., negative homogeneity). In this case the *p*-value would be equal to 1 and the practitioner would have reliable information for arguing that there were some problems/errors in the product that were evident and important. The sample's homogeneous negative behavior could be used to propose re-design of the product to solve these errors/problems, with a subsequent new evaluation of the updated design. The second case is where none of the users identify any problems (i.e., positive homogeneity). In this ideal condition the *p*-value is equal to 0 and the evaluators can report to the client that the technology is ready for release or for a large-scale evaluation. A *p*-value very close to 0 is usually the result of a test-retest process, in which the product has already been evaluated and re-designed, perhaps several times, so increasing the difficulties in, and reduced likelihood of, problem identification.

Alternatively, the sample may demonstrate heterogeneous behavior, which means that the sample has identified a certain number of problems/errors with different weights. This heterogeneity of problem identification clearly shows to practitioners that there are a certain number of problems in the product, but it cannot inform evaluators about the representativeness of the sample and the reliability of the data – this can be analyzed only by testing the sample *p*-value through the estimation models. As noted earlier, when the aimed-for percentage of discovered problems (D_{th}) is 80-85% with a planned sample size of five users, a *p*-value equal to or greater than 0.30 is required (see the discovery likelihood distribution in Figure 1).

While homogeneity of sample behavior leads practitioners to obtain reliable information (prompting re-design or release decisions), most evaluation studies will identify some degree of heterogeneity within the sample behavior. When the sample has a heterogeneous behavior, practitioners do not have enough information to make an informed re-design/release decision and consequently they have to analyze the *p*-value and, in line with their aim and budget constraints, consider adding more users to the sample in order to provide the quality of information needed to take an informed decision. This can be seen in Figure 6, which presents a model of the GP process, showing how it makes use of the table of problem/errors derived from the sample behavior to inform decision-making.

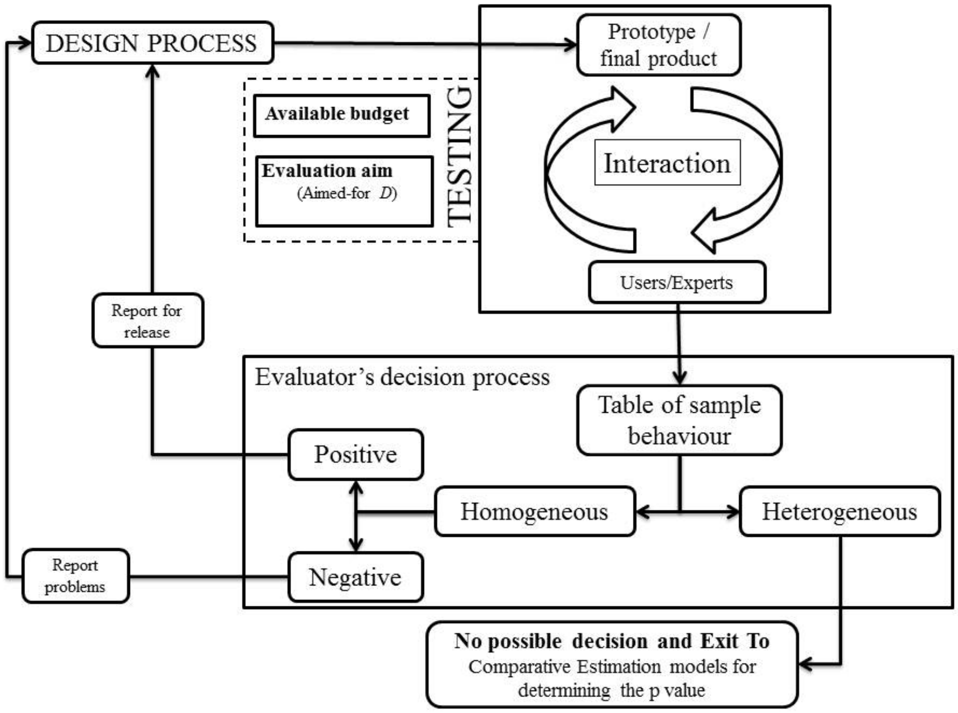

DESIGN PROCESS
Prototype / final product
Available budget
Evaluation aim
(Aimed-for D)
TESTING
Interaction
Users/Experts
Report for release
Evaluator's decision process
Table of sample behaviour
Positive
Homogeneous
Heterogeneous
Report problems
Negative
No possible decision and Exit To
Comparative Estimation models for determining the p value

Fig. 6. The application of the GP analysis of the table of sample behavior in order to drive the evaluator's decision process depending on whether the sample behavior is heterogeneous or homogeneous.

The level of heterogeneity is determined by counting the number of problems discovered only once in the sample and dividing it by the total number of problems found. Making this an explicit measure may help focus the practitioner on the degree of heterogeneity in the sample and its change as users are added. After the addition of a new user, the level of heterogeneity may decrease or increase. A decrease signifies that the new user has improved the overall representativeness of the cohort behavior, while an increase signifies that the new user has not added any new useful information for the evaluators. This may, for example, encourage evaluators to reconsider the user selection criteria if heterogeneity tends to increase as users are added (see appendix 1).

4.2. Refining the *p*-value

We have already shown how practitioners can construct a table of a study sample's behavior (Figure 5) and use this to calculate a raw *p*-value based on the average *p* of each user; in our example, *p* is equal to 0.38. By applying the ROI model with this discovery rate, we can estimate that the first four subjects will identify more than the 85% of the interface's problems. Unsurprisingly, in light of the different parameters that they consider, when the Grand-Turing, Monte Carlo analysis and the Bootstrap Discovery Behavior model are used, more conservative *p*-value estimates are produced (see Table I).

Table I. The estimation of the p-values of the sample analyzed in Figure 4 as a result of applying different estimation models: Good-Turing; Monte Carlo; and Bootstrap Discovery Behavior. The p-values estimated with the models (p_{GT}, p_{MC}, p_{BDB}) show that the discovery likelihood of this sample, composed of eight subjects, is enough to identify more than the 80% of the problems in the product under evaluation.

Grand-Turing		**Monte Carlo**		**Bootstrap Discovery Behavior**	
p_{GT}	*> 80% of the problems*	p_{MC}	*> 80% of the problems*	p_{BDB}	*> 80% of the problems*
0.235	8 subjects	0.221	8 subjects	0.215	8 subjects

In our example using Turner et al.'s [2006] data, after the first eight subjects have been studied, the practitioner can apply all of the estimation models and estimate that the study sample has a discovery likelihood ranging from 0.38 to 0.215 (M=0.265). By using as reference values the lower and the upper bounds (in this case p_{BDB}=0.215 and p_{ROI}=0.38), and the mean (p_{M}=0.265), the practitioner can determine the discovery likelihood (see appendix 2) on the basis of the data at hand and argue that, in this case, between six and eight subjects are needed to identify more than 85% of the problems associated with this product. In this case, we have applied the estimation model after the first eight subjects, but the models could be applied after the first four or five participants and the same results obtained.

We suggest that, instead of adopting a unique number provided by a specific estimation model, practitioners should rely on a range of values. This decision will, though, depend on the practitioners' budget, since the analysis of a range of *p*-values is more expensive than a test based on a single value. Practitioners can address this problem by using, as indicated above, only p_M for defining the sample likelihood so reducing the costs and the overestimation of the *p*-value.

4.3. Taking a decision on the basis of the sample behavior

As sections 4.1 and 4.2 have illustrated, in case of heterogeneous sample behavior the GP is a procedure for organizing the evaluation data, calculating the sample behavior and conducting a comparative analysis of different estimation models on the basis of the information from the tested sample (see Figure 7 for a model of this process).

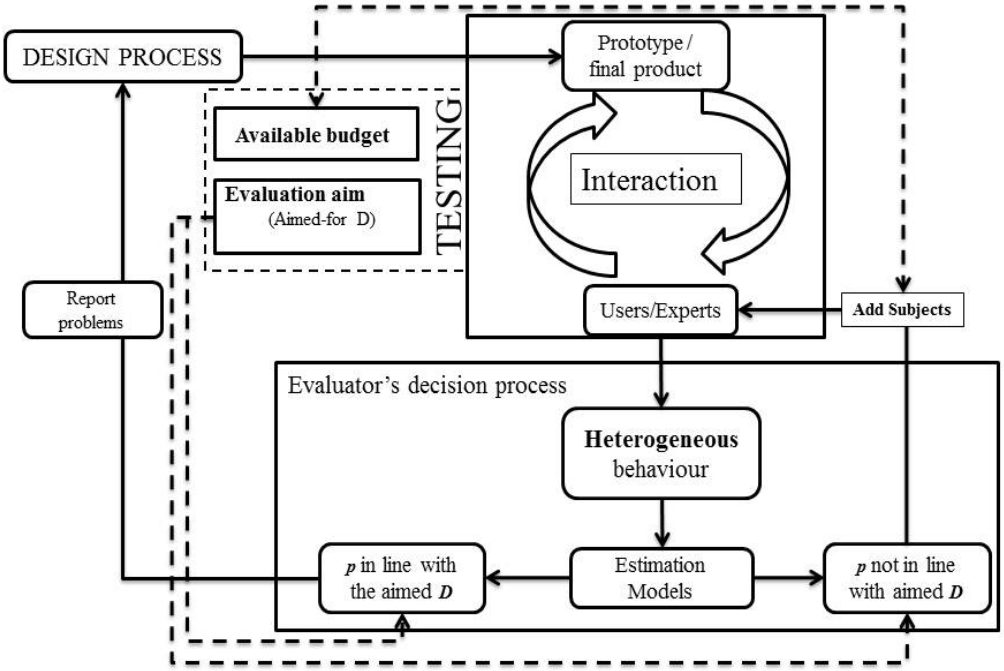


Fig. 7. The application of the GP comparative analysis of the models for estimating the sample *p*-value, and the possible actions and decisions when the *p* is, or is not, in line with the aimed-for *D* (D_{th}).

As Figure 7 shows, practitioners analyzing the data from the first X subjects, and estimating the *p*-value, may obtain two kinds of information. The first case is where the *p*-value is equal to or higher than the standard set for reaching the evaluation aim (*D*). In this case the sample has already discovered a level of problems that reaches the aimed-for *D* (D_{th}). For instance, if D_{th} is 90-95%, the *p*-value is in line with the aimed-for *D* when five users have a p=0.40, but also when seven users have a p=0.30 and so on, increasing the N in line with the available budget. Of course, the primary goal of the evaluation is to use the available budget in the optimum way by trying to obtain the aimed-for *D* with the fewest users (lowest N) possible. When the *p*-value is in line with D_{th} the practitioner is in a position to generate the evaluation report and recommend a decision to the client.

The second case is where the *p*-value is lower than the standard set for reaching the evaluation aim. In this case the sample cannot offer enough information to the evaluators for generating a report; the practitioner, respecting the available budget, has to enlarge the sample (i.e., adding users and increasing the N) in order to discover more problems until the aimed-for *D* is reached. As previously discussed, when the actual sample has a *p*-value that is not in line with the aimed-for D, it is necessary to increase the sample size (N) in order to align the *p*-value to D_{th}. For instance, if with five users the *p*-value is equal to 0.30 and D_{th} is equal to 90-95%, five users are shown to be not enough and at least other two users would have to be added to the sample.

When practitioners invest in adding new users they are seeking to improve the discovery likelihood of the sample, but this investment is a risk, as can be seen by considering the different scenarios that can arise. In the worst case, enlarging the sample by adding new users can decrease the sample's *p*-value if, for instance, these new users identify no problems at all. The sample's *p*-value will also decrease if the new users identify exclusively new problems – increasing the heterogeneity of the sample. In these two scenarios, the practitioners may have made a questionable choice in the recruitment of the additional users, which may subsequently lead them to reconsider the selection criteria used.

However, if the additional users identify only problems that have already been found by previous users, the homogeneity of the sample is increased, leading to a higher *p*-value. At the same time, if these new users also identify one or more new problems, confirming the issues identified by previous users, again both the homogeneity of the sample and the *p*-value increase. In these two scenarios, the practitioners may be argued to have made a good investment by adding these new users. At the end of this process, by comparing the obtained *p*-value with the aimed-for *D* (D_{th}) the practitioner will be in a position to decide whether to generate the evaluation report and recommend a decision to the client or to restart the evaluation cycle, again depending on available budget.

The ability of the GP to provide appropriate responses across the range of evaluation scenarios suggests, we would argue, that it is a systematic approach to the analysis of evaluation data that can be applied at different phases of, and used to inform, product development as part of a user-centered design approach (as suggested in Figure 8). For example, when the sample demonstrates positive homogeneous behavior, the practitioner can propose that the product be released, integrating the evaluation data in the product (point 4 in Figure 8). If the sample demonstrates negative homogeneous behavior, a strong redesign is required. In this case the evaluation results suggest: a) changes to the design to reflect a more realistic set of expectations about the users (point 2) ; b) re-thinking the design as a result of the gathered data (point 3); or c) integrating the outcomes of the evaluation into the product and re-evaluating it (point 4).

On the other hand, when the sample behavior is heterogeneous (see Figure 7), the practitioner has to apply the estimation models in order to estimate the sample's *p*-value. If all of the estimation models confirm that the sample matches the standard (i.e., $p \geq 0.30$) the practitioner can propose a new design cycle re-visiting points 2, 3 and 4 of the design process in the light of the evaluation results. If the *p*-value does not match the standard then the practitioner has to add new subjects, drawing on the comparative analysis from the estimation models, in order to increase the discovery rate of the sample (if the subjects find new and/or already identified errors) or to reduce the target *p*-value for a specified D_{th}.

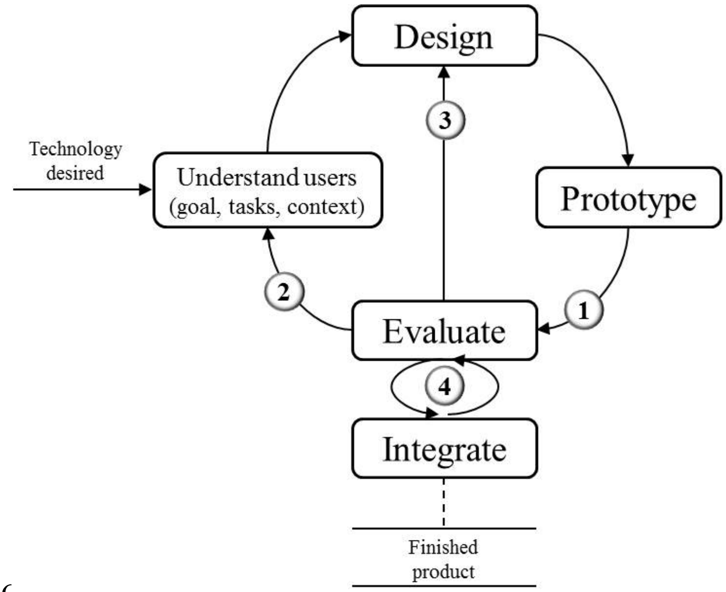


Fig. 8. The GP applied in the typical iterative User-Centered Design model and development process (adapted from Petrie and Bevan [2009]). The GP is useful at four points of the design process: 1) The evaluation of the prototype; 2) the definition and re-definition of the application desired; 3) the definition and re-definition of the design perspective; and 4) the integration of the evaluation data in the product after and before a new evaluation test.

4.4. Advantages and limitations of the Grounded Procedure

The estimation models that we have discussed in this paper have been applied with great success in the HCI field, in particular with web systems and interfaces. With these kind of technologies, a well-tested assumption is that to assess users' interaction practitioners have to consider the final user categories and sample them into multiple groups as follows [Nielsen 1995]: five subjects if testing one group of users; three or four subjects from each category if testing two groups of users; and three users from each category if testing three or more groups of users. This rule of Nielsen's [1995] is a result of the analysis of a large and reliable set of data collected by different evaluation studies.

Nielsen's comparative analysis has shown that adopting the five-user assumption is a good starting point for analyzing web systems interfaces but, as we have noted, it does not mean that five users is enough for an evaluation because the choice of sample number depends on the discovery likelihood of the sample. Nielsen's rule suggests that when computer or web interfaces are evaluated by specific evaluation techniques (e.g., the think-aloud protocol for user-based evaluation, or heuristic analysis for expert-based assessment), a sample of five users or five experts is a good starting point for the evaluation because there is a high probability (but not certainty) that such a sample has a high discovery likelihood rate (i.e., $p \geq 0.30$). In this context, the GP's value is that for a specific evaluation setting (that is, target product and chosen evaluation technique) it can help an evaluator to decide how to proceed with the evaluation after the first five users have been sampled. Away from the five-user assumption, the identification of a more reliable starting number of subjects for evaluations in different contexts is one of the most interesting features, and advantages, that widespread use of the GP would provide. The diffusion of the methodology would allow comparative analyses of different evaluation studies carried out on the same technologies, with the same evaluation techniques and with similar samples to identify more reliable initial sample sizes. As an example, Borsci et al. [2011] have undertaken a comparative analysis of the estimation models in the field of assistive technologies and have identified that a reliable starting number of blind users for the evaluation of a screen reader, when the evaluation is carried out by a Partial Concurrent Thinking Aloud evaluation technique [Federici et al. 2010], ranges from six to 20 subjects. In particular, the GP could be extremely useful in those fields that are looking for a reliable and shared evaluation framework while having to control the costs of the assessment.

As such, much as in the 1990s when the ROI model was developed to address website developers' needs to control evaluation costs, the GP can be used in the evaluation of different kinds of interactive technologies (e.g., assistive technologies, medical and industrial devices, mobile phones) which may have different requirements in terms of interaction safety. Within this context, the GP offers a way to control evaluation costs while assuring the representativeness of the sample and the associated quality of the evaluation data.

It is important to note here that the GP forces practitioners to manage and organize the gathered data in a specific way, and that the procedure of behavior analysis may be seen by evaluators as a restrictive organization of the data, and as requiring a time commitment that could prevent other kinds of analysis (e.g., environmental evaluation). We would suggest, though, that this objection may be overcome if the GP is used not as a meta-methodology but as a tool, together with other kinds of analysis in order to control the effectiveness and the efficiency of the sample evaluation.

5. CONCLUSION

By providing analysis of literature related to the *p*-value estimation and discussing the advantage and the weakness of the ROI model, this paper has presented a new perspective on the five–user assumption. We have argued that the question often posed by researchers in the field of whether five users is a sufficient number for usability testing is an unhelpful one. We have suggested instead that five subjects provides a good starting point for evaluating certain technologies (e.g., websites) with a certain evaluation technique (e.g., thinking aloud) and have shown that a five–subject sample is reliable only if it has a certain level of discovery likelihood (i.e., $p \geq 0.30$). In this sense, the only answer to the question of whether five users is or is not enough for a reliable evaluation is that it depends on the sample behavior, as this affects the reliability of the assessment and the representativeness of the gathered data.

We have proposed a method – the Grounded Procedure – that allows practitioners to analyze the reliability of the data from their usability tests, enabling them to estimate the sample size needed to identify a given proportion of interaction problems. This method provides an new perspective on the discovery likelihood and on designing evaluation studies and gives designers/manufacturers the means to use the data from their evaluations to inform critical system/product decisions, providing decision support on when to enlarge the sample, re-design, or release the product. It also allows the reliability of the evaluation to be calculated, which will help designers/manufacturers to conduct efficient evaluation studies thereby controlling costs, and will also enable them to demonstrate objectively the reliability of their evaluations to regulators and purchasers.

ACKNOWLEDGMENTS

The authors acknowledge support of this work through the MATCH programme (EPSRC Grants: EP/F063822/1 EP/G012393/1), although the views expressed are entirely their own.

REFERENCES

ALSHAMARI, M. AND MAYHEW, P. 2009. Technical Review: Current Issues of Usability Testing. *IETE Technical Review 26*, 402-406.

BIAS, R.G. AND MAYHEW, D.J. 2005. *Cost-justifying usability: An update for the Internet age*. Morgan Kaufmann Publishers, San Francisco, CA.

BORSCI, S., FEDERICI, S., MELE, M.L., POLIMENO, D. AND LONDEI, A. 2012. The Bootstrap Discovery Behaviour Model: Why Five Users are not Enough to Test User Experience. In *Cognitively Informed Intelligent Interfaces: Systems Design and Development*, E.M. ALKHALIFA AND K. GAID Eds. IGI GLobal press, Hershey, PA.

BORSCI, S., LONDEI, A. AND FEDERICI, S. 2011. The Bootstrap Discovery Behaviour (BDB): a new outlook on usability evaluation. *Cognitive Processing 12*, 23-31.

CAULTON, D.A. 2001. Relaxing the homogeneity assumption in usability testing. *Behaviour & Information Technology 20*, 1-7.

CRYSTAL, A. AND GREENBERG, J. 2005. Usability of a metadata creation application for resource authors. *Library & Information Science Research 27*, 177-189.

EFRON, B. 1979. Bootstrap Methods: Another Look at the Jackknife. *Annals of Statistics 7*, 1-26.

FAULKNER, L. 2003. Beyond the five-user assumption: Benefits of increased sample sizes in usability testing. *Behavior Research Methods 35*, 379-383.

FEDERICI, S., BORSCI, S. AND STAMERRA, G. 2010. Web usability evaluation with screen reader users: Implementation of the Partial Concurrent Thinking Aloud technique. *Cognitive Processing 11*, 263-272.

FISHMAN, G.S. 1995. *Monte Carlo: Concepts, Algorithms, and Applications*. Springer, New York.

FOOD AND DRUG ADMINISTRATION (FDA) 2011. *Draft Guidance for Industry and Food and Drug Administration Staff - Applying Human Factors and Usability Engineering to Optimize Medical Device Design*. U.S. Food and Drug Administration, Silver Spring, MD.

FOX, J. 2002. *An R and S-Plus companion to applied regression*. SAGE, California, CA.

GLASER, B.G. AND STRAUSS, A.L. 1967. The Discovery of Grounded Theory: Strategies for Qualitative Research. Aldine Publishing Company, Chicago, IL.

GOOD, I.J. 1953. The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika 40*, 237-264.

HERTZUM, M. AND JACOBSEN, N.E. 2003. The Evaluator Effect: A Chilling Fact About Usability Evaluation Methods. *International Journal of Human-Computer Interaction 15*, 183-204.

HONG, J.I., HEER, J., WATERSON, S. AND LANDAY, J.A. 2001. WebQuilt: A proxy-based approach to remote web usability testing. *ACM Transactions on Information Systems 19*, 263-285.

ISO 1998. ISO 9241-11:1998 Ergonomic requirements for office work with visual display terminals CEN, Brussels, BE.

JELINEK, F. 1997. *Statistical methods for speech recognition*. MIT Press, Cambridge, MA.

KIRAKOWSKI, J. 2005. 18 Chapter - Summative Usability Testing: Measurement and Sample Size. In *Cost-Justifying Usability (Second edition)*, R.G. BIAS AND D.J. MAYHEW Eds. Morgan Kaufmann, San Francisco, CA, 519-553.

KUROSU, M. 2007. Concept of Usability Revisited. In *Human-Computer Interaction: Interaction Design and Usability*, J. JACKO Ed. Springer, Berlin, DE, 579-586.

LEWIS, J.R. 1994. Sample Sizes for Usability Studies: Additional Considerations. *Human Factors: The Journal of the Human Factors and Ergonomics Society 36*, 368-378.

LEWIS, J.R. 2000. Validation of Monte Carlo estimation of problem discovery likelihood (Tech.Rep. No. 29.3357) IBM, Raleigh, NC.

LEWIS, J.R. 2001. Evaluation of Procedures for Adjusting Problem-Discovery Rates Estimated From Small Samples. *International Journal of Human-Computer Interaction 13*, 445-479.

LEWIS, J.R. 2006. Sample sizes for usability tests: mostly math, not magic. *Interactions 13*, 29-33.

LINDGAARD, G. AND CHATTRATICHART, J. 2007. Usability testing: what have we overlooked? In *Proceedings of the Proceedings of the SIGCHI conference on Human factors in computing systems*, San Jose, California, USA2007 ACM, New York, 1415-1424.

MANNING, C.D. AND SCHUTZE, H. 1999. *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA.

MOLICH, R. AND NIELSEN, J. 1990. Improving a human-computer dialogue. *Communications of the ACM 33*, 338-348.

NIELSEN, J. 1995. Severity Ratings for Usability Problems. http://useit.com/papers/heuristic/severityrating.html.

NIELSEN, J. 2000. Why You Only Need to Test with 5 Users. http://www.useit.com/alertbox/20000319.html.

NIELSEN, J. 2012. How Many Test Users in a Usability Study? http://www.useit.com/alertbox/number-of-test-users.html.

NIELSEN, J. AND LANDAUER, T.K. 1993. A mathematical model of the finding of usability problems. In *Proceedings of the Proceedings of the INTERACT '93 and CHI '93 Conference on Human factors in computing systems*, Amsterdam, The Netherlands, 24-29 April 1993 ACM, New York, 206-213.

NIELSEN, J. AND MACK, R.L. 1994. Usability Inspection Methods John Wiley & Sons, New York.

NORMAN, D.A. 1983. Some Observations on Mental Models. In *Mental Models*, D. GENTNER AND A. STEVEN Eds. Lawrence Earlbaum Associates, Hillsdale, NJ, 7-14.

NORMAN, D.A. 1988. *The psychology of everyday things*. Basic Books, New York.

NORMAN, D.A. AND DRAPER, S.W. 1986. *User Centered System Design: New Perspectives on Human-Computer Interaction*. Lawrence Erlbaum Associates Inc, Hillsdale, NJ.

PETRIE, H. AND BEVAN, N. 2009. The Evaluation of Accessibility, Usability, and User Experience. In *The Universal Access Handbook*, C. STEPHANIDIS Ed. CRC Press, London, UK.

SAURO, J. AND LEWIS, J.R. 2012. *Quantifying the User Experience*. Morgan Kaufmann, Waltham, MA.

SCHMETTOW, M. 2008. Heterogeneity in the usability evaluation process. In *Proceedings of the Proceedings of the 22nd British HCI Group Annual Conference on People and Computers: Culture, Creativity, Interaction - Volume 1*, Liverpool, UK2008 British Computer Society, Swinton, UK, 89-98.

SCHMETTOW, M. 2009. Controlling the usability evaluation process under varying defect visibility. In *Proceedings of the Proceedings of the 23rd British HCI Group Annual Conference on People and Computers: Celebrating People and Technology*, Cambridge, UK2009 British Computer Society, Swinton, UK, 188-197.

SCHMETTOW, M. 2012. Sample size in usability studies. *Communications of the ACM 55*, 64-70.

SHNEIDERMAN, B. 1986. *Designing the user interface: strategies for effective human-computer interaction*. Addison-Wesley Longman Publishing Co., Inc, Boston, MA.

SPOOL, J. AND SCHROEDER, W. 2001. Testing web sites: five users is nowhere near enough. In *Proceedings of the CHI '01 extended abstracts on Human factors in computing systems*, Seattle, Washington2001 ACM, New York, 285-286.

TULLIS, T. AND ALBERT, W. 2008. *Measuring the user experience: collecting, analyzing, and presenting usability metrics*. Morgan Kaufmann, Burlington, MA.

TURNER, C.W., LEWIS, J.R. AND NIELSEN, J. 2006. Determining Usability Test Sample Size. In *International Encyclopedia of Ergonomics and Human Factors*, W. KARWOWSKI Ed. CRC Press, Boca Raton, FL, 3084-3088.

VIRZI, R.A. 1990. Streamlining the Design Process: Running Fewer Subjects. In *Proceedings of the Proceedings of the Human Factors Society 34th Annual Meeting*, Santa Monica1990 ACM, New York, 291-294.

VIRZI, R.A. 1992. Refining the test phase of usability evaluation: how many subjects is enough? *Human Factors 34*, 457-468.

WOOLRYCH, A. AND COCKTON, G. 2001. Why and when five test users aren't enough. In *Proceedings of the Proceedings of IHM-HCI 2001 Conference*, Toulouse, FR, 10-14 Sept. 2001, J. VANDERDONCKT, A. BLANDFORD AND A. DERYCKE Eds. Cépaduès Editions, London, UK, 105-108.

Appendix 1. Example of heterogeneity analysis of the sample

This appendix reports an example of a small scale assessment of a website prototype. The aim of the evaluation is to identify at least 80% of the problems, with a budget available for testing no more than 10 participants. The sample analysis was carried out with the following estimation models: Return On Investment (ROI); Good – Turing (p_{GT}); Monte Carlo (p_{MC}); Bootstrap Discovery Behavior (p_{BDB}).

Table I. The test carried out by the first five participants.

	Problems										**Count**	**P**
SS	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**		
1	1	0	0	0	0	0	1	0	0	0	2	0.2
2	0	0	0	0	0	0	0	1	0	0	1	0.1
3	0	1	1	0	0	0	0	0	1	0	3	0.3
4	0	0	0	1	0	0	0	0	0	1	2	0.2
5	0	1	0	0	1	1	0	0	1	0	4	0.4
Count	**1**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**2**	**1**	**ROI**	**0.24**
											$\mathbf{p_{GT}}$	**0.08**
											$\mathbf{p_{MC}}$	**0.13**
											$\mathbf{p_{BDB}}$	**0.11**
	Heterogeneity analysis										**Level of Heterogeneity**	
	20%	40%	20%	20%	20%	20%	20%	20%	40%	20%		
	1	0	1	1	1	1	1	1	0	1	**80%**	

The data in Table I suggests to the practitioner that the sample is strongly heterogeneous, and that the estimation models show a large range of p-values (i.e., $0.08<p<0.24$). On the basis of these data, the practitioner should not stop the assessment, and they are forced to add other users. Moreover, the current high heterogeneity of the sample indicates to practitioners that there are some problems in the assessment. It could be that the website is particularly hard for the selected users to use (i.e., that the system is much too technical, or too innovative) or there is a bias in the selection criteria of the evaluation cohort. This second point may lead the practitioner to reconsider carefully the user selection criteria before moving ahead with further evaluation tests.

If the practitioner chooses the revise the selection criteria, the practitioner can face the following two scenarios:

(1) Evaluators add one or more new users which increases the overall performance of the sample. This case is exemplified in Table II, in which a new problem is identified (number 11), the level of heterogeneity is strongly reduced to well under 50%, and the estimation models support the evidence that the trend in the sample behavior is positively changed (i.e., $0.15<p<0.27$). In this case, despite the aimed-for D still not being reached, the addition of user 6 can be considered a good investment, and the practitioner can proceed to add other users. If these new users maintain the positive

trend, the practitioner will achieve the aimed-for D without the need to include the mandated 10 users, saving a part of the budget.

Table II. Example of a good investment when a sixth user is added.

	Problems											**Count**	**P**
SS	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**		
1	1	0	0	0	0	0	1	0	0	0	0	2	0.18
2	0	0	0	0	0	0	0	1	0	0	0	1	0.09
3	0	1	1	0	0	0	0	0	1	0	0	3	0.27
4	0	0	0	1	0	0	0	0	0	1	0	2	0.18
5	0	1	0	0	1	1	0	0	1	0	0	4	0.36
6	1	0	0	1	1	0	1	0	0	1	1	6	0.54
Count	2	2	1	2	2	1	2	1	2	2	1	**ROI**	**0.27**
												p_{GT}	**0.15**
												p_{MC}	**0.19**
												p_{BDB}	**0.23**
	Heterogeneity analysis											**Level of Heterogeneity**	
	40%	40%	20%	40%	40%	20%	40%	20%	40%	40%	20%		
	0	0	1	0	0	1	0	1	0	0	1	**27%**	

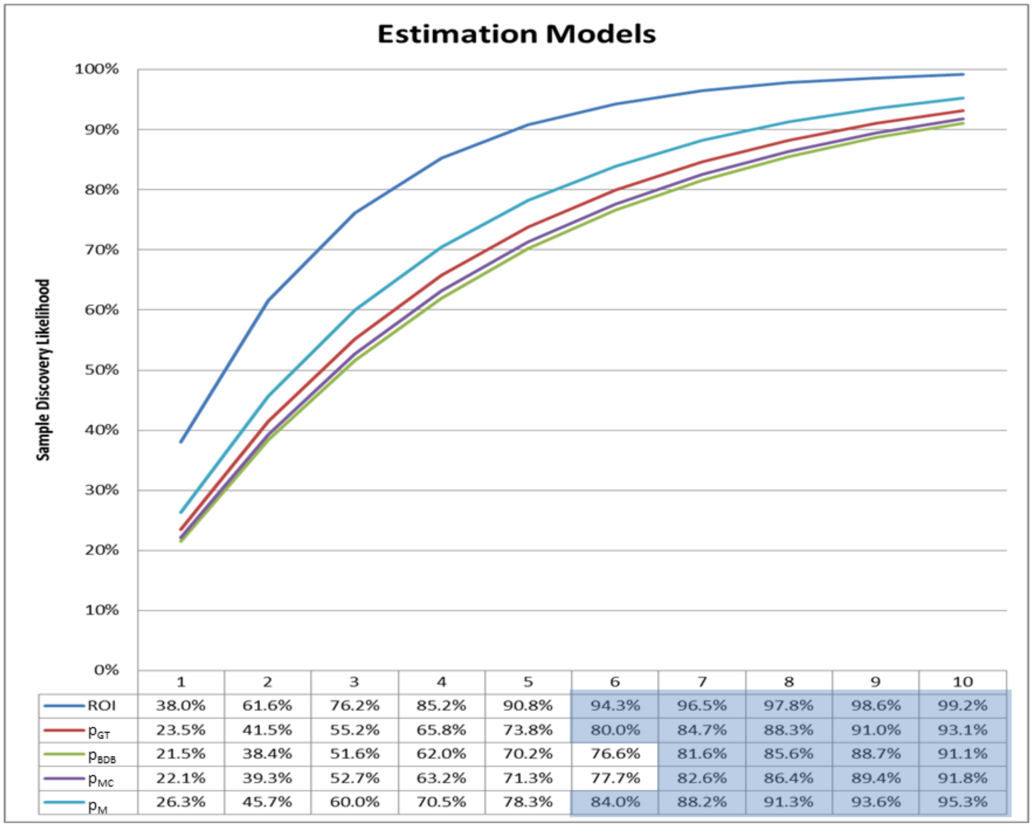
(2) Evaluators add one or more new users that decrease the overall performance of the sample. This case is exemplified in Table III, in which, though a new problem is identified (number 11), the *p*-values decrease, indicating to the practitioner that they are unlikely to achieve the aimed-D with a sample of 10 users. In such a scenario, if the practitioner cannot identify a solution that will decrease the heterogeneity of the sample to below 50%, they will be forced, at the end of the assessment process, to report to their clients that, though the evaluation data are useful indicators for the prototype redesign, a larger sample of users would be needed to obtain a more reliable set of data.

Table III. Example of a bad investment when a sixth user is added.

	Problems		

SS	1	2	3	4	5	6	7	8	9	10	11	count	p
1	1	0	0	0	0	0	1	0	0	0	0	2	0.18
2	0	0	0	0	0	0	0	1	0	0	0	1	0.09
3	0	1	1	0	0	0	0	0	1	0	0	3	0.27
4	0	0	0	1	0	0	0	0	0	1	0	2	0.18
5	0	1	0	0	1	1	0	0	1	0	0	4	0.36
6	1	0	0	0	0	0	0	0	0	0	1	2	0.18
Count	2	2	1	1	1	1	1	1	2	1	1	**ROI**	**0.21**
												p_{GT}	**0.10**
												p_{MC}	**0.15**
												p_{BDB}	**0.13**
	Heterogeneity analysis											**Level of Heterogeneity**	
	40 %	40 %	20 %	20 %	20 %	20 %	20 %	20 %	40 %	20 %	20 %		
	0	0	1	1	1	1	1	1	0	1	1	64%	

Appendix 2. Estimated discovery likelihood using Turner et al.'s data.



	1	2	3	4	5	6	7	8	9	10
ROI	38.0%	61.6%	76.2%	85.2%	90.8%	94.3%	96.5%	97.8%	98.6%	99.2%
p_{GT}	23.5%	41.5%	55.2%	65.8%	73.8%	80.0%	84.7%	88.3%	91.0%	93.1%
p_{BDB}	21.5%	38.4%	51.6%	62.0%	70.2%	76.6%	81.6%	85.6%	88.7%	91.1%
p_{MC}	22.1%	39.3%	52.7%	63.2%	71.3%	77.7%	82.6%	86.4%	89.4%	91.8%
p_M	26.3%	45.7%	60.0%	70.5%	78.3%	84.0%	88.2%	91.3%	93.6%	95.3%

Fig. 1. Graphical representation of the sample discovery likelihood estimated by a range of estimation models: Return On Investment (ROI); Good –Turing (p_{GT}); Monte Carlo (p_{MC}); Bootstrap Discovery Behavior (p_{BDB}); and the average p-value (p_M) of these models. The highlighted areas in the data indicate when the aimed-for threshold D is approximated. In this case, the practitioner can argue that with more than six subjects, the threshold of 85% is likely to be met, with the average *p*-value (p_M) being equal to 84% for six subjects. However, it is only after eight subjects have been included that all of the estimated *p*-values return a value over 85%. In light of this, the practitioner can argue that between six (lower bound) and eight (upper bound) subjects are needed to reach the aimed-for threshold.