Overestimation of *p* in Problem Discovery Usability Studies: How Serious is the Problem?

TR 29.3358

James R. Lewis

Speech Product Design and Usability

West Palm Beach, Florida

Abstract

Monte Carlo estimates of the likelihood of problem discovery (*p*) in usability studies showed that the problem of overestimation is potentially serious and can have at least two adverse consequences:

1. Leading usability practitioners to believe that a study has uncovered a greater proportion of a system's usability problems than it actually has.
2. Causing an underestimation of the required sample size to meet specific problem discovery goals.

The results of these Monte Carlo experiments demonstrate that both higher values of true *p* and a greater number of participants reduce, but do not necessarily eliminate, the overestimation.

ITIRC Keywords

Monte Carlo estimation
problem discovery likelihood
overestimation of *p*
usability evaluation
sample size estimation

Contents

Introduction

Investigations into sample size estimation have found the *p*, the likelihood of problem discovery for a product or system undergoing usability evaluation, plays a key role in determining the required sample size for a usability study (Lewis, 1994). Following the practice of using pilot studies to estimate variability when planning sample sizes for experiments based on comparison of means (Diamond, 1981; Walpole, 1976), some authors have recommended getting estimates of *p* from small sample usability studies for the purpose of estimating usability study sample sizes (Lewis, 1991, 2000b). Recently, though, Hertzum and Jacobsen (in press) have pointed out that this practice will almost always result in overestimation of the value of *p*.

For example, consider the distribution of discovered problems across participants in Table 1. An 'x' in the table indicates that this participant experienced this problem during the usability evaluation. In this hypothetical example, all participants experienced Problem 1, but only the first and tenth participants experienced Problem 10. Because the entire matrix has 100 cells (ten participants by ten problems) and 50 cells contain an 'x', the value of *p* is .5 (50/100). Note that this is the same as the estimate of *p* calculated by averaging *p* for each participant in the table.

Table 1. Hypothetical distribution of ten usability problems over ten participants

Participant	**Prob 1**	**Prob 2**	**Prob 3**	**Prob 4**	**Prob 5**	**Prob 6**	**Prob 7**	**Prob 8**	**Prob 9**	**Prob 10**	**Count**
1	x	x		x		x		x		x	6
2	x	x		x		x		x			5
3	x	x		x	x	x					5
4	x	x		x			x				4
5	x	x	x	x		x			x		6
6	x	x	x					x			4
7	x	x	x		x						4
8	x	x	x		x		x				5
9	x		x		x		x		x		5
10	x		x		x		x		x	x	6
Count	10	8	6	5	5	4	4	3	3	2	50

Suppose, though, that in this hypothetical example the usability practitioner had stopped the evaluation after the third participant. In that case, the known distribution of problems would be a subset of the set of problems discovered with ten participants, as shown in Table 2.

Table 2. Hypothetical distribution of problems discovered with first three participants

Participant	**Prob 1**	**Prob 2**	**Prob 3**	**Prob 4**	**Prob 5**	**Prob 6**	**Prob 7**	**Prob 8**	**Prob 9**	**Prob 10**	**Count**
1	x	x		x		x		x		x	6
2	x	x		x		x		x			5
3	x	x		x	x	x					5
Count	3	3	0	3	1	3	0	2	0	1	16

In Table 2, there are 30 cells (three participants by ten problems) and 16 cells containing 'x'. Dividing the number of cells containing 'x' by the total number of cells produces .533 as the estimate of *p* (which isn't much different from the estimate derived from Table 1). In this case, however, the practitioner would not know of the existence of Problems 3, 7, and 9 because none of the first three participants experienced these problems. So, when the practitioner would gather the data together for the purpose of estimating *p*, the data would not contain those columns, as shown in Table 3.

Table 3. Hypothetical problem distribution with three participants: practitioner's view

Participant	**Prob 1**	**Prob 2**	**Prob 4**	**Prob 5**	**Prob 6**	**Prob 8**	**Prob 10**	**Count**
1	x	x	x		x	x	x	6
2	x	x	x		x	x		5
3	x	x	x	x	x			5
Count	3	3	3	1	3	2	1	16

In Table 3, there are only 21 cells (seven observed problems by three participants), with sixteen of the cells containing an 'x'. This reduction in the denominator increases the estimate of *p* from .533 to .762, about a 50% overestimation.

This is a potentially serious problem because overestimation of *p* can lead usability practitioners to believe they have uncovered a greater proportion of a system's usability problems than they really have and necessarily leads to underestimation of the required sample size. The consequence of undersampling would be to fail to achieve the problem discovery goals for a usability study.

Fortunately, over the last ten years a number of researchers have published the distribution of problems discovered in usability evaluations with fairly large samples (Lewis, 1994; Nielsen & Molich, 1990; Virzi, 1990). These distributions provide a source for conducting investigations of the overestimation of *p* as a function of pilot sample size and the true value of *p*. Lewis (2000a) recently validated the use of Monte Carlo estimation to investigate the properties of *p* in problem discovery studies by showing that it produced estimates essentially identical to those obtained by complete factorial combination of a study's participants.

The primary purpose of the current experiments was to investigate the extent to which calculating *p* from small-sample usability studies results in overestimation. The sources for the evaluations include three complete problem discovery databases, subsets derived from one of the complete databases, and the hypothetical data provided in Table 1.

Method

I wrote a BASIC program that estimates the following statistics from problem discovery databases using Monte Carlo estimation with 1000 iterations (Lewis, 2000a):

- mean value of *p*
- standard deviation of *p*
- root mean square error for estimated *p* against true *p*
- standard error of the mean for *p*
- delta for a 99% confidence interval around *p*
- upper and lower bounds for a 99% confidence interval around *p*
- 1^{st}, 5^{th}, 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, 95^{th}, and 99^{th} percentiles for the distribution of *p*

I ran these programs on a Micron Millennia[1] computer (Windows[2] 95, 64 MB memory) to evaluate the following published problem discovery databases for sample sizes ranging from two to six participants:

- MACERR (Lewis, 1994; Lewis, Henry, & Mack, 1990)
- VIRZI90 (Virzi, 1990; 1992)
- MANTEL (Nielsen & Molich, 1990)

In addition to the published databases, I also evaluated the hypothetical database provided in the introduction to this report (SAMPLE) and various subsets of published databases designed to give *p* for that subset a particular value. These subsets were:

- MACERR10: 45 problems selected to produce *p*=.10
- MACERR25: 34 problems selected to produce *p*=.25
- MACERR50: 10 problems selected to produce *p*=.50
- MACERR73: the three highest frequency problems with average *p*=.73

(For copies of the databases, see Lewis, 2000a.)

[1] Micron and Millennia are trademarks or registered trademarks of Micron Inc.
[2] Windows is a trademark or registered trademark of Microsoft Corp.

Results

Estimates of Mean *p*

Tables 4 and 5 (and Figure 1) show the means of the estimates of *p* for the various published databases and the selected subsets.

Table 4. Estimates of mean p for published databases

Source	True p	Sample=2	Sample=3	Sample=4	Sample=5	Sample=6
MACERR	*0.16*	0.568	0.421	0.346	0.301	0.269
VIRZI90	*0.36*	0.661	0.544	0.484	0.448	0.425
MANTEL	*0.38*	0.724	0.622	0.572	0.536	0.511

Table 5. Estimates of mean p for hypothetical and subset data

Source	True p	Sample=2	Sample=3	Sample=4	Sample=5	Sample=6
MACERR10	*0.10*	0.523	0.357	0.276	0.225	0.193
MACERR25	*0.25*	0.550	0.410	0.341	0.307	0.282
MACERR50	*0.50*	0.645	0.558	0.519	0.509	0.502
MACERR73	*0.73*	0.785	0.742	0.734	0.733	0.733
SAMPLE	*0.50*	0.711	0.601	0.551	0.524	0.512

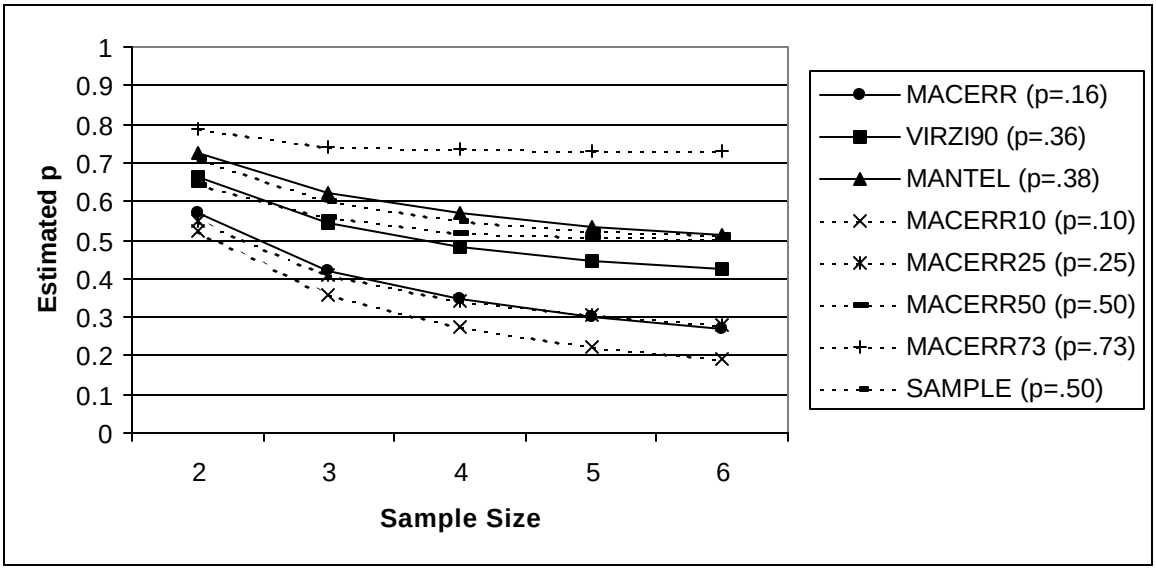

Figure 1. Means as a function of source and sample size

Mean *p* Overestimation Ratios

Table 6 and 7 (and Figure 2) show the overestimation ratios (estimated *p* divided by true *p*) for all investigated databases. A ratio of 1 indicates no overestimation. A ratio of 1.2, for example, indicates a 20% overestimation.

Table 6. Overestimation ratios for published databases

Source	True p	Sample=2	Sample=3	Sample=4	Sample=5	Sample=6
MACERR	*0.16*	3.6	2.6	2.2	1.9	1.7
VIRZI90	*0.36*	1.8	1.5	1.3	1.2	1.2
MANTEL	*0.38*	1.9	1.6	1.5	1.4	1.3

Table 7. Overestimation ratios for hypothetical and subset data

Source	True p	Sample=2	Sample=3	Sample=4	Sample=5	Sample=6
MACERR10	*0.10*	5.2	3.6	2.8	2.3	1.9
MACERR25	*0.25*	2.2	1.6	1.4	1.2	1.1
MACERR50	*0.50*	1.3	1.1	1.0	1.0	1.0
MACERR73	*0.73*	1.1	1.0	1.0	1.0	1.0
SAMPLE	*0.50*	1.4	1.2	1.1	1.0	1.0

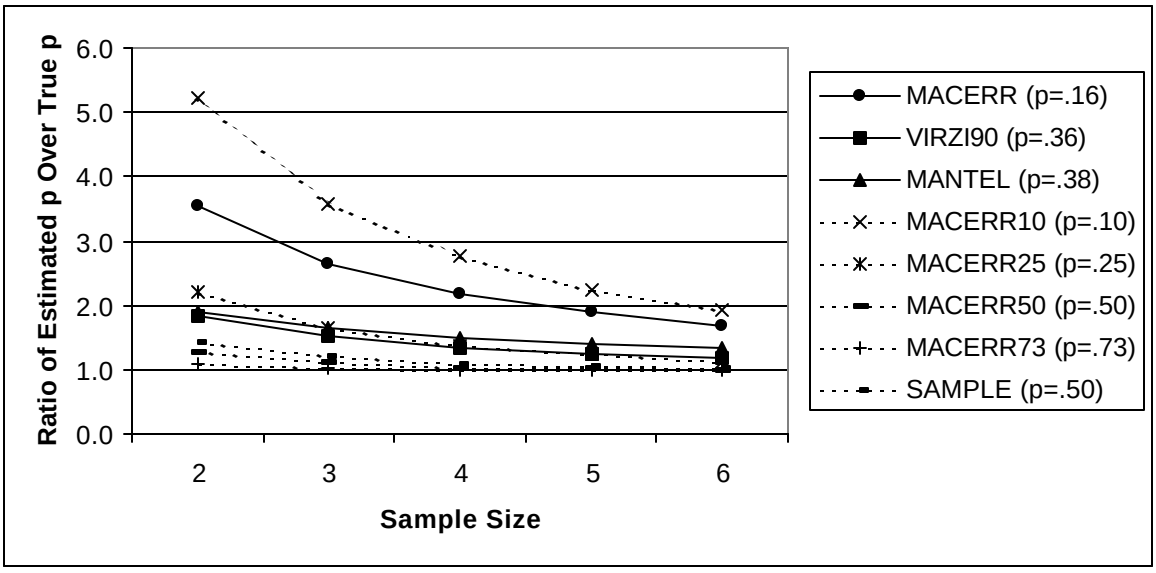

Figure 2. Overestimation ratios as a function of source and sample size

Estimates of the Root Mean Square Error

The root mean square (rms) error is similar to a standard deviation, but rather than computing the mean squared difference between each data point and the mean of a distribution (the standard deviation), the computation is the mean squared difference between each data point and the true value of *p* for the distribution[3]. Tables 8 and 9 (and Figure 3) show the rms error calculated for the published databases and selected subsets. A smaller rms error indicates more accurate estimation.

Table 8. Estimates of rms error for published databases

Source	**True p**	**Sample=2**	**Sample=3**	**Sample=4**	**Sample=5**	**Sample=6**
MACERR	*0.16*	0.406	0.259	0.185	0.140	0.107
VIRZI90	*0.36*	0.306	0.189	0.130	0.094	0.071
MANTEL	*0.38*	0.354	0.254	0.204	0.167	0.143

Table 9. Estimates of rms error for hypothetical and subset data

Source	**True p**	**Sample=2**	**Sample=3**	**Sample=4**	**Sample=5**	**Sample=6**
MACERR10	*0.10*	0.425	0.257	0.175	0.124	0.092
MACERR25	*0.25*	0.307	0.170	0.102	0.071	0.049
MACERR50	*0.50*	0.173	0.104	0.079	0.070	0.063
MACERR73	*0.73*	0.151	0.116	0.095	0.082	0.069
SAMPLE	*0.50*	0.230	0.118	0.067	0.038	0.026

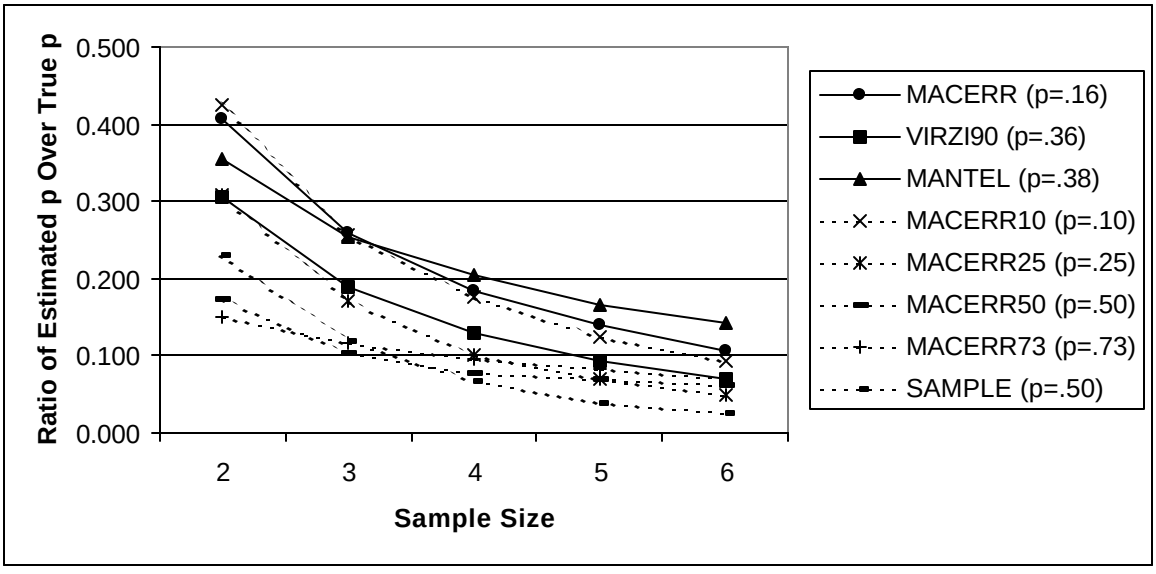

Figure 3. Root mean square error as a function of source and sample size

[3] If not true *p*, using the best estimate of *p* available, computed from the full data set.

Percentile Estimates of *p* for Published Distributions

Tables 10, 12, and 14 (and Figures 4-6) show the estimates for the 1st, 25th, 50th, 75th, and 99th percentiles for the MACERR, VIRZI90 and MANTEL problem discovery databases as a function of sample size. The 50th percentile is the median (center) of the distribution. The 25th and 75th percentiles define the interquartile range – the range that contains the central 50% of a distribution. The range between the 1st and 99th percentiles contains the central 98% of the distribution (excludes only one percent of each tail). Tables 11, 13, and 15 contain the range sizes.

Table 10. MACERR percentiles of p as a function of sample size (True p = .16)

Percentile	**Sample=2**	**Sample=3**	**Sample=4**	**Sample=5**	**Sample=6**
99th	0.646	0.475	0.390	0.340	0.303
75th	0.588	0.441	0.360	0.313	0.280
50th	0.565	0.422	0.346	0.301	0.269
25th	0.544	0.401	0.332	0.290	0.258
1st	0.500	0.364	0.298	0.260	0.231

Table 11. Interquartile and 98% ranges of p for MACERR

Range	**Sample=2**	**Sample=3**	**Sample=4**	**Sample=5**	**Sample=6**
Interquartile	0.044	0.040	0.028	0.023	0.022
98 Percent	0.146	0.111	0.092	0.080	0.072

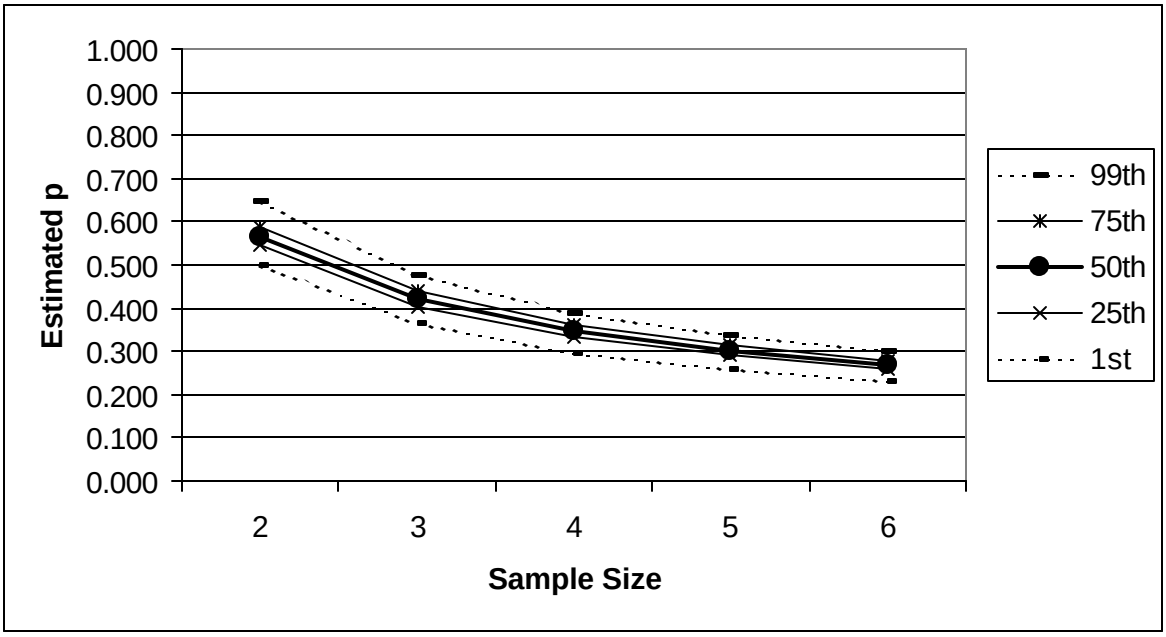

Figure 4. MACERR percentiles of p as a function of sample size

Table 12. VIRZI90 percentiles of p as a function of sample size (True p = .36)

Percentile	**Sample=2**	**Sample=3**	**Sample=4**	**Sample=5**	**Sample=6**
99th	0.773	0.640	0.565	0.526	0.494
75th	0.688	0.568	0.500	0.467	0.443
50th	0.658	0.543	0.483	0.447	0.424
25th	0.630	0.519	0.461	0.429	0.407
1st	0.563	0.464	0.414	0.389	0.369

Table 13. Interquartile and 98% ranges of p for VIRZI90

Range	**Sample=2**	**Sample=3**	**Sample=4**	**Sample=5**	**Sample=6**
Interquartile	0.058	0.049	0.039	0.038	0.036
98 Percent	0.210	0.176	0.151	0.137	0.125

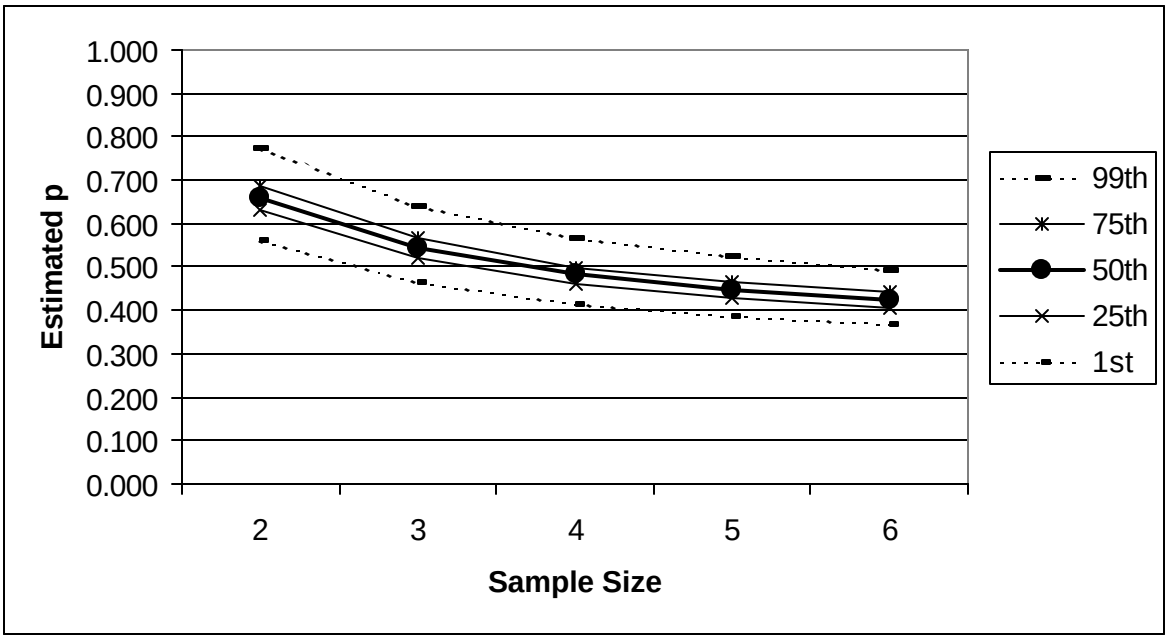

Figure 5. VIRZI90 percentiles of p as a function of sample size

Table 14. MANTEL percentiles of p as a function of sample size (True p = .38)

Percentile	**Sample=2**	**Sample=3**	**Sample=4**	**Sample=5**	**Sample=6**
99th	0.875	0.769	0.702	0.643	0.615
75th	0.767	0.667	0.605	0.567	0.540
50th	0.727	0.621	0.574	0.538	0.508
25th	0.679	0.583	0.536	0.504	0.480
1st	0.571	0.487	0.450	0.429	0.417

Table 15. Interquartile and 98% ranges of p for MANTEL

Range	**Sample=2**	**Sample=3**	**Sample=4**	**Sample=5**	**Sample=6**
Interquartile	0.088	0.084	0.069	0.063	0.060
98 Percent	0.304	0.282	0.252	0.214	0.198

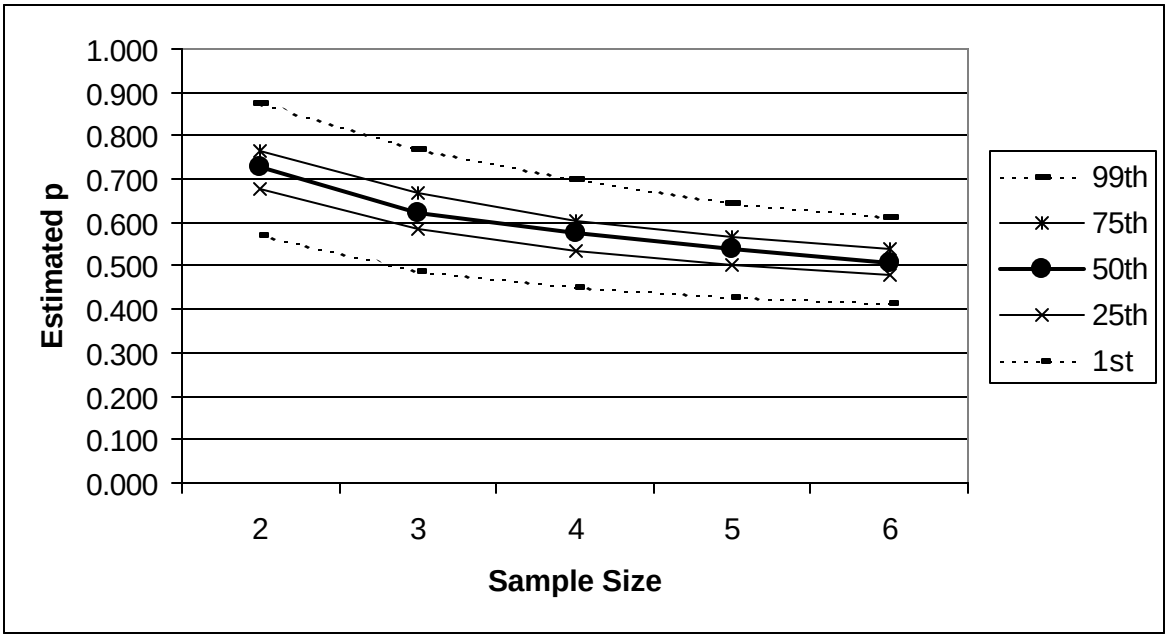

Figure 6. MANTEL percentiles of p as a function of sample size

Effect of Overestimation of *p* on Projected Sample Size Requirements

Tables 16-22 and Figures 7-13 illustrate the effect of the overestimation of *p* on projected sample size requirements using data generated from the published problem discovery databases and the MACERR subsets (MACERR10, MACERR25, MACERR50 and MACERR73). Bolded cell entries in the tables show the sample size at which a practitioner, using the mean value of *p* from the Monte Carlo experiments, would conclude that a sample size was adequate to achieve the goal of detecting 90% of the problems. Cells with a bold italic font indicate problem detection at the 95% level. Attached to each table is an analysis of the difference between the apparently required and actually required sample sizes to achieve 90% and 95% problem detection goals.

For example, in Table 16, for participants taken two at a time from the database, the mean value of *p* is .568. Using the formula $1-(1-p)^n$ (Lewis, 1994) to project problem discovery as a function of sample size produces the values in the cells for the column headed Est-*p*(2) (problem discovery estimates using *p* based on 2 participants). In that column, the first cell that exceeds .90 is the one for a projected sample size of 3 participants. The first cell that exceeds .95 is with a projected sample size of 4 participants. If the estimate of *p* came from a study with 6 participants (Est-*p*(6)), then the 90% and 95% goals appear to require 8 and 10 participants respectively. Projections based on the true *p* of .16 indicate that the true sample sizes required to meet those goals are 14 and 18 participants respectively. If a practitioner's goal is to detect 90% of the usability problems in a product and he or she uses an estimate of *p* based on 2 participants, then the discrepancy between the apparently required and actually required sample sizes is 11 participants. Similarly, if the goal is to achieve 95% problem detection and the estimate of *p* has come from a study based on 6 participants, the discrepancy between the apparently required and actually required sample sizes is 8 participants.

Table 16. Projected sample size requirements for MACERR

Projected Sample	**Est-p(2)**	**Est-p(4)**	**Est-p(6)**	**True-p**	***Sample Size***	**Goal: 90%**	**Goal: 95%**
1	0.568	0.346	0.269	0.160	*2*	11	14
2	0.813	0.572	0.466	0.294	*4*	8	10
3	**0.919**	0.720	0.609	0.407	*6*	6	8
4	***0.965***	0.817	0.714	0.502			
5	0.985	0.880	0.791	0.582			
6	0.994	**0.922**	0.847	0.649			
7	0.997	0.949	0.888	0.705			
8	0.999	***0.967***	**0.918**	0.752			
9	0.999	0.978	0.940	0.792			
10	1.000	0.986	***0.956***	0.825			
11	1.000	0.991	0.968	0.853			
12	1.000	0.994	0.977	0.877			
13	1.000	0.996	0.983	0.896			
14	1.000	0.997	0.988	**0.913**			
15	1.000	0.998	0.991	0.927			
16	1.000	0.999	0.993	0.939			
17	1.000	0.999	0.995	0.948			
18	1.000	1.000	0.996	***0.957***			
19	1.000	1.000	0.997	0.964			
20	1.000	1.000	0.998	0.969			

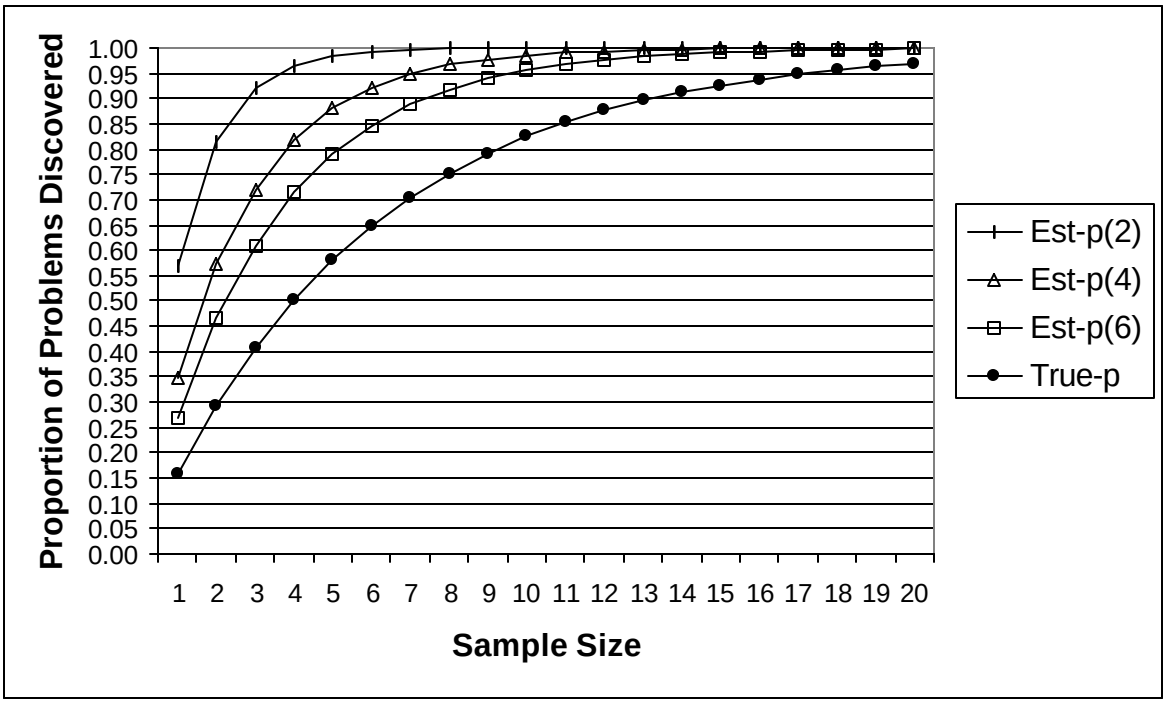

Figure 7. Projected problem discovery curves for MACERR

Table 17. Projected sample size requirements for VIRZI90

Projected Sample	**Est-p(2)**	**Est-p(4)**	**Est-p(6)**	**True-p**	***Sample Size***	**Goal: 90%**	**Goal: 95%**
1	0.661	0.484	0.425	0.360	*2*	3	4
2	0.885	0.734	0.669	0.590	*4*	2	2
3	***0.961***	0.863	0.810	0.738	*6*	1	1
4	0.987	**0.929**	0.891	0.832			
5	0.996	***0.963***	**0.937**	0.893			
6	0.998	0.981	***0.964***	**0.931**			
7	0.999	0.990	0.979	***0.956***			
8	1.000	0.995	0.988	0.972			
9	1.000	0.997	0.993	0.982			
10	1.000	0.999	0.996	0.988			
11	1.000	0.999	0.998	0.993			
12	1.000	1.000	0.999	0.995			
13	1.000	1.000	0.999	0.997			
14	1.000	1.000	1.000	0.998			
15	1.000	1.000	1.000	0.999			
16	1.000	1.000	1.000	0.999			
17	1.000	1.000	1.000	0.999			
18	1.000	1.000	1.000	1.000			
19	1.000	1.000	1.000	1.000			
20	1.000	1.000	1.000	1.000			

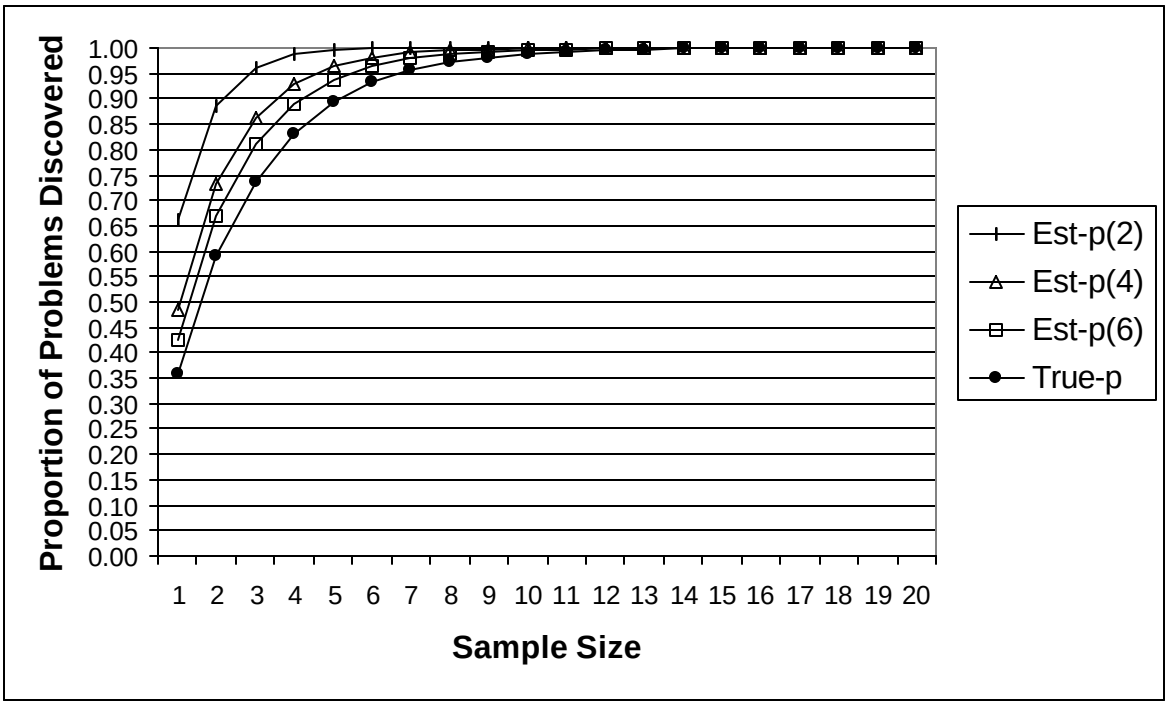

Figure 8. Projected problem discovery curves for VIRZI90

Table 18. Projected sample size requirements for MANTEL

Projected Sample	**Est-p(2)**	**Est-p(4)**	**Est-p(6)**	**True-p**	***Sample Size***	**Goal: 90%**	**Goal: 95%**
1	0.724	0.572	0.511	0.380	*2*	3	4
2	**0.924**	0.817	0.761	0.616	*4*	2	3
3	***0.979***	**0.922**	0.883	0.762	*6*	1	2
4	0.994	***0.966***	**0.943**	0.852			
5	0.998	0.986	***0.972***	**0.908**			
6	1.000	0.994	0.986	0.943			
7	1.000	0.997	0.993	***0.965***			
8	1.000	0.999	0.997	0.978			
9	1.000	1.000	0.998	0.986			
10	1.000	1.000	0.999	0.992			
11	1.000	1.000	1.000	0.995			
12	1.000	1.000	1.000	0.997			
13	1.000	1.000	1.000	0.998			
14	1.000	1.000	1.000	0.999			
15	1.000	1.000	1.000	0.999			
16	1.000	1.000	1.000	1.000			
17	1.000	1.000	1.000	1.000			
18	1.000	1.000	1.000	1.000			
19	1.000	1.000	1.000	1.000			
20	1.000	1.000	1.000	1.000			

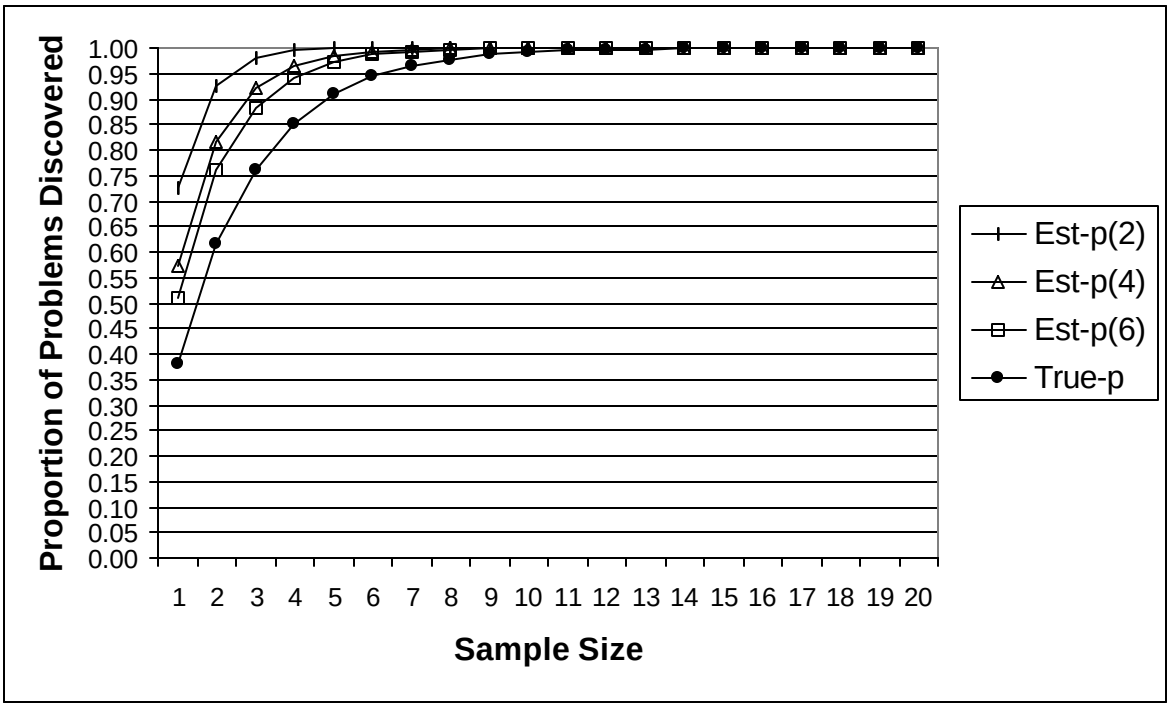

Figure 9. Projected problem discovery curves for MANTEL

Table 19. Projected sample size requirements for MACERR10

Projected Sample	**Est-p(2)**	**Est-p(4)**	**Est-p(6)**	**True-p**	***Sample Size***	**Goal: 90%**	**Goal: 95%**
1	0.523	0.276	0.193	0.100	*2*	18	25
2	0.772	0.476	0.349	0.190	*4*	14	19
3	0.891	0.620	0.474	0.271	*6*	11	15
4	**0.948**	0.725	0.576	0.344			
5	***0.975***	0.801	0.658	0.410			
6	0.988	0.856	0.724	0.469			
7	0.994	0.896	0.777	0.522			
8	0.997	**0.925**	0.820	0.570			
9	0.999	0.945	0.855	0.613			
10	0.999	***0.960***	0.883	0.651			
11	1.000	0.971	**0.905**	0.686			
12	1.000	0.979	0.924	0.718			
13	1.000	0.985	0.938	0.746			
14	1.000	0.989	***0.950***	0.771			
15	1.000	0.992	0.960	0.794			
16	1.000	0.994	0.968	0.815			
17	1.000	0.996	0.974	0.833			
18	1.000	0.997	0.979	0.850			
19	1.000	0.998	0.983	0.865			
20	1.000	0.998	0.986	0.878			
21	1.000	0.999	0.989	0.891			
22	1.000	0.999	0.991	**0.902**			
23	1.000	0.999	0.993	0.911			
24	1.000	1.000	0.994	0.920			
25	1.000	1.000	0.995	0.928			
26	1.000	1.000	0.996	0.935			
27	1.000	1.000	0.997	0.942			
28	1.000	1.000	0.998	0.948			
29	1.000	1.000	0.998	***0.953***			

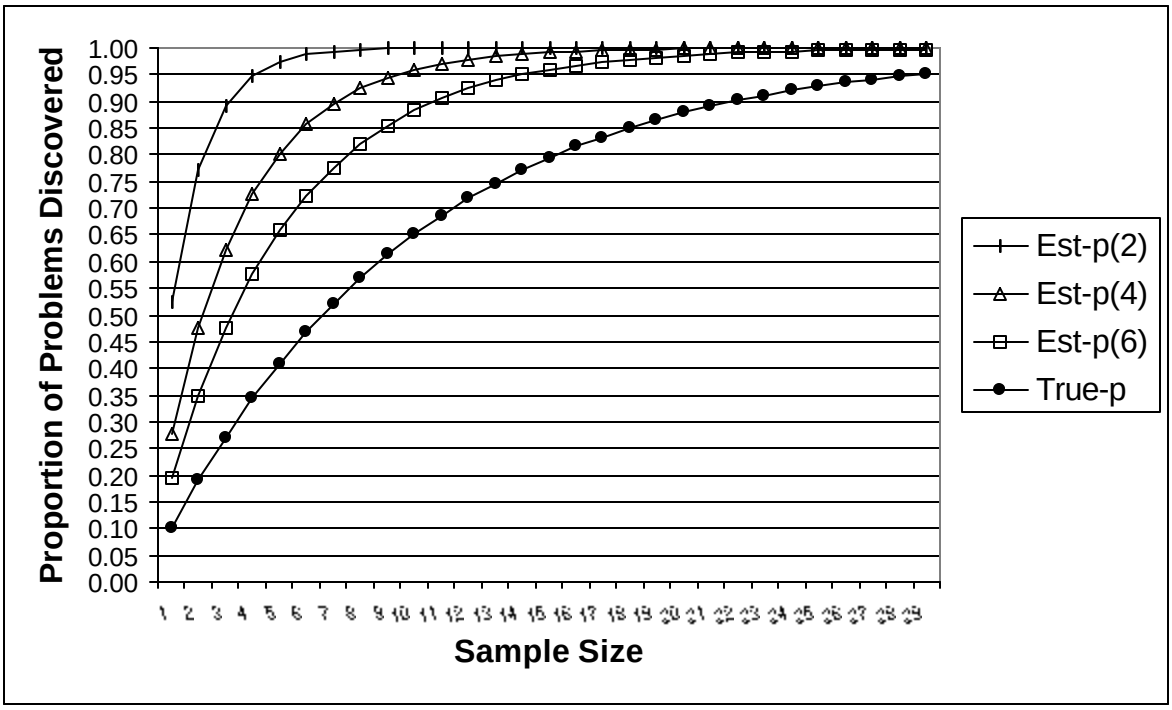

Figure 10. Projected problem discovery curves for MACERR10

Table 20. Projected sample size requirements for MACERR25

Projected Sample	**Est-p(2)**	**Est-p(4)**	**Est-p(6)**	**True-p**	***Sample Size***	**Goal: 90%**	**Goal: 95%**
1	0.550	0.341	0.282	0.250	*2*	5	7
2	0.798	0.566	0.484	0.438	*4*	2	3
3	**0.909**	0.714	0.630	0.578	*6*	1	1
4	***0.959***	0.811	0.734	0.684			
5	0.982	0.876	0.809	0.763			
6	0.992	**0.918**	0.863	0.822			
7	0.996	0.946	**0.902**	0.867			
8	0.998	***0.964***	0.929	**0.900**			
9	0.999	0.977	0.949	0.925			
10	1.000	0.985	***0.964***	0.944			
11	1.000	0.990	0.974	***0.958***			
12	1.000	0.993	0.981	0.968			
13	1.000	0.996	0.987	0.976			
14	1.000	0.997	0.990	0.982			
15	1.000	0.998	0.993	0.987			
16	1.000	0.999	0.995	0.990			
17	1.000	0.999	0.996	0.992			
18	1.000	0.999	0.997	0.994			
19	1.000	1.000	0.998	0.996			
20	1.000	1.000	0.999	0.997			

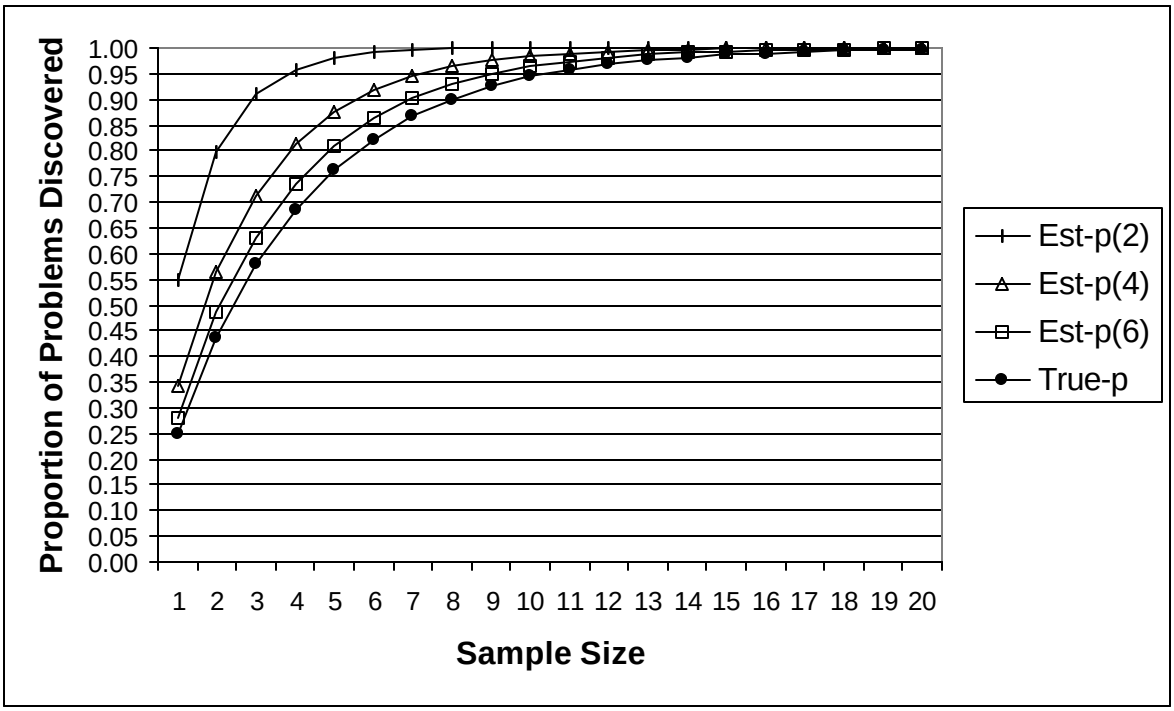

Figure 11. Projected problem discovery curves for MACERR25

Table 21. Projected sample size requirements for MACERR50

Projected Sample	**Est-p(2)**	**Est-p(4)**	**Est-p(6)**	**True-p**	***Sample Size***	**Goal: 90%**	**Goal: 95%**
1	0.645	0.519	0.502	0.500	*2*	1	2
2	0.874	0.769	0.752	0.750	*4*	0	0
3	***0.955***	0.889	0.876	0.875	*6*	0	0
4	0.984	**0.946**	**0.938**	**0.938**			
5	0.994	***0.974***	***0.969***	***0.969***			
6	0.998	0.988	0.985	0.984			
7	0.999	0.994	0.992	0.992			
8	1.000	0.997	0.996	0.996			
9	1.000	0.999	0.998	0.998			
10	1.000	0.999	0.999	0.999			
11	1.000	1.000	1.000	1.000			
12	1.000	1.000	1.000	1.000			
13	1.000	1.000	1.000	1.000			
14	1.000	1.000	1.000	1.000			
15	1.000	1.000	1.000	1.000			
16	1.000	1.000	1.000	1.000			
17	1.000	1.000	1.000	1.000			
18	1.000	1.000	1.000	1.000			
19	1.000	1.000	1.000	1.000			
20	1.000	1.000	1.000	1.000			

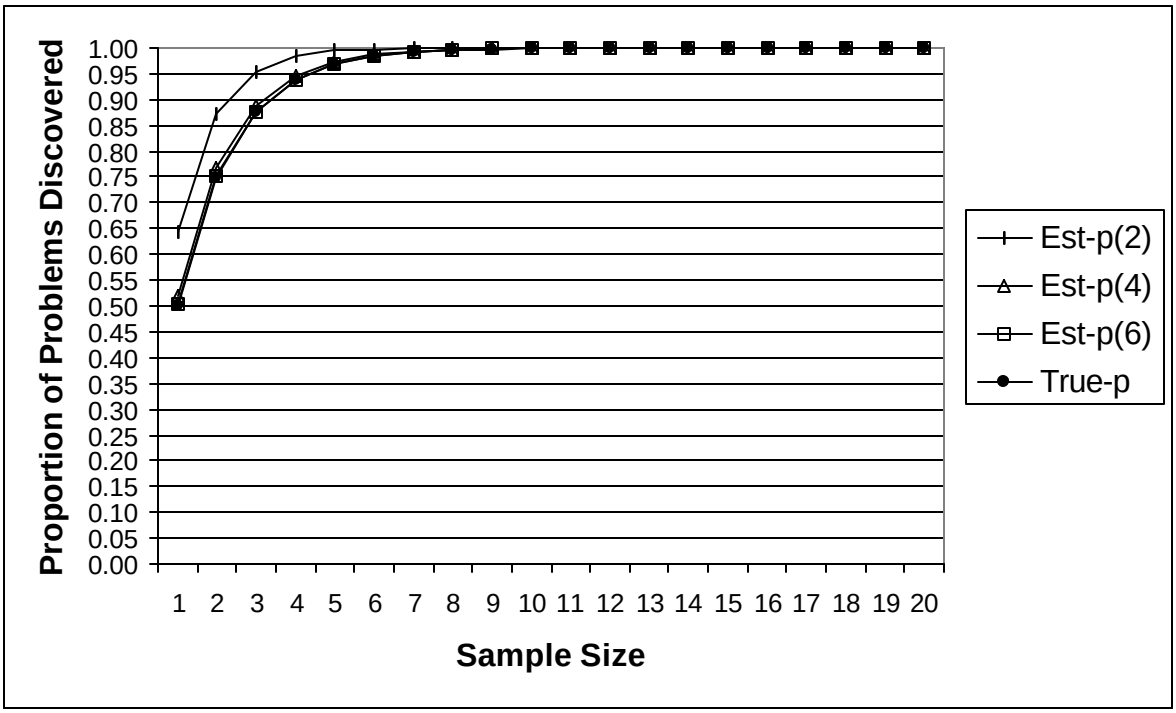

Figure 12. Projected problem discovery curves for MACERR50

Table 22. Projected sample size requirements for MACERR73

Projected Sample	**Est-p(2)**	**Est-p(4)**	**Est-p(6)**	**True-p**	***Sample Size***	**Goal: 90%**	**Goal: 95%**
1	0.785	0.734	0.733	0.730	*2*	0	1
2	***0.954***	**0.929**	**0.929**	**0.927**	*4*	0	0
3	0.990	***0.981***	***0.981***	***0.980***	*6*	0	0
4	0.998	0.995	0.995	0.995			
5	1.000	0.999	0.999	0.999			
6	1.000	1.000	1.000	1.000			
7	1.000	1.000	1.000	1.000			
8	1.000	1.000	1.000	1.000			
9	1.000	1.000	1.000	1.000			
10	1.000	1.000	1.000	1.000			
11	1.000	1.000	1.000	1.000			
12	1.000	1.000	1.000	1.000			
13	1.000	1.000	1.000	1.000			
14	1.000	1.000	1.000	1.000			
15	1.000	1.000	1.000	1.000			
16	1.000	1.000	1.000	1.000			
17	1.000	1.000	1.000	1.000			
18	1.000	1.000	1.000	1.000			
19	1.000	1.000	1.000	1.000			
20	1.000	1.000	1.000	1.000			

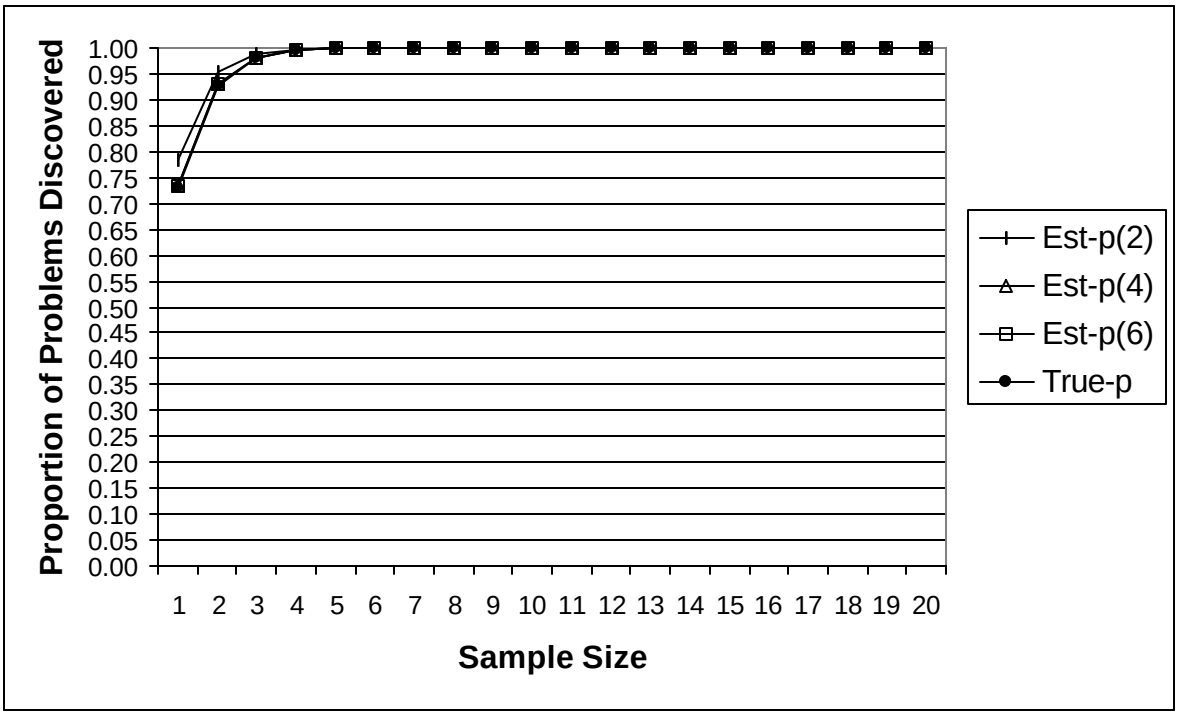

Figure 13. Projected problem discovery curves for MACERR73

Additional Analyses Using 98% Ranges

Tables 23 and 24 contain the results of additional analyses focusing on the 98% ranges of estimates of *p* as a function of sample size.

Table 23. 98% overestimation ratio ranges as a function of sample size

		n=2		n=4		n=6	
Source	***True p***	**99th**	**1st**	**99th**	**1st**	**99th**	**1st**
MACERR	*0.16*	4.0	3.1	2.4	1.9	1.9	1.4
VIRZI90	*0.36*	2.1	1.6	1.6	1.2	1.4	1.0
MANTEL	*0.38*	2.3	1.5	1.8	1.2	1.6	1.1
SAMPLE	*0.50*	1.8	1.1	1.3	1.0	1.1	0.9
MACERR10	*0.10*	7.5	5.0	3.8	2.5	2.3	1.7
MACERR25	*0.25*	2.7	2.0	1.7	1.0	1.5	0.8
MACERR50	*0.50*	1.8	1.0	1.4	0.7	1.3	0.7
MACERR73	*0.73*	1.4	0.7	1.3	0.7	1.2	0.8

Table 24. 98% ranges of p for sample sizes of 3 and 6

		n=3			n=6		
Source	***True p***	**Lower**	**Upper**	**Difference**	**Lower**	**Upper**	**Difference**
MACERR	*0.16*	0.364	0.475	0.111	0.231	0.303	0.072
VIRZI90	*0.36*	0.464	0.640	0.176	0.369	0.494	0.125
MANTEL	*0.38*	0.487	0.769	0.282	0.417	0.615	0.198
SAMPLE	*0.50*	0.500	0.762	0.262	0.467	0.574	0.107
MACERR10	*0.10*	0.333	0.500	0.167	0.167	0.227	0.060
MACERR25	*0.25*	0.333	0.536	0.203	0.205	0.374	0.169
MACERR50	*0.50*	0.375	0.741	0.366	0.367	0.650	0.283
MACERR73	*0.73*	0.444	1.000	0.556	0.556	0.889	0.333

Discussion

Estimates of Mean *p*

As shown in Figure 1, as the sample size increases, the estimate of mean *p* decreases for all databases. The slope of the decrease is greater for databases with smaller values of true *p*. For the published databases, the value of estimated *p* at a sample size of six participants did not reach the value of true *p*. Estimated *p* for the subset databases derived from MACERR did converge with true *p* except for MACERR10, which had the very low true *p* of .10. This difference in convergence property is most likely due to a reduced variability of *p* in the subset data caused by the nonrandom selection of problems to go into the subset database.

Mean *p* Overestimation Ratios

The curves in Figure 2 show that overestimation is very serious for small samples and small values of true *p*. For a sample size of two participants and true *p* equal to .10, the estimated value of *p* is over five times the value of true *p*. When the sample size is six participants, the overestimation has diminished considerably, with the greatest overestimation (for true *p* equal to .10) a little less than twice the value of true *p*.

Estimates of the Root Mean Square Error

The results for rms error (shown in Figure 3) mirror those for overestimation ratios. The values for rms error range from a high of .425 (true *p* equal to .10 with two-participant samples) to a low of .026 (true *p* equal to .73 with six-participant samples).

Percentile Estimates of *p* for Published Distributions

The interquartile ranges for the three published distributions (Figures 4-6) demonstrate that the variability of estimated *p* is quite low, even with sample sizes of two participants. MANTEL had the largest interquartile ranges, which spanned from a high of .088 with two-participant samples to a low of .060 with six-participant samples. One way to interpret these ranges is that an estimate of *p* based on two participants plus or minus half of the interquartile range is 50% likely to contain the median value of estimated *p*.

Expanding the coverage of the range to 98% of the distribution necessarily increases the size of the range. With two-participant samples, the size of this range varied from .146 (MACERR) to .304 (MANTEL). With six-participant samples, the variance was from .072 (MACERR) to .198 (MANTEL), with VIRZI90 in between (.125). This indicates that it is at least 98% likely that any single estimate of *p* based on six participants will be within .1 of the median estimate of *p*.

Effect of Overestimation of *p* on Projected Sample Size Requirements

The data in Tables 16-18 illustrate underestimation for sample sizes projected to reach goals of 90% or 95% problem discovery using estimated *p* even when the sample size is six participants. For the VIRZI90 and MANTEL databases, the underestimation is not serious when the sample size is six (one to two participants short), and isn't too bad when the sample size is four (two to

three participants short). For the MACERR database, which has a true *p* of .16 (which is about half that of the other two databases), the underestimation is serious even with the estimate of *p* based on six participants (six to eight participants short).

Tables 19-22, based on subsets of the MACERR database designed to achieve specific values of true *p* of .10, .25., .50 and .73, also illustrate the potential for sample size underestimation due to overestimation of *p*. Underestimation of the required sample size is not a problem (even with two-participant samples) when the true value of *p* exceeds .50. With six-participant samples, the underestimation for MACERR25 is only one participant for both 90% and 95% problem discovery goals. Underestimation is severe for MACERR10 (11 and 15 participants for 90% and 95% problem discovery goals, respectively).

Additional Analyses Using 98% Ranges

Most of the preceding analyses refer to the mean value of estimated *p*. Any single usability study, though, will provide only a single value of *p*. That value is more likely to be close to the mean of *p* than to be distant from it, but there is no guarantee.

Table 23 shows that with a sample size of two, even the 1st percentile of the distribution of estimated values for *p* is greater than true *p* for the published databases. The 99th percentile for MACERR10 with two-participant samples has an estimated value of *p* that is 7.5 times its true value. With six-participant samples, the value of estimated *p* is almost never more than twice the value of true *p*.

Table 24 shows that for the published studies and a sample size of three participants, the 98% ranges for VIRZI90 and MANTEL are similar, and are distinctly different from the range for MACERR (which is reasonable because true *p* for VIRZI90 and MANTEL is .36 and .38 respectively, in contrast to .16 for MACERR). The MACERR10 and MACERR25 subsets are similar, as are the MACERR50 and MACERR73 subsets. With six-participant samples, the distributions' differences are clearer, with no overlap between the 98% ranges for MACERR and either VIRZI90 or MANTEL (which do overlap). For the MACERR subsets, the 98% intervals still have some overlap, but the overlap is small enough that all four distributions appear to be different. This suggests that it might be possible to apply discounting methods such as Laplace's Law of Succession or Good-Turing estimation (Jelinek, 1997; Manning & Schutze, 1999) or multiple regression methods to calculate an adjusted value for small sample estimates of *p* to bring the estimate closer to the true value of *p*.

Conclusions

The primary conclusions to draw from these experiments and their analyses appear to be:

- *p* is a biased estimator of problem discovery likelihood, with the extent of the bias (overestimation) a function of sample size.
- Increasing the sample size from two to six participants reduces overestimation of *p*.
- Higher values of true *p* lead to reduced overestimation.

- When true *p* is reasonably high (> .50), the mean estimate of *p* rapidly converges on true *p* as the sample size increases from two to six participants.
- Overestimation of *p* is a serious problem whenever true *p* is very low (< .25) or when true *p* is moderately low (< .50) and the sample size is small (fewer than six participants).
- The magnitude of overestimation is sufficient to result in underestimation of required sample sizes for problem discovery usability studies, even when the study includes six participants.

A practitioner can never know the true value of *p* from a small-sample study. One consequence of this is that projecting sample size requirements from small-sample estimates of *p* is likely to result in underestimation of the required sample size. Unless true *p* is very small (< .25), though, the underestimated sample size will be fairly close to the true sample size requirement, especially if the estimate of *p* is based on six participants. For both VIRZI90 and MANTEL using a discovery criterion of 90% and estimates of *p* based on six participants, the projected sample size requirement was only one participant short of the true sample size requirement. The overestimation ratio for the 99th percentile of MACERR, the published study with the smallest value for true *p*, was 1.9. Thus, it seems reasonable (unless the cost of running participants is very high) to estimate *p* from small-sample usability studies six participants, then project the required sample size to accomplish the study's problem discovery goals using *p*/2 as an estimate of true *p*. Because this rule of thumb is likely to overcompensate in most cases, future work in this area should concentrate on the application of discounting methods or multiple regression to adjust small-sample estimates of *p* to address the overestimation problem more systematically.

References

Diamond, W. J. (1981). *Practical experiment designs for engineers and scientists.* Belmont, CA: Lifetime Learning Publications.

Hertzum, M., & Jacobsen, N. (In press). The evaluator effect in usability evaluation methods: A chilling fact about a burning issue. To appear in *The International Journal of Human-Computer Interaction*.

Jelinek, F. (1997). *Statistical methods for speech recognition*. Cambridge, MA: MIT Press.

Manning, C. D., & Schutze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. *In Conference Proceedings on Human Factors in Computing Systems – CHI90* (pp. 249-256). New York, NY: ACM.

Lewis, J. R. (1991). *Legitimate use of small sample sizes in usability studies: Three examples* (Tech. Report 54.594). Boca Raton, FL: International Business Machines Corp.

Lewis, J. R. (1994). Sample sizes for usability studies: Additional considerations. *Human Factors*, *36*, 368-378.

Lewis, J. R. (2000a). *Validation of Monte Carlo estimation of problem discovery likelihood* (Tech. Report in press). Raleigh, NC: International Business Machines Corp.

Lewis, J. R. (2000b). *Sample size estimation and use of substitute audiences* (Tech. report in press). Raleigh, NC: International Business Machines Corp.

Lewis, J. R., Henry, S. C., & Mack, R. L. (1990). Integrated office scenario benchmarks: A case study. In *Human Computer Interaction -- INTERACT '90* (pp. 337-343). Cambridge, England: Elsevier Science Publishers, IFIP.

Virzi, R. A. (1990). Streamlining the design process: Running fewer subjects. In *Proceedings of the Human Factors Society 34th Annual Meeting* (pp. 291-294). Santa Monica, CA: Human Factors Society.

Virzi, R. A. (1992). Refining the test phase of usability evaluation: How many subjects is enough? *Human Factors*, *34*, 443-451.

Walpole, R. E. (1976). *Elementary statistical concepts.* New York, NY: Macmillan.