Evaluation of Problem Discovery Rate Adjustment Procedures for Sample Sizes from Two to Ten

TR 29.3362

James R. Lewis

Speech Product Design and Usability

West Palm Beach, Florida

Abstract

Overestimation of problem discovery rates (*p*) in usability studies leads to underestimation of required sample sizes. Previous experiments indicated that combining normalization and Good-Turing discounting resulted in accurate adjustment of *p*. The previous experiments stopped at a sample size of six, however, and the data suggested that the procedure might become less accurate at larger sample sizes. The current experiment demonstrated that the procedure remains accurate through a sample size of ten. Iterative estimation of required sample sizes at sample sizes of two and four using the combined adjustment of *p* resulted in extremely accurate estimation of required sample sizes.

ITIRC Keywords

Monte Carlo estimation
problem discovery likelihood
overestimation of *p*
usability evaluation
sample size estimation
discounting
Good-Turing estimator
multiple regression
normalization of *p*

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, 2000c). Recently, though, Hertzum and Jacobsen (in press) pointed out that this practice will almost always result in overestimation of the value of *p*, with consequent overestimation of the proportion of problems discovered and underestimation of required sample sizes.

Overestimation 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 (2000e) 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. A follow-on Monte Carlo study (Lewis, 2000a) demonstrated that the extent of overestimation of *p* led to underestimation

of required sample size. Additional follow-on Monte Carlo studies (Lewis, 2000b, 2000d) evaluated the use of discounting techniques, normalization, and linear regression to adjust the observed value of p to be closer to the true value of p when estimating p from sample sizes that ranged from two to six.

Discounting *p* with the Good-Turing Estimator

One way to adjust an observed estimate of p to bring it closer to the true value of p is to use a discounting procedure. There are many discounting procedures, all of which attempt to allocate some amount of probability space to unseen events. Discounting procedures receive wide use in the field of statistical natural language processing, especially in the construction of language models (Manning and Schutze, 1999). One of the best known discounting procedures is the Good-Turing estimator (Jelinek, 1997; Manning and Schutze, 1999). There are a number of paths that lead to the derivation of the Good-Turing estimator, but the end result is that the total probability mass reserved for unseen events is $E(N_1)/N$, where $E(N_1)$ is the expected number of events that happen exactly once and N is the total number of events. For a given sample, the usual value used for $E(N_1)$ is the actually observed number of events that occurred once. In the context of a problem discovery usability study, the events are problems. Applying this to the example shown in Table 3, $E(N_1)$ would be the observed number of problems that happened exactly once (2 in the example) and N would be the total number of problems (7 in the example). Thus, $E(N_1)/N$ is 2/7, or .286. To add this to the total probability space and adjust the original estimate of p would result in .762/(1+.286), or .592 – still an overestimate, but much closer to the true value of p.

A Monte Carlo study (Lewis, 2000d) investigated the use of the Good-Turing estimator to discount (adjust to a lower value) the original estimate of p. Using this adjusted value of p led to more accurate projection of required sample sizes. Estimates of p adjusted with the Good-Turing method still tended to overestimate p slightly, resulting in a slight underestimation of required sample sizes, especially when the sample size used to estimate p included fewer than five participants.

Normalizing *p*

In their discussion of the problem of overestimation of p, Hertzum and Jacobsen (in press) pointed out that the smallest possible value of p from a small sample problem discovery study is $1/n$. Suppose that an investigator stops data collection after observing one participant. The discovery rate is necessarily 1.000 because the participant by problem matrix will have one row with an 'x' in every cell. Suppose the investigator stops after observing two participants, and each participant has experienced five problems. If there is no overlap, then the matrix will have two rows and ten columns with ten cells containing an 'x' for a discovery rate of .500. If there is any overlap (participants experienced at least one common problem), then the number of cells containing an 'x' will continue to be ten, but the number of columns will be less than ten, inflating the estimate of p. With larger sample sizes, the effect of this limit on the lowest possible value of p becomes less important. For example, if a study includes 20 participants, then the lower limit for p is 1/20, or .05, which is reasonably close to 0.

With the knowledge of this lower limit determined by the sample size, it is possible to normalize a small sample estimate of *p* in the following way. Subtract from the original estimate of *p* the lower limit, 1/*n*. Then, to normalize this value to a scale of 0 to 1, multiply it by (1 – 1/*n*). For the estimate of *p* generated from the data in Table 3, the first step would result in the subtraction of .333 from .762, or .429. The second step would be the multiplication of .429 by .667, resulting in .286. In this particular case, the result underestimated true *p* by a fair amount. Monte Carlo experiments (Lewis, 2000b) indicated that normalization has a tendency to underestimate true *p*.

Combining Normalized and Good-Turing Estimates of *p*

Because normalization is fairly accurate but tends to underestimate *p* and the Good-Turing method is fairly accurate but tends to overestimate *p*, it is reasonable to hypothesize that taking the average of these estimates should provide a very accurate estimate of true *p*. Monte Carlo experiments (Lewis, 2000b) supported this hypothesis. The formula for this combined estimate is:

[1] $truep = \frac{1}{2}[(estp - 1/n)(1 - 1/n)] + \frac{1}{2}[estp/(1+GTadj)]$

where *GTadj* is the Good-Turing adjustment to probability space, which is the proportion of the number of problems that occurred once divided by the number of different problems. The *estp*/(1+*GTadj*) component in the equation produces the Good-Turing adjusted estimate of *p* by dividing the observed, unadjusted estimate of *p* (*estp*) by the Good-Turing adjustment to probability space. The (*estp* – 1/*n*)(1 – 1/*n*) component in the equation produces the normalized estimate of *p* from the observed, unadjusted estimate of *p* and *n* (the sample size used to estimate *p*).

Adjusting *p* with Linear Regression

Regression equations created from Monte Carlo sampling of subsets of a problem discovery database were not as effective as discounting or normalization (or their combination) for adjusting observed values of *p*, at least not within the range of investigated sample sizes (Lewis, 2000b). Of the six evaluated regression equations in that study, the two most accurate were Regression 2 (Reg2) and Regression 5 (Reg5). The equations for these regressions were:

[2] Reg2: $truep = .16 + .823(normp)$

[3] Reg5: $truep = .210 + .829(normp) - .013(n)$

where *truep* is the true value of *p*, *normp* is the normalized estimate of the observed value of *p* and *n* is the sample size used to estimate *p*.

An Important Limitation of the Previous Studies

The experiments in Lewis (2000b) supported the use of the normalization/Good-Turing combined estimator for sample sizes from two to six. Figure 1 shows one of the results from this study.

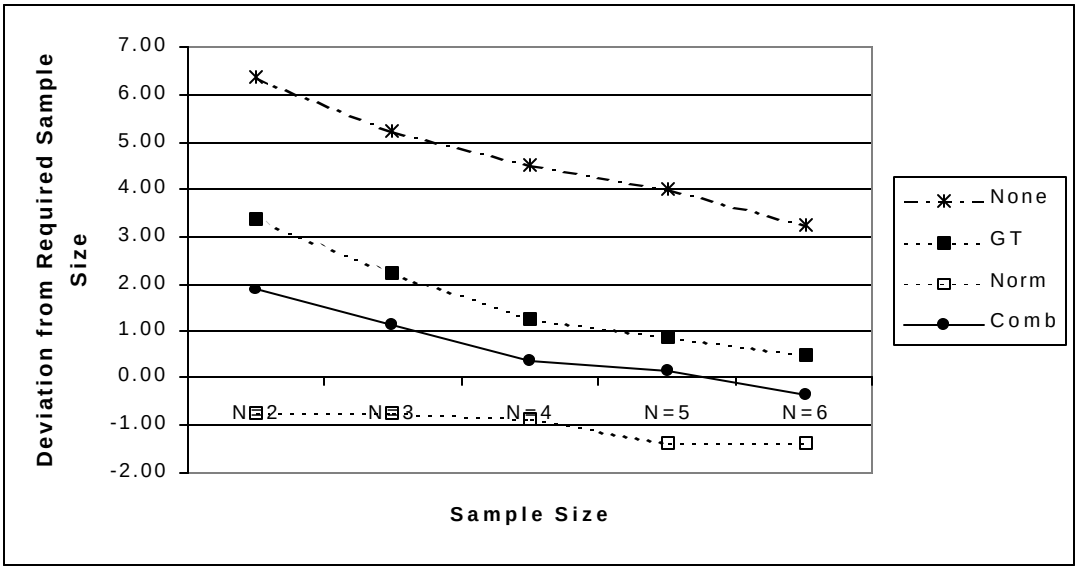

Figure 1. Sample size by adjustment type interaction (from Lewis, 2000b)

The lines in the figure show the mean deviation from the sample size required to achieve 90% and 95% problem discovery goals from four published usability problem databases. Normalization appeared to work best at sample sizes of two and three, with the combination method working best for sample sizes from four to six. The apparent trends in Figure 1 suggest, however, that these results might not generalize beyond six participants. At six participants, normalization continues to underestimate *p*, and the combination approach has begun to underestimate *p* slightly. The Good-Turing approach appears to be getting closer to true *p* and, as the sample size continues to increase, the unadjusted estimate of *p* should continue to approach true *p*. It isn't clear from the figure whether the combination approach will continue to be the best approach for slightly larger sample sizes because the estimate might be converging at a value just below 0.00 or might continue to decline.

Purposes of the Current Experiment

The primary purposes of the current experiment were to:

- Investigate a variety of approaches (regression, normalization, Good-Turing, the combination estimator) for adjusting observed estimates of *p* to bring them closer to true *p* using sample sizes from two to ten participants to estimate *p*. In particular, does the combination approach continue to provide accurate estimates of true *p* for sample sizes from seven to ten participants?
- Investigate the extent to which inaccuracy in estimating problem discovery sample size using these methods affects the true proportion of discovered problems. The previous investigations assessed the deviation from the sample size required to achieve 90% and 95% problem discovery goals, but did not assess the deviation from the problem discovery goals caused by overestimating or underestimating the required sample size.

Method

Measurements

I modified the BASIC program used in Lewis (2000b, Experiment 2) so it would produce the following measurements from Monte Carlo iterations for each member of a set of four problem discovery databases estimating *p* with sample sizes ranging from two to ten:

- mean value of *p*
- standard deviation of *p*
- overestimation ratios of calculated *p* over true *p*
- deviation scores for calculated *p* minus true *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*
- the interquartile range (the size of the interval encompassing the central 50% of the distribution of *p*)
- the 90% range (the size of the interval encompassing the central 90% of the distribution of *p*)

The program produced this set of statistics for the unadjusted estimate of *p* and the following adjusted estimates (based on 1000 Monte Carlo iterations for each problem discovery database at each level of sample size from two to ten):

- Normalization
- Regression formula 2
- Regression formula 5
- Good-Turing estimation
- Combined normalization/Good-Turing

Problem Discovery Databases

I ran the program on a Micron Millennia computer (Windows 95, 64 MB memory) to evaluate the following published problem discovery databases:

- MACERR (Lewis, 1994; Lewis, Henry, & Mack, 1990): This database came from a scenario-driven problem-discovery usability study conducted to develop usability benchmark values for an integrated office system (word processor, mail application, calendar application and spreadsheet). Fifteen employees of a temporary employee agency, observed by a highly experienced usability practitioner, completed 11 scenarios-of-use with the system. Participants typically worked on the scenarios for about six hours, and the study uncovered 145 different usability problems. The problem discovery rate (*p*) for this study was .16. Participants did not think aloud in this study.

- VIRZI90 (Virzi, 1990; 1992): The problems in this database came from a scenario-driven problem-discovery usability study conducted to evaluate a computer-based appointment calendar. The participants were 20 university undergraduates with little or no computer experience. The participants completed 21 scenarios-of-use under a think-aloud protocol, observed by two experimenters. The experimenters identified 40 separate usability problems. The problem discovery rate (*p*) for this study was .36.
- MANTEL (Nielsen & Molich, 1990): These usability problems came from 76 submissions to a contest presented in the Danish edition of *Computerworld*. The evaluators were primarily computer professionals who evaluated a written specification (not a working program) for a design of a small information system with which users could dial in to find the name and address associated with a telephone number. The specification contained a single screen and a few system messages, which the participants evaluated using a set of heuristics. The evaluators described 30 distinct usability problems. The problem discovery rate (*p*) for this study was .38.
- SAVE (Nielsen & Molich, 1990): For this study, 34 computer science students taking a course in user interface design performed heuristic evaluations of an interactive voice response system (working and deployed) designed to give banking customers information such as their account balances and currency exchange rates. The participants uncovered 48 different usability problems with a problem discovery rate (*p*) of .26.

(For copies of the MACERR, VIRZI90, and MANTEL databases, see Lewis, 2000e. For the SAVE database, see Lewis, 2000d.)

Because these usability problem databases had considerable variation in total number of participants, total number of usability problems uncovered, basic problem discovery rate, and method of execution (observational vs. heuristic, with differences in error classification procedures), these results stand a good chance of generalizing to other problem discovery databases (Chapanis, 1988).

Results

Adequacy of Randomness of Monte Carlo Sampling

To check the adequacy of random sampling of participants from the problem-discovery databases with the Monte Carlo program, I conducted χ^2 analyses for each combination of source database and sample size. As shown in Table 4, The observed significance levels (osl) for the χ^2 tests indicate adequately random sampling by the Monte Carlo program.

Table 4. Results of c^2 tests to check adequacy of random sampling

Source	*Sample Size*	Chi-Squared	osl
MACERR (df=14)	*2*	8.5	0.86
	3	9.2	0.82
	4	17.9	0.21
	5	12.9	0.53
	6	8.1	0.89
	7	5.9	0.97
	8	3.7	1.00
	9	8.7	0.85
	10	2.2	1.00
VIRZI90 (df=20)	*2*	25.0	0.16
	3	13.6	0.81
	4	17.5	0.55
	5	14.1	0.78
	6	13.8	0.79
	7	11.2	0.92
	8	19.1	0.45
	9	10.0	0.95
	10	8.8	0.98
MANTEL (df=75)	*2*	59.2	0.91
	3	61.1	0.88
	4	72.3	0.57
	5	60.1	0.89
	6	60.3	0.89
	7	74.3	0.50
	8	69.5	0.66
	9	70.1	0.64
	10	54.4	0.97
SAVE (df=33)	*2*	28.5	0.69
	3	20.2	0.96
	4	24.3	0.86
	5	26.0	0.80
	6	14.0	1.00
	7	21.5	0.94
	8	31.4	0.55
	9	25.7	0.81
	10	28.7	0.68

Estimates of Problem Discovery Rates (*p*)

An analysis of variance (within-subjects ANOVA treating problem discovery databases as subjects) conducted on the problem discovery rates for each of the six estimation methods at each of the nine levels of sample size used to estimate *p* revealed:

- a significant main effect for type of adjustment ($F(8,24)=92.6$, $p=.00000015$)
- a significant main effect of sample size ($F(5,15)=138.5$, $p=.00000015$)
- a significant adjustment type by sample size interaction ($F(40,120)=72.8$, $p=.0001$)

The patterns for these effects appear in Table 5, with the interaction represented by the interior of the table and the main effects represented as row (main effect of sample size) and column (main effect of adjustment type) totals. Figure 2 illustrates the interaction.

Table 5. Problem discovery rate effects

Sample	**None**	**Norm**	**Reg2**	**Reg5**	**GT**	**Comb**		**Average**
2	0.646	0.291	0.425	0.400	0.385	0.338		0.414
3	0.524	0.286	0.408	0.395	0.342	0.314		0.378
4	0.461	0.282	0.391	0.391	0.318	0.300		0.357
5	0.424	0.280	0.377	0.390	0.305	0.292		0.345
6	0.396	0.275	0.360	0.386	0.293	0.284		0.332
7	0.377	0.274	0.346	0.385	0.287	0.280		0.325
8	0.362	0.271	0.331	0.383	0.280	0.276		0.317
9	0.351	0.270	0.316	0.382	0.277	0.273		0.311
10	0.342	0.269	0.303	0.381	0.275	0.271		0.307
Average	0.431	0.277	0.362	0.388	0.307	0.292		0.343

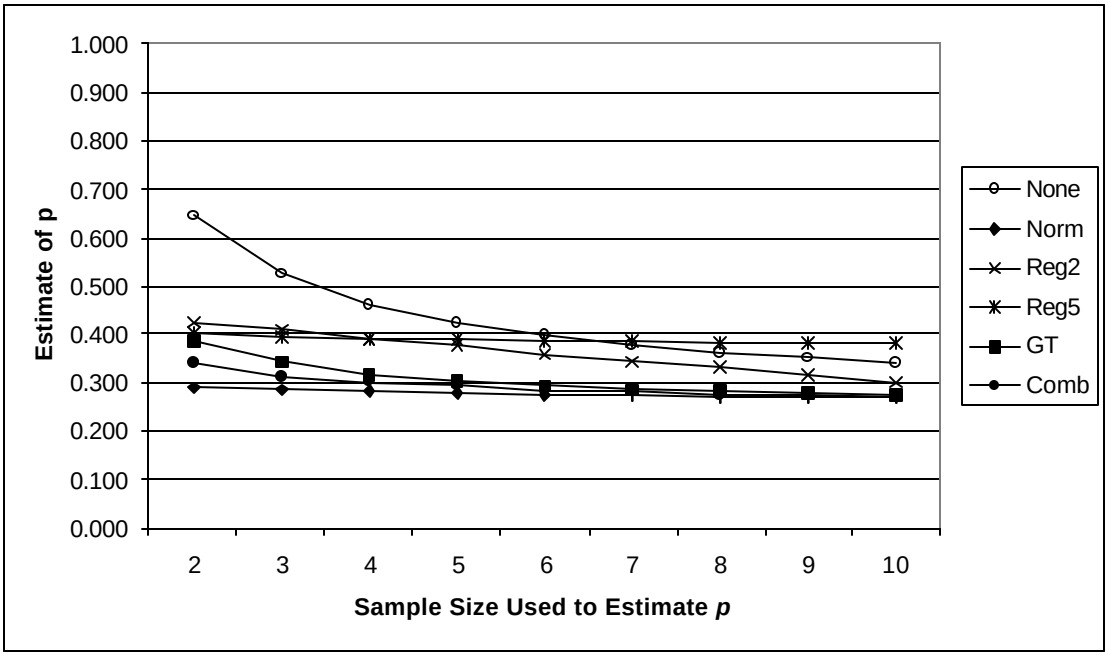

Figure 2. Adjustment type by sample size interaction for problem discovery rates

Estimation Accuracy

Overestimation Ratios

The overestimation ratio is the ratio of the estimate of *p* over the known true value of *p* for a given problem discovery database. An analysis of variance conducted on the overestimation ratios revealed:

- a significant main effect for type of adjustment ($F(8,24)=11.3$, $p=.000002$)
- a significant main effect of sample size ($F(5,15)=13.0$, $p=.00006$)
- a significant adjustment type by sample size interaction ($F(40,120)=10.8$, $p=.0007$)

The patterns for these effects appear in Table 6 and Figure 3.

Table 6. Overestimation ratio effects

Sample	**None**	**Norm**	**Reg2**	**Reg5**	**GT**	**Comb**		**Average**
2	2.430	0.980	1.525	1.428	1.418	1.200		1.497
3	1.938	0.968	1.463	1.415	1.228	1.095		1.351
4	1.683	0.950	1.403	1.403	1.125	1.035		1.266
5	1.530	0.948	1.345	1.398	1.068	1.008		1.216
6	1.420	0.928	1.280	1.385	1.020	0.975		1.168
7	1.345	0.920	1.228	1.380	0.990	0.960		1.137
8	1.283	0.913	1.168	1.373	0.968	0.940		1.107
9	1.238	0.905	1.113	1.365	0.950	0.930		1.083
10	1.200	0.900	1.058	1.360	0.938	0.920		1.063
Average	1.563	0.934	1.287	1.389	1.078	1.007		1.210

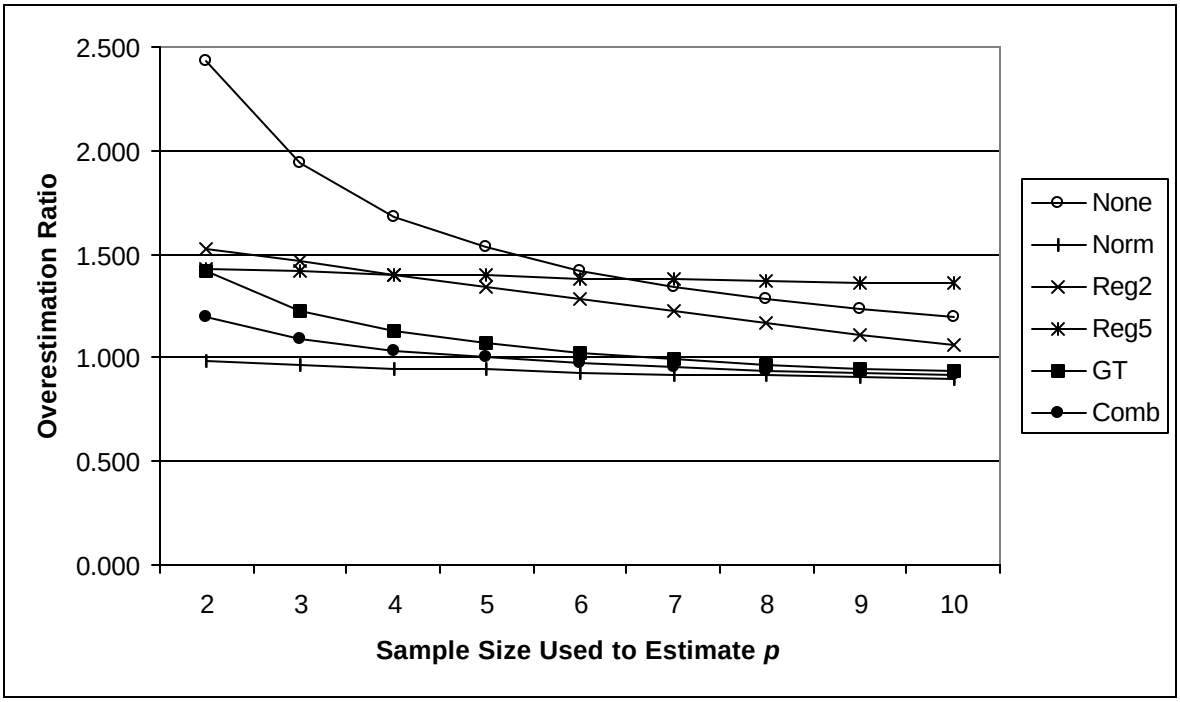

Figure 3. Adjustment type by sample size interaction for overestimation ratios

Deviation from True *p*

The deviation from true *p* is the measure of the observed estimate of *p* minus the known true value of *p*. An analysis of variance conducted on these deviation scores revealed:

- a significant main effect for type of adjustment ($F(8,24)=92.6$, $p=.00000015$)
- a significant main effect of sample size ($F(5,15)=138.5$, $p=.00000015$)
- a significant adjustment type by sample size interaction ($F(40,120)=72.8$, $p=.0001$)

The patterns for these effects appear in Table 7 and Figure 4.

Table 7. Effects for deviation from true p

Sample	**None**	**Norm**	**Reg2**	**Reg5**	**GT**	**Comb**		**Average**
2	0.356	0.001	0.135	0.110	0.095	0.048		0.124
3	0.234	-0.004	0.118	0.105	0.052	0.024		0.088
4	0.171	-0.009	0.101	0.101	0.028	0.010		0.067
5	0.134	-0.010	0.087	0.100	0.015	0.002		0.055
6	0.106	-0.015	0.070	0.096	0.003	-0.006		0.042
7	0.087	-0.017	0.056	0.095	-0.003	-0.010		0.035
8	0.072	-0.019	0.041	0.093	-0.010	-0.015		0.027
9	0.061	-0.021	0.026	0.092	-0.013	-0.017		0.021
10	0.052	-0.022	0.013	0.091	-0.016	-0.019		0.017
Average	0.141	-0.013	0.072	0.098	0.017	0.002		0.053

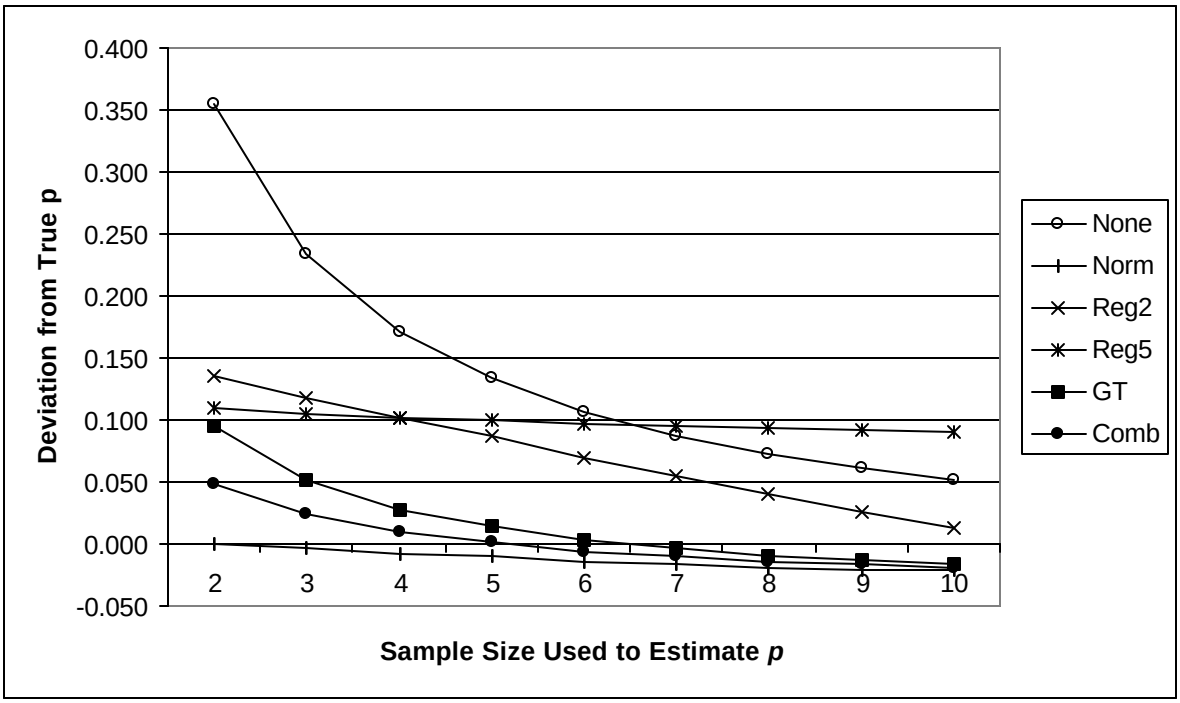

Figure 4. Adjustment type by sample size interaction for deviation from true p

Root Mean Square (rms) Error

The root mean square (rms) error is the average squared deviation of estimates of *p* from the known true value of *p* in these databases. Whereas overestimation ratios and deviation scores are primarily sensitive to mean differences, rms error is sensitive to both mean differences and differences in distributional variance. Thus, the rms error is an excellent measure of accuracy to use to assess adjustment procedures. An analysis of variance conducted on rms error revealed:

- a significant main effect for type of adjustment ($F(8,24)=120.6$, $p=.00000004$)
- a significant main effect of sample size ($F(5,15)=27.2$, $p=.0000006$)
- a significant adjustment type by sample size interaction ($F(40,120)=32.9$, $p=.0000001$)

The patterns for these effects appear in Table 8 and Figure 5.

Table 8. Effects for rms error

Sample	**None**	**Norm**	**Reg2**	**Reg5**	**GT**	**Comb**		**Average**
2	0.361	0.106	0.160	0.139	0.115	0.095		0.162
3	0.240	0.074	0.132	0.120	0.078	0.067		0.118
4	0.178	0.062	0.112	0.112	0.059	0.054		0.096
5	0.140	0.056	0.096	0.108	0.050	0.049		0.083
6	0.112	0.051	0.078	0.103	0.042	0.045		0.072
7	0.094	0.049	0.064	0.101	0.041	0.044		0.065
8	0.079	0.047	0.051	0.098	0.040	0.043		0.060
9	0.067	0.045	0.040	0.097	0.040	0.042		0.055
10	0.058	0.044	0.032	0.095	0.039	0.041		0.051
Average	0.148	0.059	0.085	0.108	0.056	0.053		0.085

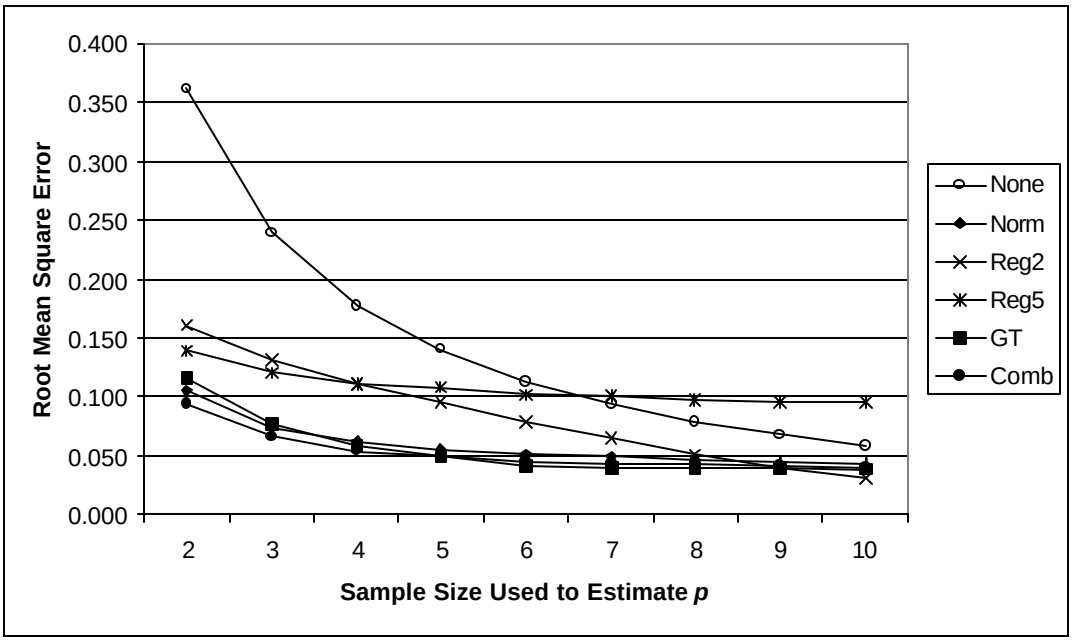

Figure 5. Adjustment type by sample size interaction for rms error

Estimation Variability

Interquartile Ranges

The interquartile range is the size of the interval that contains the central 50% of a distribution (the range from the 25th to the 75th percentile). The smaller this range, the less variable is the distribution. An analysis of variance conducted on interquartile ranges revealed:

- a significant main effect for type of adjustment ($F(8,24)=109.1$, $p=.0000003$)
- a significant main effect of sample size ($F(5,15)=10.3$, $p=.0002$)
- a significant adjustment type by sample size interaction ($F(40,120)=69.8$, $p=.00000001$)

The patterns for these effects appear in Table 9 and Figure 6.

Table 9. Effects for interquartile range

Sample	**None**	**Norm**	**Reg2**	**Reg5**	**GT**	**Comb**		**Average**
2	0.066	0.131	0.109	0.108	0.070	0.100		0.097
3	0.058	0.087	0.072	0.072	0.064	0.075		0.071
4	0.050	0.067	0.056	0.055	0.056	0.061		0.057
5	0.046	0.058	0.047	0.047	0.053	0.055		0.051
6	0.040	0.048	0.040	0.039	0.048	0.047		0.044
7	0.038	0.044	0.037	0.037	0.046	0.044		0.041
8	0.033	0.038	0.031	0.031	0.041	0.040		0.035
9	0.031	0.035	0.029	0.029	0.040	0.037		0.034
10	0.029	0.032	0.027	0.027	0.038	0.035		0.031
Average	0.043	0.060	0.050	0.049	0.051	0.055		0.051

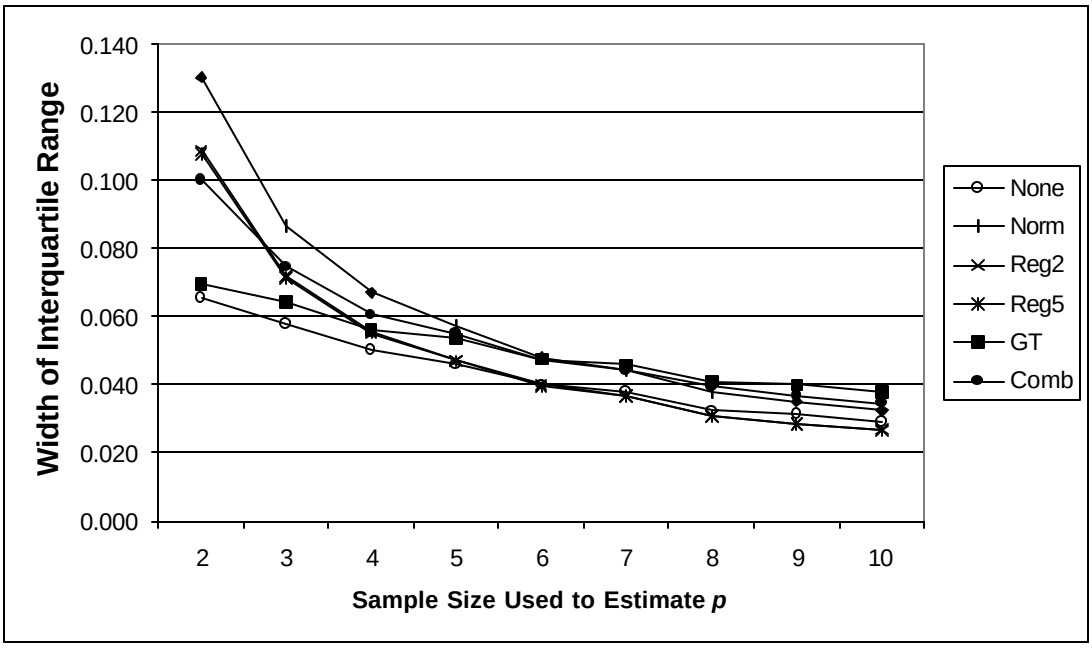

Figure 6. Adjustment type by sample size interaction for interquartile range

90% Ranges

The 90% range is the size of the interval that contains the central 90% of a distribution (the range from the 5th to the 95th percentile). The smaller this range, the less variable is the distribution. An analysis of variance conducted on 90% ranges revealed:

- a significant main effect for type of adjustment ($F(8,24)=71.6$, $p=.00000015$)
- a significant main effect of sample size ($F(5,15)=11.1$, $p=.0001$)
- a significant adjustment type by sample size interaction ($F(40,120)=44.1$, $p=.00000001$)

The patterns for these effects appear in Table 10 and Figure 7.

Table 10. Effects for 90% range

Sample	**None**	**Norm**	**Reg2**	**Reg5**	**GT**	**Comb**		**Average**
2	0.151	0.303	0.252	0.249	0.166	0.234		0.226
3	0.140	0.210	0.174	0.173	0.158	0.182		0.173
4	0.122	0.163	0.134	0.133	0.138	0.149		0.140
5	0.108	0.135	0.112	0.112	0.126	0.129		0.120
6	0.094	0.113	0.094	0.093	0.116	0.114		0.104
7	0.091	0.106	0.088	0.087	0.109	0.106		0.098
8	0.080	0.092	0.076	0.075	0.103	0.096		0.087
9	0.076	0.086	0.071	0.071	0.097	0.090		0.082
10	0.072	0.079	0.066	0.065	0.092	0.083		0.076
Average	0.104	0.143	0.118	0.118	0.123	0.131		0.123

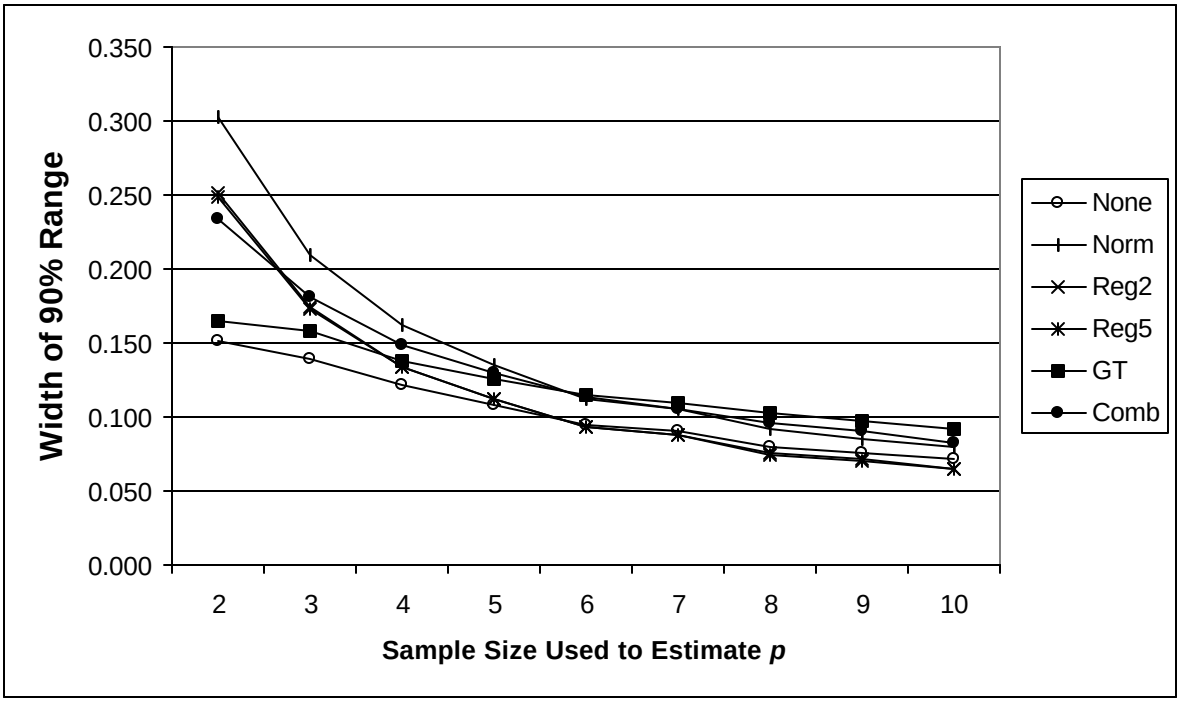

Figure 7. Adjustment type by sample size interaction for 90% range

Projecting Required Sample Sizes

The following analyses show (Tables 11-46, Figures 8-19), for each type of adjustment, for each database, and for estimates based on sample sizes from two to ten participants, the difference in projected sample sizes for studies having the goal of uncovering 90% and 95% of the usability problems in a product for unadjusted estimates of p and p adjusted by averaging estimates of p from the various procedures. The tables and figures use the cumulative binomial probability formula $1-(1-p)^n$ to project the proportion of problem discovery as a function of sample size (Lewis, 1982, 1994; Nielsen & Landauer, 1993; Virzi, 1992). Each subsection provides projections for two of the adjustment procedures evaluated in this experiment. The proportion of discovery in every table has a precision of three significant digits, and a cell with bold text indicates the smallest projected sample size for that row to achieve 90% problem discovery. Bold italic text indicates the projected sample size for 95% problem discovery. Additional tables in each section provide information about:

- the deviation in projected sample size from the truly required sample size as a function of problem discovery goal and sample size
- the deviation in projected proportion of discovered problems as a function of problem discovery goal and sample size

The values for the deviation in projected sample size from the truly required sample size come from analysis of the entries in the sample size projection tables. For example, in Table 11 the value for None (no adjustment) for MACERR for a sample size of two and a problem discovery goal of 90% is 11 (an underestimation of 11 participants). The value of 11 is the difference between the sample size required to achieve at least 90% problem discovery from the "True p" column in the table (14 participants) and the projected sample size required for 90% problem discovery from the "None 2" column in the table (3 participants). Positive numbers indicate underestimation of the required sample size; negative numbers indicate overestimation.

The values for the deviation in projected proportion of discovered problems also comes from analysis of the entries in the sample size projection tables. Using the same example (Table 11, MACERR, no adjustment, sample size of 2), the deviation is -.493. This is the difference between the true proportion of problems discovered with three participants as shown in the "True p" column (.407) minus the stated discovery goal of a proportion of .90 (90%). Negative numbers indicate underachievement of the problem discovery goal; positive numbers indicate overachievement.

No Adjustment and Normalization/Good-Turing Combined Estimation

Tables 11, 14, 17, and 20, and Figures 8-11 illustrate the sample size projections for unadjusted *p* (None) and *p* adjusted with the combination normalization/Good-Turing estimate (Comb). Tables 12, 15, 18, and 21 and Tables 13, 16, 19, and 22 show the deviations from required sample size and proportion of discovered problems, respectively, for each problem discovery database with these adjustment procedures.

Table 11. Unadjusted and combination estimate problem discovery projections for MACERR

N	None 2	Comb 2	None 3	Comb 3	None 4	Comb 4	None 5	Comb 5	None 6	Comb 6	None 7	Comb 7	None 8	Comb 8	None 9	Comb 9	None 10	Comb 10	True *p*
1	0.566	0.218	0.421	0.185	0.346	0.165	0.301	0.154	0.269	0.143	0.245	0.136	0.227	0.130	0.213	0.125	0.201	0.121	0.160
2	0.812	0.388	0.665	0.336	0.572	0.303	0.511	0.284	0.466	0.266	0.430	0.254	0.402	0.243	0.381	0.234	0.362	0.227	0.294
3	**0.918**	0.522	0.806	0.459	0.720	0.418	0.658	0.395	0.609	0.371	0.570	0.355	0.538	0.341	0.513	0.330	0.490	0.321	0.407
4	***0.965***	0.626	0.888	0.559	0.817	0.514	0.761	0.488	0.714	0.461	0.675	0.443	0.643	0.427	0.616	0.414	0.592	0.403	0.502
5	0.985	0.708	**0.935**	0.640	0.880	0.594	0.833	0.567	0.791	0.538	0.755	0.519	0.724	0.502	0.698	0.487	0.674	0.475	0.582
6	0.993	0.771	***0.962***	0.707	**0.922**	0.661	0.883	0.633	0.847	0.604	0.815	0.584	0.787	0.566	0.762	0.551	0.740	0.539	0.649
7	0.997	0.821	0.978	0.761	0.949	0.717	**0.918**	0.690	0.888	0.660	0.860	0.641	0.835	0.623	0.813	0.607	0.792	0.595	0.705
8	0.999	0.860	0.987	0.805	***0.967***	0.764	0.943	0.738	**0.918**	0.709	0.894	0.689	0.873	0.672	0.853	0.656	0.834	0.644	0.752
9	0.999	0.891	0.993	0.841	0.978	0.803	***0.960***	0.778	0.940	0.751	**0.920**	0.732	**0.901**	0.714	0.884	0.699	0.867	0.687	0.792
10	1.000	**0.914**	0.996	0.871	0.986	0.835	0.972	0.812	***0.956***	0.786	0.940	0.768	0.924	0.752	**0.909**	0.737	0.894	0.725	0.825
11	1.000	0.933	0.998	0.895	0.991	0.862	0.981	0.841	0.968	0.817	***0.955***	0.800	0.941	0.784	0.928	0.770	**0.915**	0.758	0.853
12	1.000	0.948	0.999	**0.914**	0.994	0.885	0.986	0.866	0.977	0.843	0.966	0.827	***0.954***	0.812	0.944	0.799	0.932	0.787	0.877
13	1.000	***0.959***	0.999	0.930	0.996	**0.904**	0.990	0.886	0.983	0.865	0.974	0.850	0.965	0.836	***0.956***	0.824	0.946	0.813	0.896
14	1.000	0.968	1.000	0.943	0.997	0.920	0.993	**0.904**	0.988	0.885	0.980	0.871	0.973	0.858	0.965	0.846	***0.957***	0.836	**0.913**
15	1.000	0.975	1.000	***0.954***	0.998	0.933	0.995	0.919	0.991	**0.901**	0.985	0.888	0.979	0.876	0.972	0.865	0.965	0.856	0.927
16	1.000	0.980	1.000	0.962	0.999	0.944	0.997	0.931	0.993	0.915	0.989	**0.904**	0.984	0.892	0.978	0.882	0.972	0.873	0.939
17	1.000	0.985	1.000	0.969	0.999	***0.953***	0.998	0.942	0.995	0.927	0.992	0.917	0.987	**0.906**	0.983	0.897	0.978	0.888	0.948
18	1.000	0.988	1.000	0.975	1.000	0.961	0.998	***0.951***	0.996	0.938	0.994	0.928	0.990	0.918	0.987	**0.910**	0.982	**0.902**	***0.957***
19	1.000	0.991	1.000	0.979	1.000	0.967	0.999	0.958	0.997	0.947	0.995	0.938	0.992	0.929	0.989	0.921	0.986	0.914	0.964
20	1.000	0.993	1.000	0.983	1.000	0.973	0.999	0.965	0.998	***0.954***	0.996	0.946	0.994	0.938	0.992	0.931	0.989	0.924	0.969
21	1.000	0.994	1.000	0.986	1.000	0.977	0.999	0.970	0.999	0.961	0.997	***0.954***	0.996	0.946	0.993	0.939	0.991	0.933	0.974
22	1.000	0.996	1.000	0.989	1.000	0.981	1.000	0.975	0.999	0.966	0.998	0.960	0.997	***0.953***	0.995	0.947	0.993	0.941	0.978
23	1.000	0.997	1.000	0.991	1.000	0.984	1.000	0.979	0.999	0.971	0.998	0.965	0.997	0.959	0.996	***0.954***	0.994	0.949	0.982
24	1.000	0.997	1.000	0.993	1.000	0.987	1.000	0.982	0.999	0.975	0.999	0.970	0.998	0.965	0.997	0.959	0.995	***0.955***	0.985
25	1.000	0.998	1.000	0.994	1.000	0.989	1.000	0.985	1.000	0.979	0.999	0.974	0.998	0.969	0.997	0.965	0.996	0.960	0.987

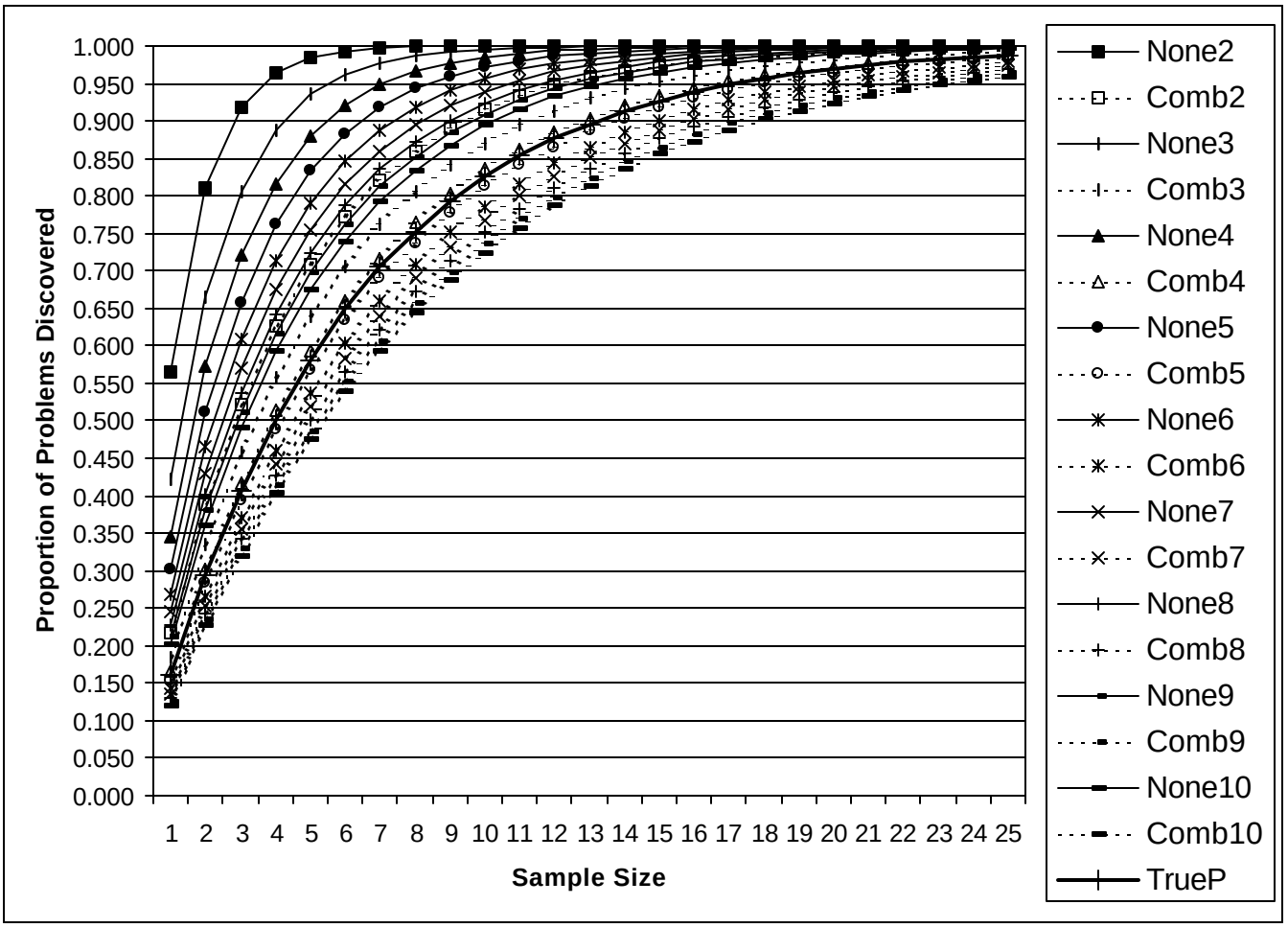

Figure 8. MACERR unadjusted and combination estimate problem discovery projections

Table 12. Unadjusted and combination estimate deviation from required sample size for MACERR

N	None90	Comb90	None95	Comb95
2	11	4	14	5
3	9	2	12	3
4	8	1	10	1
5	7	0	9	0
6	6	-1	8	-2
7	5	-2	7	-3
8	5	-3	6	-4
9	4	-4	5	-5
10	3	-4	4	-6

Note: A positive number indicates underestimation of the required sample size.

Table 13. Unadjusted and combination estimate deviation from problem discovery goal for MACERR

N	None90	Comb90	None95	Comb95
2	-0.493	-0.075	-0.448	-0.054
3	-0.318	-0.023	-0.301	-0.023
4	-0.251	-0.004	-0.198	-0.002
5	-0.195	0.013	-0.158	0.007
6	-0.148	0.027	-0.125	0.019
7	-0.108	0.039	-0.097	0.024
8	-0.108	0.048	-0.073	0.028
9	-0.075	0.057	-0.054	0.032
10	-0.047	0.057	-0.037	0.035

Note: A positive number indicates overachievement of the problem discovery goal

Table 14. Unadjusted and combination estimate problem discovery projections for VIRZI90

N	None 2	Comb 2	None 3	Comb 3	None 4	Comb 4	None 5	Comb 5	None 6	Comb 6	None 7	Comb 7	None 8	Comb 8	None 9	Comb 9	None 10	Comb 10	True *p*
1	0.662	0.361	0.545	0.338	0.485	0.328	0.449	0.321	0.425	0.318	0.409	0.318	0.396	0.316	0.388	0.318	0.381	0.319	0.359
2	0.886	0.592	0.793	0.562	0.735	0.548	0.696	0.539	0.669	0.535	0.651	0.535	0.635	0.532	0.625	0.535	0.617	0.536	0.589
3	***0.961***	0.739	**0.906**	0.710	0.863	0.697	0.833	0.687	0.810	0.683	0.794	0.683	0.780	0.680	0.771	0.683	0.763	0.684	0.737
4	0.987	0.833	***0.957***	0.808	**0.930**	0.796	**0.908**	0.787	0.891	0.784	0.878	0.784	0.867	0.781	0.860	0.784	0.853	0.785	0.831
5	0.996	0.893	0.980	0.873	***0.964***	0.863	0.949	0.856	**0.937**	0.852	**0.928**	0.852	**0.920**	0.850	**0.914**	0.852	**0.909**	0.854	0.892
6	0.999	**0.932**	0.991	**0.916**	0.981	**0.908**	***0.972***	**0.902**	***0.964***	0.899	***0.957***	0.899	***0.951***	0.898	0.947	0.899	0.944	**0.900**	**0.931**
7	0.999	***0.956***	0.996	0.944	0.990	0.938	0.985	0.933	0.979	**0.931**	0.975	**0.931**	0.971	**0.930**	***0.968***	**0.931**	***0.965***	0.932	***0.956***
8	1.000	0.972	0.998	***0.963***	0.995	***0.958***	0.992	***0.955***	0.988	***0.953***	0.985	***0.953***	0.982	***0.952***	0.980	***0.953***	0.978	***0.954***	0.971
9	1.000	0.982	0.999	0.976	0.997	0.972	0.995	0.969	0.993	0.968	0.991	0.968	0.989	0.967	0.988	0.968	0.987	0.968	0.982
10	1.000	0.989	1.000	0.984	0.999	0.981	0.997	0.979	0.996	0.978	0.995	0.978	0.994	0.978	0.993	0.978	0.992	0.979	0.988
11	1.000	0.993	1.000	0.989	0.999	0.987	0.999	0.986	0.998	0.985	0.997	0.985	0.996	0.985	0.995	0.985	0.995	0.985	0.992
12	1.000	0.995	1.000	0.993	1.000	0.992	0.999	0.990	0.999	0.990	0.998	0.990	0.998	0.990	0.997	0.990	0.997	0.990	0.995
13	1.000	0.997	1.000	0.995	1.000	0.994	1.000	0.993	0.999	0.993	0.999	0.993	0.999	0.993	0.998	0.993	0.998	0.993	0.997
14	1.000	0.998	1.000	0.997	1.000	0.996	1.000	0.996	1.000	0.995	0.999	0.995	0.999	0.995	0.999	0.995	0.999	0.995	0.998
15	1.000	0.999	1.000	0.998	1.000	0.997	1.000	0.997	1.000	0.997	1.000	0.997	0.999	0.997	0.999	0.997	0.999	0.997	0.999
16	1.000	0.999	1.000	0.999	1.000	0.998	1.000	0.998	1.000	0.998	1.000	0.998	1.000	0.998	1.000	0.998	1.000	0.998	0.999
17	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.998	1.000	0.999	1.000	0.999	0.999
18	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000
19	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000
20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000	1.000	1.000	1.000

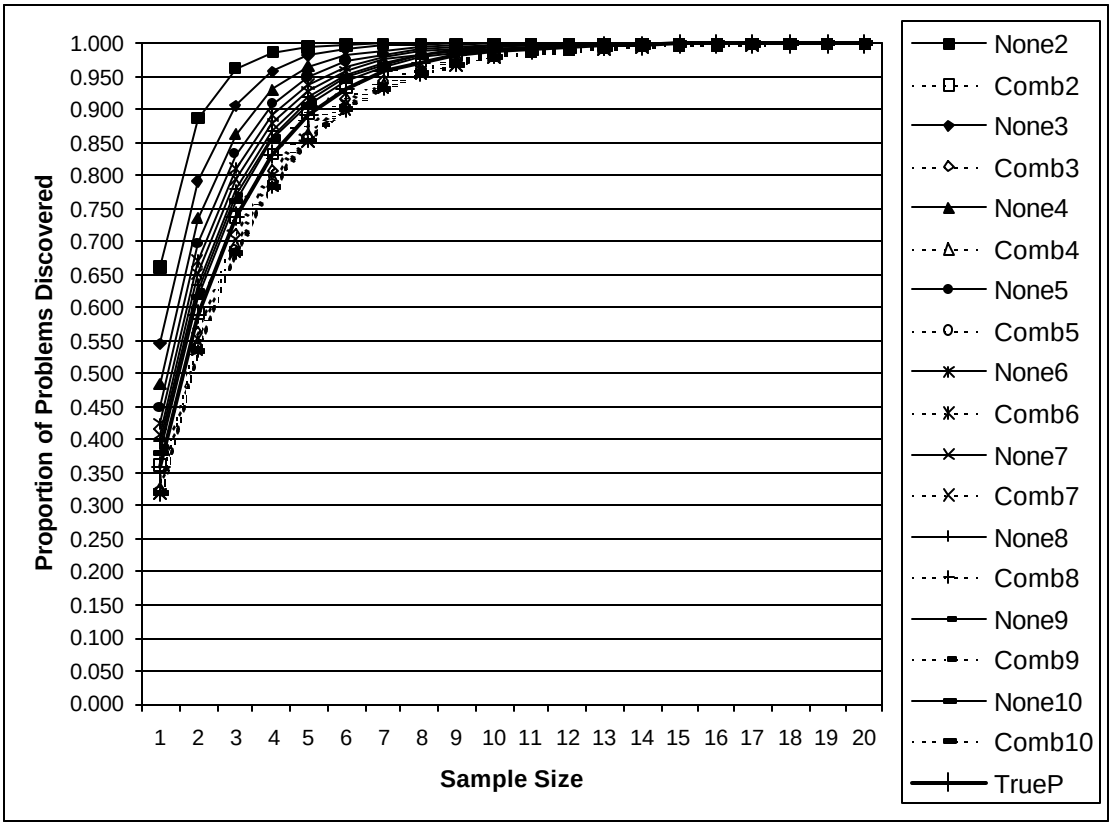

Figure 9. VIRZI90 unadjusted and combination estimate problem discovery projections

Table 15. Unadjusted and combination estimation deviation from required sample size for VIRZI90

N	None90	Comb90	None95	Comb95
2	3	0	4	0
3	3	0	3	-1
4	2	0	2	-1
5	2	0	1	-1
6	1	-1	1	-1
7	1	-1	1	-1
8	1	-1	1	-1
9	1	-1	0	-1
10	1	0	0	-1

Note: A positive number indicates underestimation of the required sample size.

Table 16. Unadjusted and combination estimation deviation from problem discovery goal for VIRZI90

N	None90	Comb90	None95	Comb95
2	-0.163	0.031	-0.213	0.006
3	-0.163	0.031	-0.119	0.021
4	-0.069	0.031	-0.058	0.021
5	-0.069	0.031	-0.019	0.021
6	-0.008	0.056	-0.019	0.021
7	-0.008	0.056	-0.019	0.021
8	-0.008	0.056	-0.019	0.021
9	-0.008	0.056	0.006	0.021
10	-0.008	0.031	0.021	0.021

Note: A positive number indicates overachievement of the problem discovery goal

Table 17. Unadjusted and combination estimate problem discovery projections for MANTEL

N	None 2	Comb 2	None 3	Comb 3	None 4	Comb 4	None 5	Comb 5	None 6	Comb 6	None 7	Comb 7	None 8	Comb 8	None 9	Comb 9	None 10	Comb 10	True *p*
1	0.725	0.462	0.626	0.444	0.571	0.429	0.539	0.421	0.510	0.407	0.492	0.401	0.476	0.393	0.465	0.390	0.455	0.386	0.375
2	**0.924**	0.711	0.860	0.691	0.816	0.674	0.787	0.665	0.760	0.648	0.742	0.641	0.725	0.632	0.714	0.628	0.703	0.623	0.609
3	***0.979***	0.844	**0.948**	0.828	**0.921**	0.814	**0.902**	0.806	0.882	0.791	0.869	0.785	0.856	0.776	0.847	0.773	0.838	0.769	0.756
4	0.994	**0.916**	***0.980***	**0.904**	***0.966***	0.894	***0.955***	0.888	**0.942**	0.876	**0.933**	0.871	**0.925**	0.864	**0.918**	0.862	**0.912**	0.858	0.847
5	0.998	***0.955***	0.993	0.947	0.985	**0.939**	0.979	**0.935**	***0.972***	**0.927**	***0.966***	**0.923**	***0.960***	**0.918**	***0.956***	**0.916**	***0.952***	**0.913**	**0.905**
6	1.000	0.976	0.997	***0.970***	0.994	***0.965***	0.990	***0.962***	0.986	***0.957***	0.983	***0.954***	0.979	***0.950***	0.977	0.948	0.974	0.946	0.940
7	1.000	0.987	0.999	0.984	0.997	0.980	0.996	0.978	0.993	0.974	0.991	0.972	0.989	0.970	0.987	***0.969***	0.986	***0.967***	***0.963***
8	1.000	0.993	1.000	0.991	0.999	0.989	0.998	0.987	0.997	0.985	0.996	0.983	0.994	0.982	0.993	0.981	0.992	0.980	0.977
9	1.000	0.996	1.000	0.995	1.000	0.994	0.999	0.993	0.998	0.991	0.998	0.990	0.997	0.989	0.996	0.988	0.996	0.988	0.985
10	1.000	0.998	1.000	0.997	1.000	0.996	1.000	0.996	0.999	0.995	0.999	0.994	0.998	0.993	0.998	0.993	0.998	0.992	0.991
11	1.000	0.999	1.000	0.998	1.000	0.998	1.000	0.998	1.000	0.997	0.999	0.996	0.999	0.996	0.999	0.996	0.999	0.995	0.994
12	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.998	1.000	0.998	1.000	0.997	0.999	0.997	0.999	0.997	0.996
13	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.998	1.000	0.998	1.000	0.998	0.998
14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	0.999
15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.999	0.999
16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
19	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

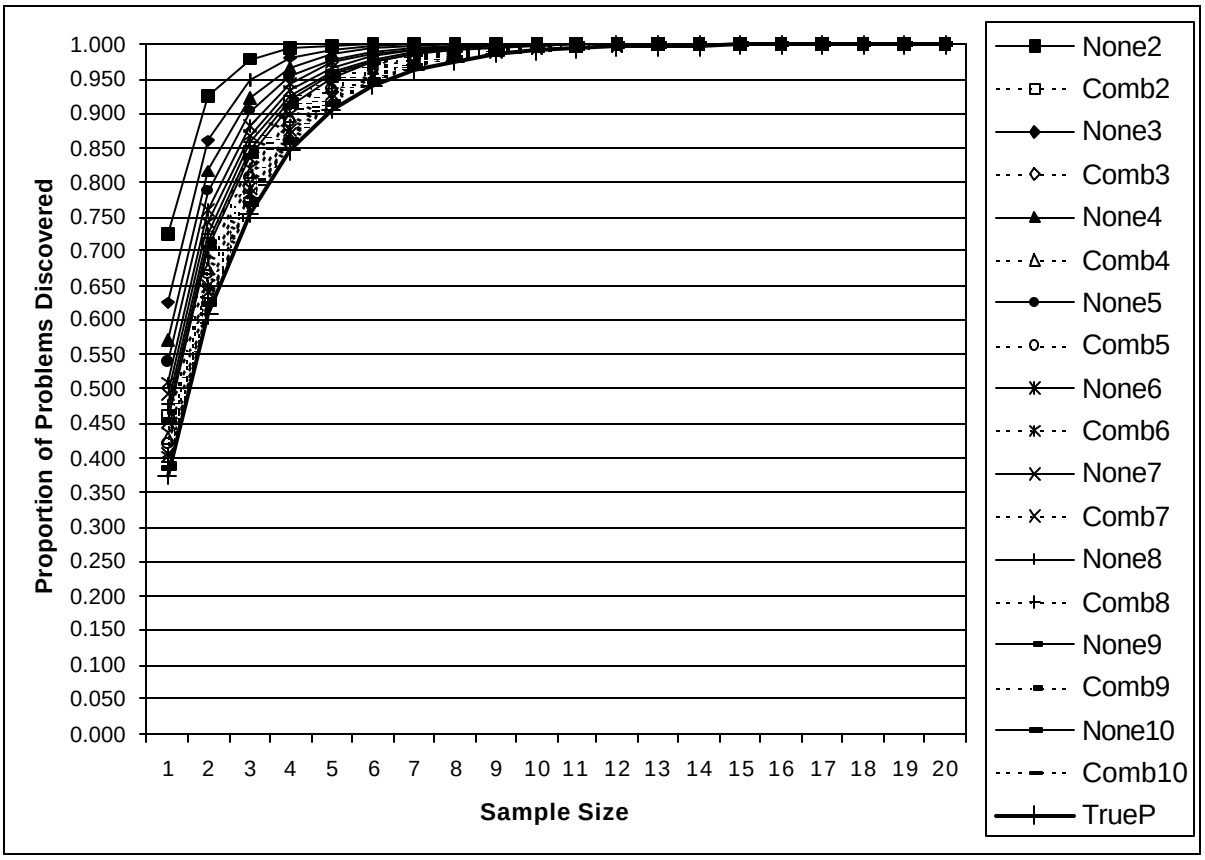

Figure 10. MANTEL unadjusted and combination estimate problem discovery projections

Table 18. Unadjusted and combination estimation deviation from required sample size for MANTEL

N	None90	Comb90	None95	Comb95
2	3	1	4	2
3	2	1	3	1
4	2	0	3	1
5	2	0	3	1
6	1	0	2	1
7	1	0	2	1
8	1	0	2	1
9	1	0	2	0
10	1	0	2	0

Note: A positive number indicates underestimation of the required sample size.

Table 19. Unadjusted and combination estimation deviation from problem discovery goal for MANTEL

N	None90	Comb90	None95	Comb95
2	-0.291	-0.053	-0.194	-0.045
3	-0.144	-0.053	-0.103	-0.010
4	-0.144	0.005	-0.103	-0.010
5	-0.144	0.005	-0.103	-0.010
6	-0.053	0.005	-0.045	-0.010
7	-0.053	0.005	-0.045	-0.010
8	-0.053	0.005	-0.045	-0.010
9	-0.053	0.005	-0.045	0.013
10	-0.053	0.005	-0.045	0.013

Note: A positive number indicates overachievement of the problem discovery goal

Table 20. Unadjusted and combination estimate problem discovery projections for SAVE

N	None 2	Comb 2	None 3	Comb 3	None 4	Comb 4	None 5	Comb 5	None 6	Comb 6	None 7	Comb 7	None 8	Comb 8	None 9	Comb 9	None 10	Comb 10	True *p*
1	0.629	0.311	0.505	0.288	0.442	0.277	0.406	0.273	0.380	0.267	0.362	0.265	0.349	0.263	0.337	0.260	0.329	0.259	0.256
2	0.862	0.525	0.755	0.493	0.689	0.477	0.647	0.471	0.616	0.463	0.593	0.460	0.576	0.457	0.560	0.452	0.550	0.451	0.446
3	**0.949**	0.673	0.879	0.639	0.826	0.622	0.790	0.616	0.762	0.606	0.740	0.603	0.724	0.600	0.709	0.595	0.698	0.593	0.588
4	***0.981***	0.775	**0.940**	0.743	**0.903**	0.727	0.876	0.721	0.852	0.711	0.834	0.708	0.820	0.705	0.807	0.700	0.797	0.699	0.694
5	0.993	0.845	***0.970***	0.817	0.946	0.802	**0.926**	0.797	**0.908**	0.788	0.894	0.785	0.883	0.783	0.872	0.778	0.864	0.777	0.772
6	0.997	0.893	0.985	0.870	***0.970***	0.857	***0.956***	0.852	0.943	0.845	**0.933**	0.842	**0.924**	0.840	**0.915**	0.836	**0.909**	0.834	0.830
7	0.999	**0.926**	0.993	**0.907**	0.983	0.897	0.974	0.893	***0.965***	0.886	***0.957***	0.884	***0.950***	0.882	0.944	0.878	0.939	0.877	0.874
8	1.000	0.949	0.996	0.934	0.991	**0.925**	0.985	**0.922**	0.978	**0.917**	0.973	**0.915**	0.968	**0.913**	***0.963***	**0.910**	***0.959***	**0.909**	**0.906**
9	1.000	***0.965***	0.998	***0.953***	0.995	0.946	0.991	0.943	0.986	0.939	0.982	0.937	0.979	0.936	0.975	0.933	0.972	0.933	0.930
10	1.000	0.976	0.999	0.967	0.997	***0.961***	0.995	***0.959***	0.992	***0.955***	0.989	***0.954***	0.986	***0.953***	0.984	***0.951***	0.981	***0.950***	0.948
11	1.000	0.983	1.000	0.976	0.998	0.972	0.997	0.970	0.995	0.967	0.993	0.966	0.991	0.965	0.989	0.964	0.988	0.963	***0.961***
12	1.000	0.989	1.000	0.983	0.999	0.980	0.998	0.978	0.997	0.976	0.995	0.975	0.994	0.974	0.993	0.973	0.992	0.973	0.971
13	1.000	0.992	1.000	0.988	0.999	0.985	0.999	0.984	0.998	0.982	0.997	0.982	0.996	0.981	0.995	0.980	0.994	0.980	0.979
14	1.000	0.995	1.000	0.991	1.000	0.989	0.999	0.988	0.999	0.987	0.998	0.987	0.998	0.986	0.997	0.985	0.996	0.985	0.984
15	1.000	0.996	1.000	0.994	1.000	0.992	1.000	0.992	0.999	0.991	0.999	0.990	0.998	0.990	0.998	0.989	0.997	0.989	0.988
16	1.000	0.997	1.000	0.996	1.000	0.994	1.000	0.994	1.000	0.993	0.999	0.993	0.999	0.992	0.999	0.992	0.998	0.992	0.991
17	1.000	0.998	1.000	0.997	1.000	0.996	1.000	0.996	1.000	0.995	1.000	0.995	0.999	0.994	0.999	0.994	0.999	0.994	0.993
18	1.000	0.999	1.000	0.998	1.000	0.997	1.000	0.997	1.000	0.996	1.000	0.996	1.000	0.996	0.999	0.996	0.999	0.995	0.995
19	1.000	0.999	1.000	0.998	1.000	0.998	1.000	0.998	1.000	0.997	1.000	0.997	1.000	0.997	1.000	0.997	0.999	0.997	0.996
20	1.000	0.999	1.000	0.999	1.000	0.998	1.000	0.998	1.000	0.998	1.000	0.998	1.000	0.998	1.000	0.998	1.000	0.998	0.997

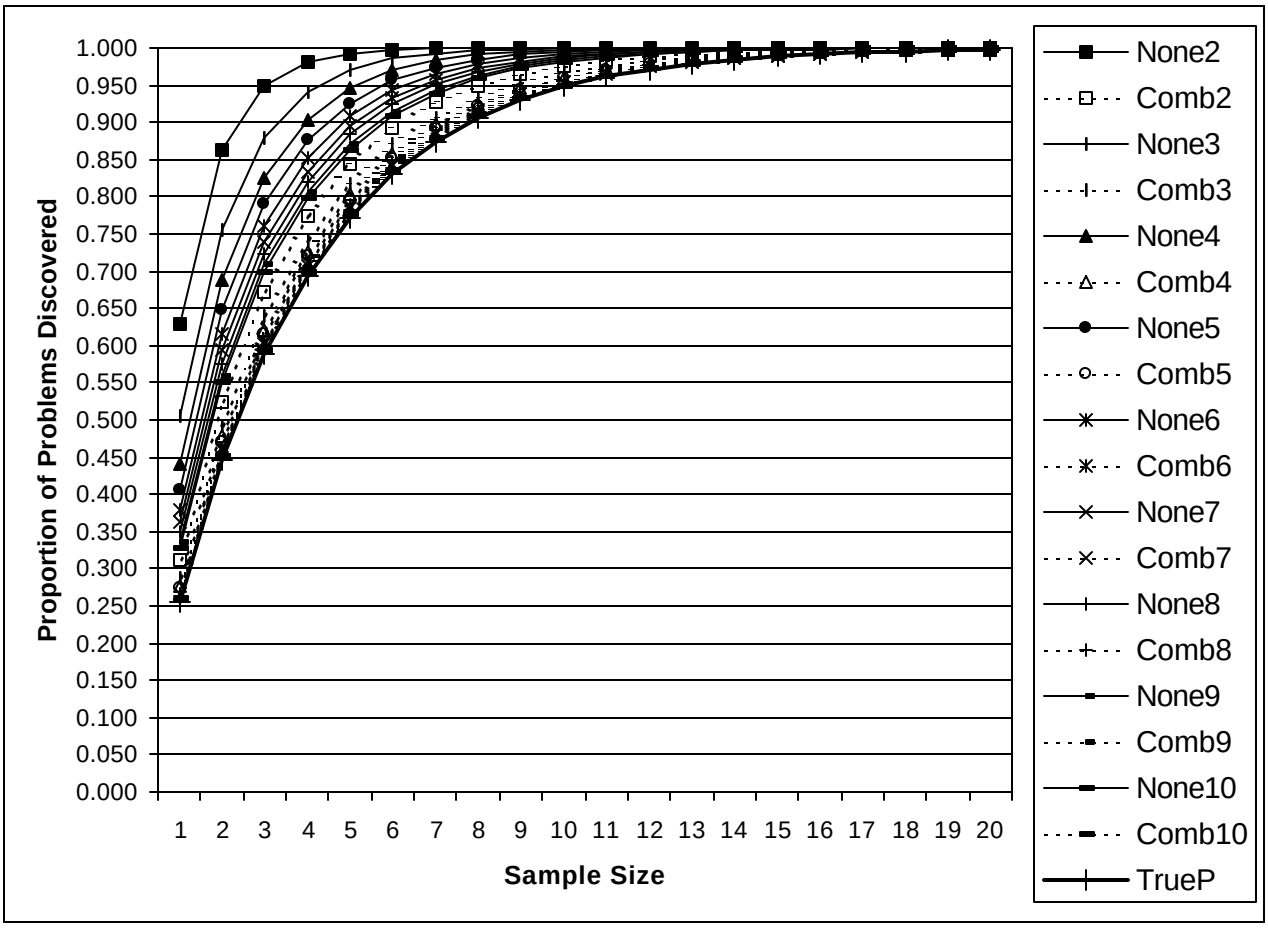

Figure 11. SAVE unadjusted and combination estimate problem discovery projections

Table 21. Unadjusted and combination estimation deviation from required sample size for SAVE

N	None90	Comb90	None95	Comb95
2	5	1	7	2
3	4	1	6	2
4	4	0	5	1
5	3	0	5	1
6	3	0	4	1
7	2	0	4	1
8	2	0	4	1
9	2	0	3	1
10	2	0	3	1

Note: A positive number indicates underestimation of the required sample size.

Table 22. Unadjusted and combination estimation deviation from problem discovery goal for SAVE

N	None90	Comb90	None95	Comb95
2	-0.312	-0.026	-0.256	-0.020
3	-0.206	-0.026	-0.178	-0.020
4	-0.206	0.006	-0.120	-0.002
5	-0.128	0.006	-0.120	-0.002
6	-0.128	0.006	-0.076	-0.002
7	-0.070	0.006	-0.076	-0.002
8	-0.070	0.006	-0.076	-0.002
9	-0.070	0.006	-0.044	-0.002
10	-0.070	0.006	-0.044	-0.002

Note: A positive number indicates overachievement of the problem discovery goal

Normalization and Good-Turing Estimation

Tables 23, 26, 29, and 32, and Figures 12-15 illustrate the sample size projections for normalized *p* (Norm) and *p* adjusted with the Good-Turing estimate (GT). Tables 24, 27, 30, and 33 and Tables 25, 28, 31, and 34 show the deviations from required sample size and proportion of discovered problems, respectively, for each problem discovery database with these adjustment procedures.

Table 23. Normalized and Good-Turing problem discovery projections for MACERR

N	Norm 2	GT 2	Norm 3	GT 3	Norm 4	GT 4	Norm 5	GT 5	Norm 6	GT 6	Norm 7	GT 7	Norm 8	GT 8	Norm 9	GT 9	Norm 10	GT 10	True *p*
1	0.132	0.304	0.131	0.238	0.128	0.202	0.127	0.181	0.122	0.164	0.120	0.152	0.117	0.143	0.114	0.135	0.112	0.129	0.160
2	0.247	0.516	0.245	0.419	0.240	0.363	0.238	0.329	0.229	0.301	0.226	0.281	0.220	0.266	0.215	0.252	0.211	0.241	0.294
3	0.346	0.663	0.344	0.558	0.337	0.492	0.335	0.451	0.323	0.416	0.319	0.390	0.312	0.371	0.304	0.353	0.300	0.339	0.407
4	0.432	0.765	0.430	0.663	0.422	0.594	0.419	0.550	0.406	0.512	0.400	0.483	0.392	0.461	0.384	0.440	0.378	0.424	0.502
5	0.507	0.837	0.504	0.743	0.496	0.676	0.493	0.632	0.478	0.592	0.472	0.561	0.463	0.538	0.454	0.516	0.448	0.499	0.582
6	0.572	0.886	0.569	0.804	0.560	0.742	0.557	0.698	0.542	0.659	0.536	0.628	0.526	0.604	0.516	0.581	0.510	0.563	0.649
7	0.629	**0.921**	0.626	0.851	0.617	0.794	0.614	0.753	0.598	0.715	0.591	0.685	0.581	0.660	0.571	0.638	0.565	0.620	0.705
8	0.678	0.945	0.675	0.886	0.666	0.836	0.663	0.798	0.647	0.761	0.640	0.733	0.630	0.709	0.620	0.687	0.613	0.669	0.752
9	0.720	***0.962***	0.717	**0.913**	0.708	0.869	0.705	0.834	0.690	0.801	0.684	0.773	0.674	0.751	0.664	0.729	0.657	0.711	0.792
10	0.757	0.973	0.754	0.934	0.746	0.895	0.743	0.864	0.728	0.833	0.721	0.808	0.712	0.786	0.702	0.765	0.695	0.749	0.825
11	0.789	0.981	0.787	***0.950***	0.778	**0.916**	0.776	0.889	0.761	0.861	0.755	0.837	0.746	0.817	0.736	0.797	0.729	0.781	0.853
12	0.817	0.987	0.815	0.962	0.807	0.933	0.804	**0.909**	0.790	0.883	0.784	0.862	0.775	0.843	0.766	0.825	0.760	0.809	0.877
13	0.841	0.991	0.839	0.971	0.831	0.947	0.829	0.925	0.816	**0.903**	0.810	0.883	0.802	0.865	0.793	0.848	0.787	0.834	0.896
14	0.862	0.994	0.860	0.978	0.853	***0.958***	0.851	0.939	0.838	0.919	0.833	**0.901**	0.825	0.885	0.816	0.869	0.810	0.855	**0.913**
15	0.880	0.996	0.878	0.983	0.872	0.966	0.870	***0.950***	0.858	0.932	0.853	0.916	0.845	**0.901**	0.837	0.886	0.832	0.874	0.927
16	0.896	0.997	0.894	0.987	0.888	0.973	0.886	0.959	0.875	0.943	0.871	0.928	0.863	0.915	0.856	**0.902**	0.851	0.890	0.939
17	**0.910**	0.998	**0.908**	0.990	**0.903**	0.978	**0.901**	0.966	0.891	***0.952***	0.886	0.939	0.879	0.927	0.872	0.915	0.867	**0.904**	0.948
18	0.922	0.999	0.920	0.992	0.915	0.983	0.913	0.973	**0.904**	0.960	**0.900**	0.949	0.894	0.938	0.887	0.926	0.882	0.917	***0.957***
19	0.932	0.999	0.931	0.994	0.926	0.986	0.924	0.977	0.916	0.967	0.912	***0.956***	**0.906**	0.947	**0.900**	0.936	0.895	0.927	0.964
20	0.941	0.999	0.940	0.996	0.935	0.989	0.934	0.982	0.926	0.972	0.922	0.963	0.917	***0.954***	0.911	0.945	**0.907**	0.937	0.969
21	0.949	1.000	0.948	0.997	0.944	0.991	0.942	0.985	0.935	0.977	0.932	0.969	0.927	0.961	0.921	***0.952***	0.917	0.945	0.974
22	***0.956***	1.000	***0.954***	0.997	***0.951***	0.993	***0.950***	0.988	0.943	0.981	0.940	0.973	0.935	0.966	0.930	0.959	0.927	***0.952***	0.978
23	0.961	1.000	0.960	0.998	0.957	0.994	0.956	0.990	***0.950***	0.984	0.947	0.977	0.943	0.971	0.938	0.964	0.935	0.958	0.982
24	0.967	1.000	0.966	0.999	0.963	0.996	0.962	0.992	0.956	0.986	***0.953***	0.981	***0.950***	0.975	0.945	0.969	0.942	0.964	0.985
25	0.971	1.000	0.970	0.999	0.967	0.996	0.966	0.993	0.961	0.989	0.959	0.984	0.955	0.979	***0.951***	0.973	0.949	0.968	0.987
26	0.975	1.000	0.974	0.999	0.972	0.997	0.971	0.994	0.966	0.991	0.964	0.986	0.961	0.982	0.957	0.977	***0.954***	0.972	0.989

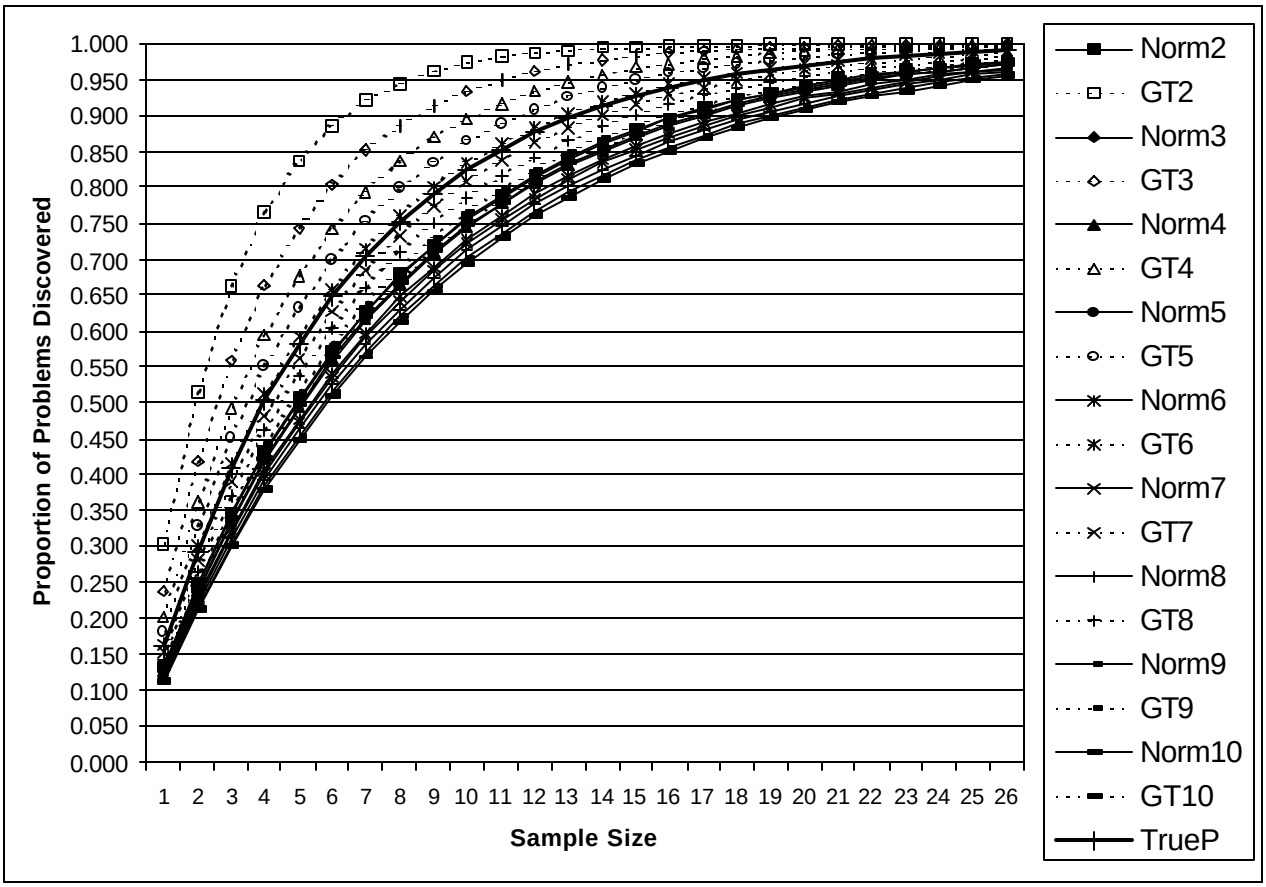

Figure 12. MACERR normalized and Good-Turing problem discovery projections

Table 24. Normalized and Good-Turing estimate deviation from required sample size for MACERR

N	Norm90	GT90	Norm95	GT95
2	-3	7	-4	9
3	-3	5	-4	7
4	-3	3	-4	4
5	-3	2	-4	3
6	-4	1	-5	1
7	-4	0	-6	-1
8	-5	-1	-6	-2
9	-5	-2	-7	-3
10	-6	-3	-8	-4

Note: A positive number indicates underestimation of the required sample size.

Table 25. Normalized and Good-Turing estimate deviation from problem discovery goal for MACERR

N	Norm90	GT90	Norm95	GT95
2	0.048	-0.195	0.028	-0.158
3	0.048	-0.108	0.028	-0.097
4	0.048	-0.047	0.028	-0.037
5	0.048	-0.023	0.028	-0.023
6	0.057	-0.004	0.032	-0.002
7	0.057	0.013	0.035	0.014
8	0.064	0.027	0.035	0.019
9	0.064	0.039	0.037	0.024
10	0.069	0.048	0.039	0.028

Note: A positive number indicates overachievement of the problem discovery goal

Table 26. Normalized and Good-Turing problem discovery projections for VIRZI90

N	Norm 2	GT 2	Norm 3	GT 3	Norm 4	GT 4	Norm 5	GT 5	Norm 6	GT 6	Norm 7	GT 7	Norm 8	GT 8	Norm 9	GT 9	Norm 10	GT 10	True *p*
1	0.324	0.398	0.317	0.359	0.314	0.341	0.311	0.332	0.310	0.326	0.311	0.325	0.310	0.323	0.312	0.325	0.313	0.326	0.359
2	0.543	0.638	0.534	0.589	0.529	0.566	0.525	0.554	0.524	0.546	0.525	0.544	0.524	0.542	0.527	0.544	0.528	0.546	0.589
3	0.691	0.782	0.681	0.737	0.677	0.714	0.673	0.702	0.671	0.694	0.673	0.692	0.671	0.690	0.674	0.692	0.676	0.694	0.737
4	0.791	0.869	0.782	0.831	0.779	0.811	0.775	0.801	0.773	0.794	0.775	0.792	0.773	0.790	0.776	0.792	0.777	0.794	0.831
5	0.859	**0.921**	0.851	0.892	0.848	0.876	0.845	0.867	0.844	0.861	0.845	0.860	0.844	0.858	0.846	0.860	0.847	0.861	0.892
6	**0.905**	***0.952***	0.898	**0.931**	0.896	**0.918**	0.893	**0.911**	0.892	**0.906**	0.893	**0.905**	0.892	**0.904**	0.894	**0.905**	0.895	**0.906**	**0.931**
7	0.935	0.971	**0.931**	***0.956***	**0.929**	0.946	**0.926**	0.941	**0.926**	0.937	**0.926**	0.936	**0.926**	0.935	**0.927**	0.936	**0.928**	0.937	***0.956***
8	***0.956***	0.983	***0.953***	0.971	***0.951***	***0.964***	0.949	***0.960***	0.949	***0.957***	0.949	***0.957***	0.949	***0.956***	***0.950***	***0.957***	***0.950***	***0.957***	0.971
9	0.971	0.990	0.968	0.982	0.966	0.977	***0.965***	0.974	***0.965***	0.971	***0.965***	0.971	***0.965***	0.970	0.965	0.971	0.966	0.971	0.982
10	0.980	0.994	0.978	0.988	0.977	0.985	0.976	0.982	0.976	0.981	0.976	0.980	0.976	0.980	0.976	0.980	0.977	0.981	0.988
11	0.987	0.996	0.985	0.992	0.984	0.990	0.983	0.988	0.983	0.987	0.983	0.987	0.983	0.986	0.984	0.987	0.984	0.987	0.992
12	0.991	0.998	0.990	0.995	0.989	0.993	0.989	0.992	0.988	0.991	0.989	0.991	0.988	0.991	0.989	0.991	0.989	0.991	0.995
13	0.994	0.999	0.993	0.997	0.993	0.996	0.992	0.995	0.992	0.994	0.992	0.994	0.992	0.994	0.992	0.994	0.992	0.994	0.997
14	0.996	0.999	0.995	0.998	0.995	0.997	0.995	0.996	0.994	0.996	0.995	0.996	0.994	0.996	0.995	0.996	0.995	0.996	0.998
15	0.997	1.000	0.997	0.999	0.996	0.998	0.996	0.998	0.996	0.997	0.996	0.997	0.996	0.997	0.996	0.997	0.996	0.997	0.999
16	0.998	1.000	0.998	0.999	0.998	0.999	0.997	0.998	0.997	0.998	0.997	0.998	0.997	0.998	0.997	0.998	0.998	0.998	0.999
17	0.999	1.000	0.998	0.999	0.998	0.999	0.998	0.999	0.998	0.999	0.998	0.999	0.998	0.999	0.998	0.999	0.998	0.999	0.999
18	0.999	1.000	0.999	1.000	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	1.000
19	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	1.000
20	1.000	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000	1.000

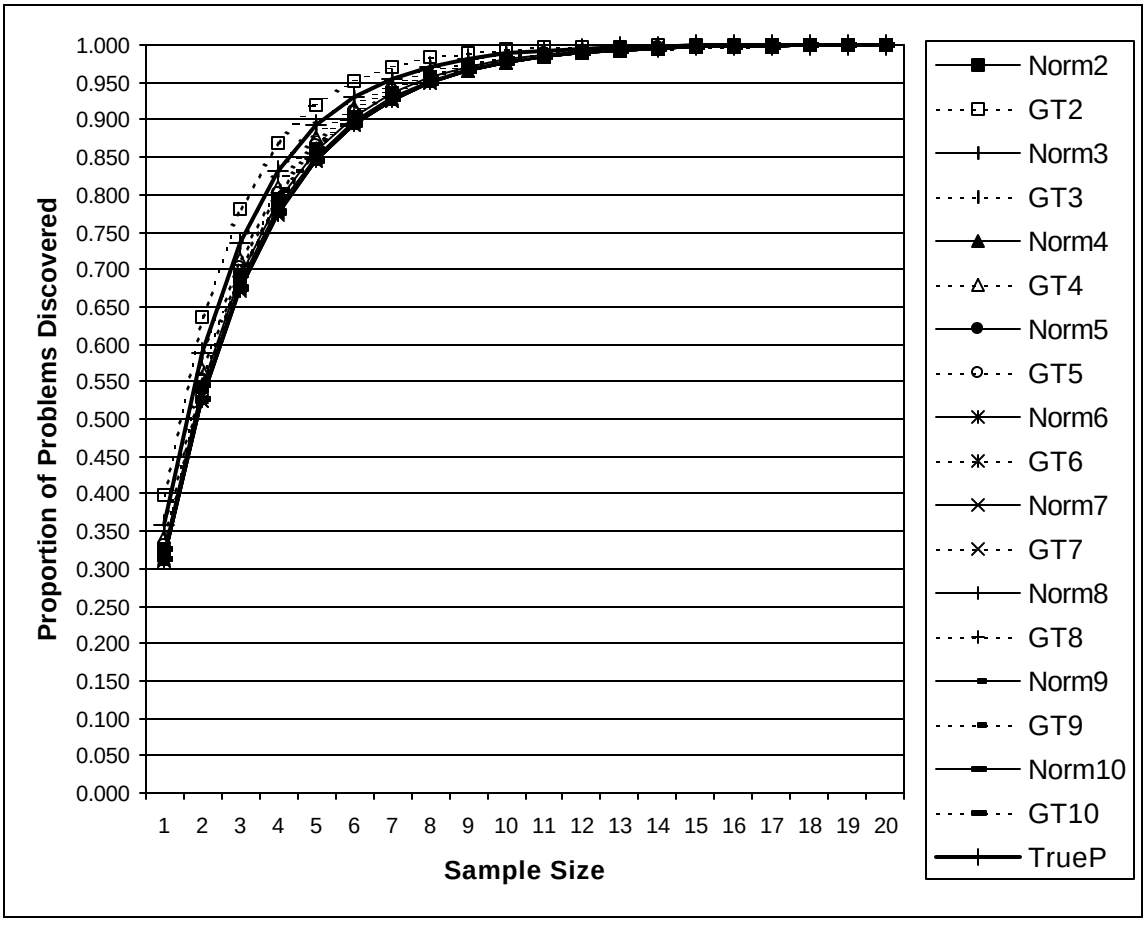

Figure 13. VIRZI90 normalized and Good-Turing problem discovery projections

Table 27. Normalized and Good-Turing estimation deviation from required sample size for VIRZI90

N	Norm90	GT90	Norm95	GT95
2	0	1	-1	1
3	-1	0	-1	0
4	-1	0	-1	-1
5	-1	0	-2	-1
6	-1	0	-2	-1
7	-1	0	-2	-1
8	-1	0	-2	-1
9	-1	0	-1	-1
10	-1	0	-1	-1

Note: A positive number indicates underestimation of the required sample size.

Table 28. Normalized and Good-Turing estimation deviation from problem discovery goal for VIRZI90

N	Norm90	GT90	Norm95	GT95
2	0.031	-0.008	0.021	-0.019
3	0.056	0.031	0.021	0.006
4	0.056	0.031	0.021	0.021
5	0.056	0.031	0.032	0.021
6	0.056	0.031	0.032	0.021
7	0.056	0.031	0.032	0.021
8	0.056	0.031	0.032	0.021
9	0.056	0.031	-0.021	0.021
10	0.056	0.031	-0.021	0.021

Note: A positive number indicates overachievement of the problem discovery goal

Table 29. Normalized and Good-Turing problem discovery projections for MANTEL

N	Norm 2	GT 2	Norm 3	GT 3	Norm 4	GT 4	Norm 5	GT 5	Norm 6	GT 6	Norm 7	GT 7	Norm 8	GT 8	Norm 9	GT 9	Norm 10	GT 10	True *p*
1	0.449	0.474	0.439	0.450	0.428	0.429	0.423	0.418	0.412	0.401	0.407	0.395	0.402	0.385	0.398	0.382	0.395	0.378	0.375
2	0.696	0.723	0.685	0.698	0.673	0.674	0.667	0.661	0.654	0.641	0.648	0.634	0.642	0.622	0.638	0.618	0.634	0.613	0.609
3	0.833	0.854	0.823	0.834	0.813	0.814	0.808	0.803	0.797	0.785	0.791	0.779	0.786	0.767	0.782	0.764	0.779	0.759	0.756
4	**0.908**	**0.923**	**0.901**	**0.908**	0.893	0.894	0.889	0.885	0.880	0.871	0.876	0.866	0.872	0.857	0.869	0.854	0.866	0.850	0.847
5	0.949	***0.960***	0.944	***0.950***	**0.939**	**0.939**	**0.936**	**0.933**	**0.930**	**0.923**	**0.927**	**0.919**	**0.924**	**0.912**	**0.921**	**0.910**	**0.919**	**0.907**	**0.905**
6	***0.972***	0.979	***0.969***	0.972	***0.965***	***0.965***	***0.963***	***0.961***	***0.959***	***0.954***	***0.957***	***0.951***	***0.954***	0.946	***0.952***	0.944	***0.951***	0.942	0.940
7	0.985	0.989	0.983	0.985	0.980	0.980	0.979	0.977	0.976	0.972	0.974	0.970	0.973	***0.967***	0.971	***0.966***	0.970	***0.964***	***0.963***
8	0.992	0.994	0.990	0.992	0.989	0.989	0.988	0.987	0.986	0.983	0.985	0.982	0.984	0.980	0.983	0.979	0.982	0.978	0.977
9	0.995	0.997	0.994	0.995	0.993	0.994	0.993	0.992	0.992	0.990	0.991	0.989	0.990	0.987	0.990	0.987	0.989	0.986	0.985
10	0.997	0.998	0.997	0.997	0.996	0.996	0.996	0.996	0.995	0.994	0.995	0.993	0.994	0.992	0.994	0.992	0.993	0.991	0.991
11	0.999	0.999	0.998	0.999	0.998	0.998	0.998	0.997	0.997	0.996	0.997	0.996	0.997	0.995	0.996	0.995	0.996	0.995	0.994
12	0.999	1.000	0.999	0.999	0.999	0.999	0.999	0.998	0.998	0.998	0.998	0.998	0.998	0.997	0.998	0.997	0.998	0.997	0.996
13	1.000	1.000	0.999	1.000	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.998	0.999	0.998	0.999	0.998	0.998
14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999
15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.999	0.999	0.999	0.999
16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	0.999
17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
19	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

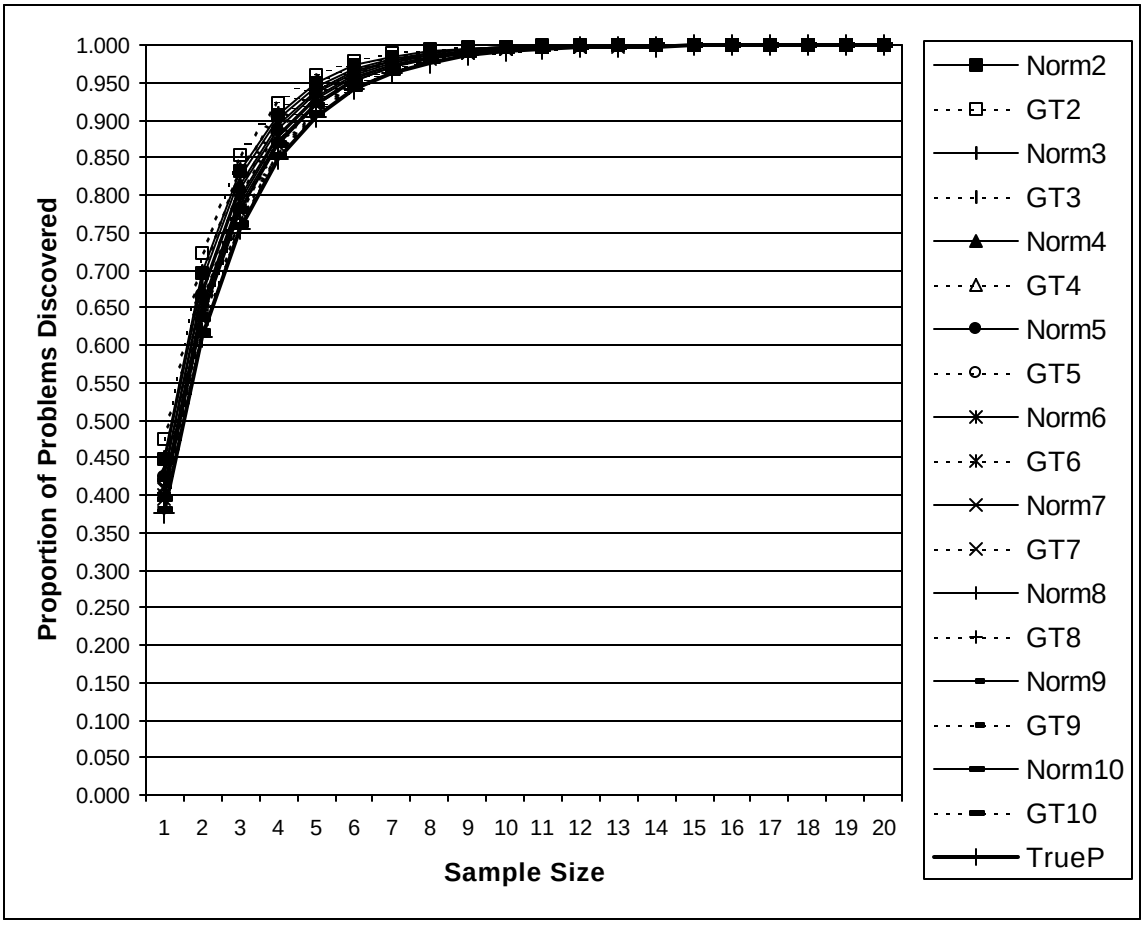

Figure 14. MANTEL Normalized and Good-Turing problem discovery projections

Table 30. Normalized and Good-Turing estimation deviation from required sample size for MANTEL

N	Norm90	GT90	Norm95	GT95
2	1	1	1	2
3	1	1	1	2
4	0	0	1	1
5	0	0	1	1
6	0	0	1	1
7	0	0	1	1
8	0	0	1	0
9	0	0	1	0
10	0	0	1	0

Note: A positive number indicates underestimation of the required sample size.

Table 31. Normalized and Good-Turing estimation deviation from problem discovery goal for MANTEL

N	Norm90	GT90	Norm95	GT95
2	-0.053	-0.053	-0.010	-0.045
3	-0.053	-0.053	-0.010	-0.045
4	0.005	0.005	-0.010	-0.010
5	0.005	0.005	-0.010	-0.010
6	0.005	0.005	-0.010	-0.010
7	0.005	0.005	-0.010	-0.010
8	0.005	0.005	-0.010	0.013
9	0.005	0.005	-0.010	0.013
10	0.005	0.005	-0.010	0.013

Note: A positive number indicates overachievement of the problem discovery goal

Table 32. Normalized and Good-Turing problem discovery projections for SAVE

N	Norm 2	GT 2	Norm 3	GT 3	Norm 4	GT 4	Norm 5	GT 5	Norm 6	GT 6	Norm 7	GT 7	Norm 8	GT 8	Norm 9	GT 9	Norm 10	GT 10	True *p*
1	0.258	0.364	0.257	0.320	0.256	0.298	0.258	0.288	0.256	0.279	0.256	0.275	0.255	0.270	0.254	0.266	0.254	0.265	0.256
2	0.449	0.596	0.448	0.538	0.446	0.507	0.449	0.493	0.446	0.480	0.446	0.474	0.445	0.467	0.443	0.461	0.443	0.460	0.446
3	0.591	0.743	0.590	0.686	0.588	0.654	0.591	0.639	0.588	0.625	0.588	0.619	0.587	0.611	0.585	0.605	0.585	0.603	0.588
4	0.697	0.836	0.695	0.786	0.694	0.757	0.697	0.743	0.694	0.730	0.694	0.724	0.692	0.716	0.690	0.710	0.690	0.708	0.694
5	0.775	0.896	0.774	0.855	0.772	0.830	0.775	0.817	0.772	0.805	0.772	0.800	0.771	0.793	0.769	0.787	0.769	0.785	0.772
6	0.833	**0.934**	0.832	**0.901**	0.830	0.880	0.833	0.870	0.830	0.860	0.830	0.855	0.829	0.849	0.828	0.844	0.828	0.842	0.830
7	0.876	***0.958***	0.875	0.933	0.874	**0.916**	0.876	**0.907**	0.874	0.899	0.874	0.895	0.873	0.890	0.871	0.885	0.871	0.884	0.874
8	**0.908**	0.973	**0.907**	***0.954***	**0.906**	0.941	**0.908**	0.934	**0.906**	**0.927**	**0.906**	**0.924**	**0.905**	**0.919**	**0.904**	**0.916**	**0.904**	**0.915**	**0.906**
9	0.932	0.983	0.931	0.969	0.930	***0.959***	0.932	***0.953***	0.930	0.947	0.930	0.945	0.929	0.941	0.928	0.938	0.928	0.937	0.930
10	0.949	0.989	0.949	0.979	0.948	0.971	0.949	0.967	0.948	***0.962***	0.948	***0.960***	0.947	***0.957***	0.947	***0.955***	0.947	***0.954***	0.948
11	***0.962***	0.993	***0.962***	0.986	***0.961***	0.980	***0.962***	0.976	***0.961***	0.973	***0.961***	0.971	***0.961***	0.969	***0.960***	0.967	***0.960***	0.966	***0.961***
12	0.972	0.996	0.972	0.990	0.971	0.986	0.972	0.983	0.971	0.980	0.971	0.979	0.971	0.977	0.970	0.976	0.970	0.975	0.971
13	0.979	0.997	0.979	0.993	0.979	0.990	0.979	0.988	0.979	0.986	0.979	0.985	0.978	0.983	0.978	0.982	0.978	0.982	0.979
14	0.985	0.998	0.984	0.995	0.984	0.993	0.985	0.991	0.984	0.990	0.984	0.989	0.984	0.988	0.983	0.987	0.983	0.987	0.984
15	0.989	0.999	0.988	0.997	0.988	0.995	0.989	0.994	0.988	0.993	0.988	0.992	0.988	0.991	0.988	0.990	0.988	0.990	0.988
16	0.992	0.999	0.991	0.998	0.991	0.997	0.992	0.996	0.991	0.995	0.991	0.994	0.991	0.993	0.991	0.993	0.991	0.993	0.991
17	0.994	1.000	0.994	0.999	0.993	0.998	0.994	0.997	0.993	0.996	0.993	0.996	0.993	0.995	0.993	0.995	0.993	0.995	0.993
18	0.995	1.000	0.995	0.999	0.995	0.998	0.995	0.998	0.995	0.997	0.995	0.997	0.995	0.997	0.995	0.996	0.995	0.996	0.995
19	0.997	1.000	0.996	0.999	0.996	0.999	0.997	0.998	0.996	0.998	0.996	0.998	0.996	0.997	0.996	0.997	0.996	0.997	0.996
20	0.997	1.000	0.997	1.000	0.997	0.999	0.997	0.999	0.997	0.999	0.997	0.998	0.997	0.998	0.997	0.998	0.997	0.998	0.997

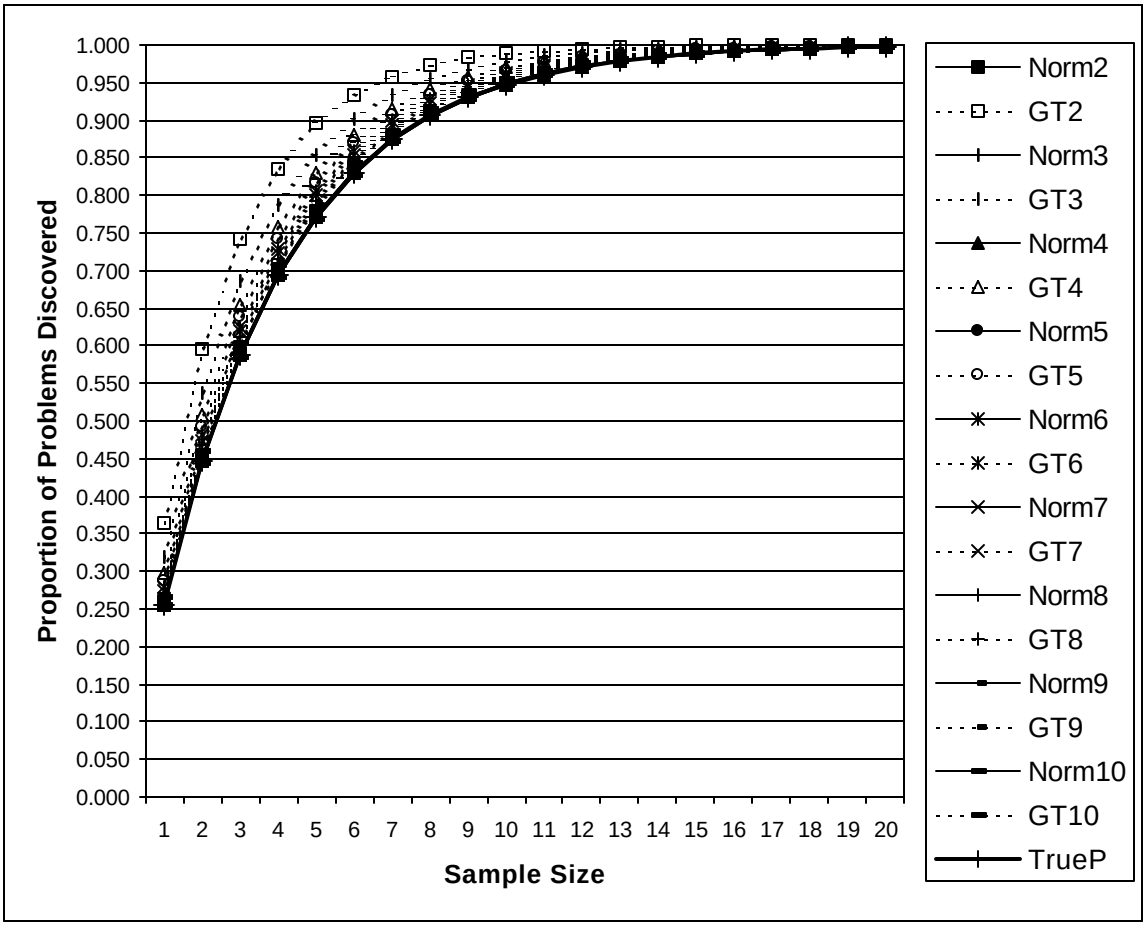

Figure 15. SAVE normalized and Good-Turing problem discovery projections

Table 33. Normalized and Good-Turing estimation deviation from required sample size for SAVE

N	Norm90	GT90	Norm95	GT95
2	0	2	0	4
3	0	2	0	3
4	0	1	0	2
5	0	1	0	2
6	0	0	0	1
7	0	0	0	1
8	0	0	0	1
9	0	0	0	1
10	0	0	0	1

Note: A positive number indicates underestimation of the required sample size.

Table 34. Normalized and Good-Turing estimation deviation from problem discovery goal for SAVE

N	Norm90	GT90	Norm95	GT95
2	0.006	-0.070	0.011	-0.076
3	0.006	-0.070	0.011	-0.044
4	0.006	-0.026	0.011	-0.020
5	0.006	-0.026	0.011	-0.020
6	0.006	0.006	0.011	-0.002
7	0.006	0.006	0.011	-0.002
8	0.006	0.006	0.011	-0.002
9	0.006	0.006	0.011	-0.002
10	0.006	0.006	0.011	-0.002

Note: A positive number indicates overachievement of the problem discovery goal

Regression Equations 2 and 5

Tables 35, 38, 41, and 44, and Figures 16-19 illustrate the sample size projections for *p* adjusted with Regression Equation 2 (Reg2) and *p* adjusted with Regression Equation 5 (Reg5). Tables 36, 39, 42, and 45 and Tables 37, 40, 43, and 46 show the deviations from required sample size and proportion of discovered problems, respectively, for each problem discovery database with these adjustment procedures.

Table 35. Regression equation 2 and 5 problem discovery projections for MACERR

N	Reg 2 2	Reg 5 2	Reg 2 3	Reg 5 3	Reg 2 4	Reg 5 4	Reg 2 5	Reg 5 5	Reg 2 6	Reg 5 6	Reg 2 7	Reg 5 7	Reg 2 8	Reg 5 8	Reg 2 9	Reg 5 9	Reg 2 10	Reg 5 10	True *p*
1	0.293	0.269	0.280	0.268	0.264	0.265	0.250	0.264	0.233	0.261	0.218	0.258	0.203	0.256	0.188	0.254	0.173	0.252	0.160
2	0.500	0.466	0.482	0.464	0.458	0.460	0.438	0.458	0.412	0.454	0.388	0.449	0.365	0.446	0.341	0.443	0.316	0.440	0.294
3	0.647	0.609	0.627	0.608	0.601	0.603	0.578	0.601	0.549	0.596	0.522	0.591	0.494	0.588	0.465	0.585	0.434	0.581	0.407
4	0.750	0.714	0.731	0.713	0.707	0.708	0.684	0.707	0.654	0.702	0.626	0.697	0.597	0.694	0.565	0.690	0.532	0.687	0.502
5	0.823	0.791	0.807	0.790	0.784	0.785	0.763	0.784	0.735	0.780	0.708	0.775	0.678	0.772	0.647	0.769	0.613	0.766	0.582
6	0.875	0.847	0.861	0.846	0.841	0.842	0.822	0.841	0.796	0.837	0.771	0.833	0.744	0.830	0.713	0.828	0.680	0.825	0.649
7	**0.912**	0.888	**0.900**	0.887	0.883	0.884	0.867	0.883	0.844	0.880	0.821	0.876	0.796	0.874	0.767	0.871	0.735	0.869	0.705
8	0.938	**0.918**	0.928	**0.918**	**0.914**	**0.915**	**0.900**	**0.914**	0.880	**0.911**	0.860	**0.908**	0.837	**0.906**	0.811	**0.904**	0.781	**0.902**	0.752
9	***0.956***	0.940	0.948	0.940	0.937	0.937	0.925	0.937	**0.908**	0.934	0.891	0.932	0.870	0.930	0.847	0.928	0.819	0.927	0.792
10	0.969	***0.956***	***0.963***	***0.956***	***0.953***	***0.954***	0.944	***0.953***	0.930	***0.951***	**0.914**	0.949	0.897	0.948	0.875	0.947	0.850	0.945	0.825
11	0.978	0.968	0.973	0.968	0.966	0.966	***0.958***	0.966	0.946	0.964	0.933	***0.962***	**0.918**	***0.961***	0.899	***0.960***	0.876	***0.959***	0.853
12	0.984	0.977	0.981	0.976	0.975	0.975	0.968	0.975	***0.959***	0.973	0.948	0.972	0.934	0.971	**0.918**	0.970	0.898	0.969	0.877
13	0.989	0.983	0.986	0.983	0.981	0.982	0.976	0.981	0.968	0.980	***0.959***	0.979	0.948	0.979	0.933	0.978	**0.915**	0.977	0.896
14	0.992	0.988	0.990	0.987	0.986	0.987	0.982	0.986	0.976	0.986	0.968	0.985	***0.958***	0.984	0.946	0.983	0.930	0.983	**0.913**
15	0.994	0.991	0.993	0.991	0.990	0.990	0.987	0.990	0.981	0.989	0.975	0.989	0.967	0.988	***0.956***	0.988	0.942	0.987	0.927
16	0.996	0.993	0.995	0.993	0.993	0.993	0.990	0.993	0.986	0.992	0.980	0.992	0.973	0.991	0.964	0.991	***0.952***	0.990	0.939
17	0.997	0.995	0.996	0.995	0.995	0.995	0.992	0.995	0.989	0.994	0.985	0.994	0.979	0.993	0.971	0.993	0.960	0.993	0.948
18	0.998	0.996	0.997	0.996	0.996	0.996	0.994	0.996	0.992	0.996	0.988	0.995	0.983	0.995	0.976	0.995	0.967	0.995	***0.957***
19	0.999	0.997	0.998	0.997	0.997	0.997	0.996	0.997	0.994	0.997	0.991	0.997	0.987	0.996	0.981	0.996	0.973	0.996	0.964
20	0.999	0.998	0.999	0.998	0.998	0.998	0.997	0.998	0.995	0.998	0.993	0.997	0.989	0.997	0.984	0.997	0.978	0.997	0.969
21	0.999	0.999	0.999	0.999	0.998	0.998	0.998	0.998	0.996	0.998	0.994	0.998	0.991	0.998	0.987	0.998	0.981	0.998	0.974
22	1.000	0.999	0.999	0.999	0.999	0.999	0.998	0.999	0.997	0.999	0.996	0.999	0.993	0.999	0.990	0.998	0.985	0.998	0.978
23	1.000	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.998	0.999	0.997	0.999	0.995	0.999	0.992	0.999	0.987	0.999	0.982
24	1.000	0.999	1.000	0.999	0.999	0.999	0.999	0.999	0.998	0.999	0.997	0.999	0.996	0.999	0.993	0.999	0.990	0.999	0.985
25	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	0.999	0.998	0.999	0.997	0.999	0.995	0.999	0.991	0.999	0.987

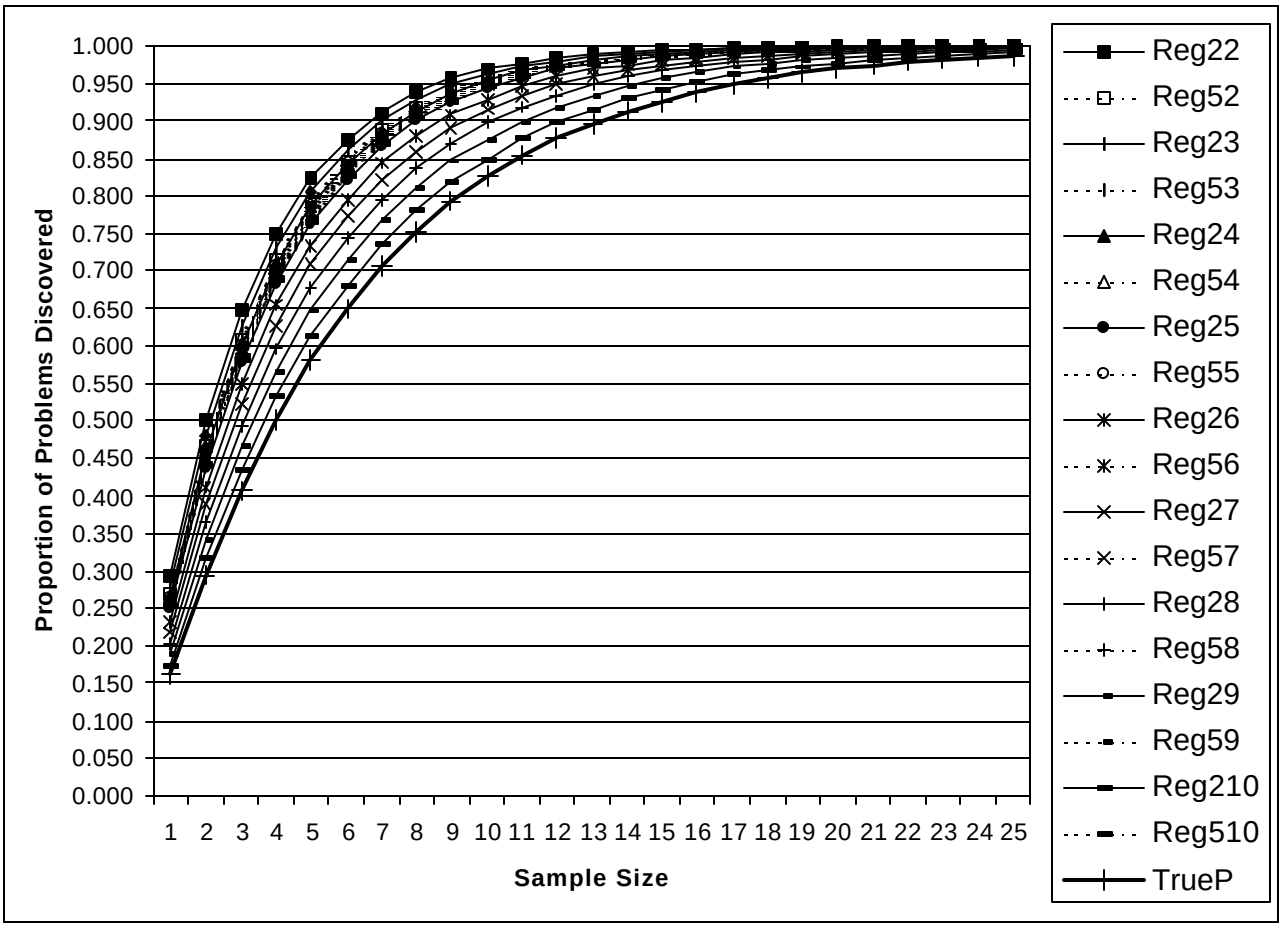

Figure 16. MACERR regression equation 2 and 5 problem discovery projections

Table 36. Regression equation 2 and 5 estimate deviation from required sample size for MACERR

N	Reg2-90	Reg5-90	Reg2-95	Reg5-95
2	7	6	9	8
3	7	6	8	8
4	6	6	8	8
5	6	6	7	8
6	5	6	6	8
7	4	6	5	7
8	3	6	4	7
9	2	6	3	7
10	1	6	2	7

Note: A positive number indicates underestimation of the required sample size.

Table 37. Regression equation 2 and 5 estimate deviation from problem discovery goal for MACERR

N	Reg2-90	Reg5-90	Reg2-95	Reg5-95
2	-0.195	-0.148	-0.158	-0.125
3	-0.195	-0.148	-0.125	-0.125
4	-0.148	-0.148	-0.125	-0.125
5	-0.148	-0.148	-0.097	-0.125
6	-0.108	-0.148	-0.073	-0.125
7	-0.075	-0.148	-0.054	-0.097
8	-0.047	-0.148	-0.037	-0.097
9	-0.023	-0.148	-0.023	-0.097
10	-0.004	-0.148	-0.011	-0.097

Note: A positive number indicates overachievement of the problem discovery goal

Table 38. Regression equation 2 and 5 problem discovery projections for VIRZI90

N	Reg 2 2	Reg 5 2	Reg 2 3	Reg 5 3	Reg 2 4	Reg 5 4	Reg 2 5	Reg 5 5	Reg 2 6	Reg 5 6	Reg 2 7	Reg 5 7	Reg 2 8	Reg 5 8	Reg 2 9	Reg 5 9	Reg 2 10	Reg 5 10	True *p*
1	0.452	0.427	0.434	0.421	0.418	0.418	0.403	0.416	0.389	0.415	0.376	0.416	0.363	0.415	0.351	0.417	0.339	0.417	0.359
2	0.700	0.672	0.680	0.665	0.661	0.661	0.644	0.659	0.627	0.658	0.611	0.659	0.594	0.658	0.579	0.660	0.563	0.660	0.589
3	0.835	0.812	0.819	0.806	0.803	0.803	0.787	0.801	0.772	0.800	0.757	0.801	0.742	0.800	0.727	0.802	0.711	0.802	0.737
4	**0.910**	0.892	0.897	0.888	0.885	0.885	0.873	0.884	0.861	0.883	0.848	0.884	0.835	0.883	0.823	0.884	0.809	0.884	0.831
5	***0.951***	**0.938**	**0.942**	**0.935**	**0.933**	**0.933**	**0.924**	**0.932**	**0.915**	**0.931**	**0.905**	**0.932**	0.895	**0.931**	0.885	**0.933**	0.874	**0.933**	0.892
6	0.973	***0.965***	***0.967***	***0.962***	***0.961***	***0.961***	***0.955***	***0.960***	***0.948***	***0.960***	***0.941***	***0.960***	**0.933**	***0.960***	**0.925**	***0.961***	**0.917**	***0.961***	**0.931**
7	0.985	0.980	0.981	0.978	0.977	0.977	0.973	0.977	0.968	0.977	0.963	0.977	***0.957***	0.977	***0.952***	0.977	0.945	0.977	***0.956***
8	0.992	0.988	0.989	0.987	0.987	0.987	0.984	0.986	0.981	0.986	0.977	0.986	0.973	0.986	0.969	0.987	***0.964***	0.987	0.971
9	0.996	0.993	0.994	0.993	0.992	0.992	0.990	0.992	0.988	0.992	0.986	0.992	0.983	0.992	0.980	0.992	0.976	0.992	0.982
10	0.998	0.996	0.997	0.996	0.996	0.996	0.994	0.995	0.993	0.995	0.991	0.995	0.989	0.995	0.987	0.995	0.984	0.995	0.988
11	0.999	0.998	0.998	0.998	0.997	0.997	0.997	0.997	0.996	0.997	0.994	0.997	0.993	0.997	0.991	0.997	0.989	0.997	0.992
12	0.999	0.999	0.999	0.999	0.998	0.998	0.998	0.998	0.997	0.998	0.997	0.998	0.996	0.998	0.994	0.998	0.993	0.998	0.995
13	1.000	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.998	0.999	0.998	0.999	0.997	0.999	0.996	0.999	0.995	0.999	0.997
14	1.000	1.000	1.000	1.000	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.998	0.999	0.998	0.999	0.997	0.999	0.998
15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.998	1.000	0.998	1.000	0.999
16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999
17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.999
18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	1.000
19	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

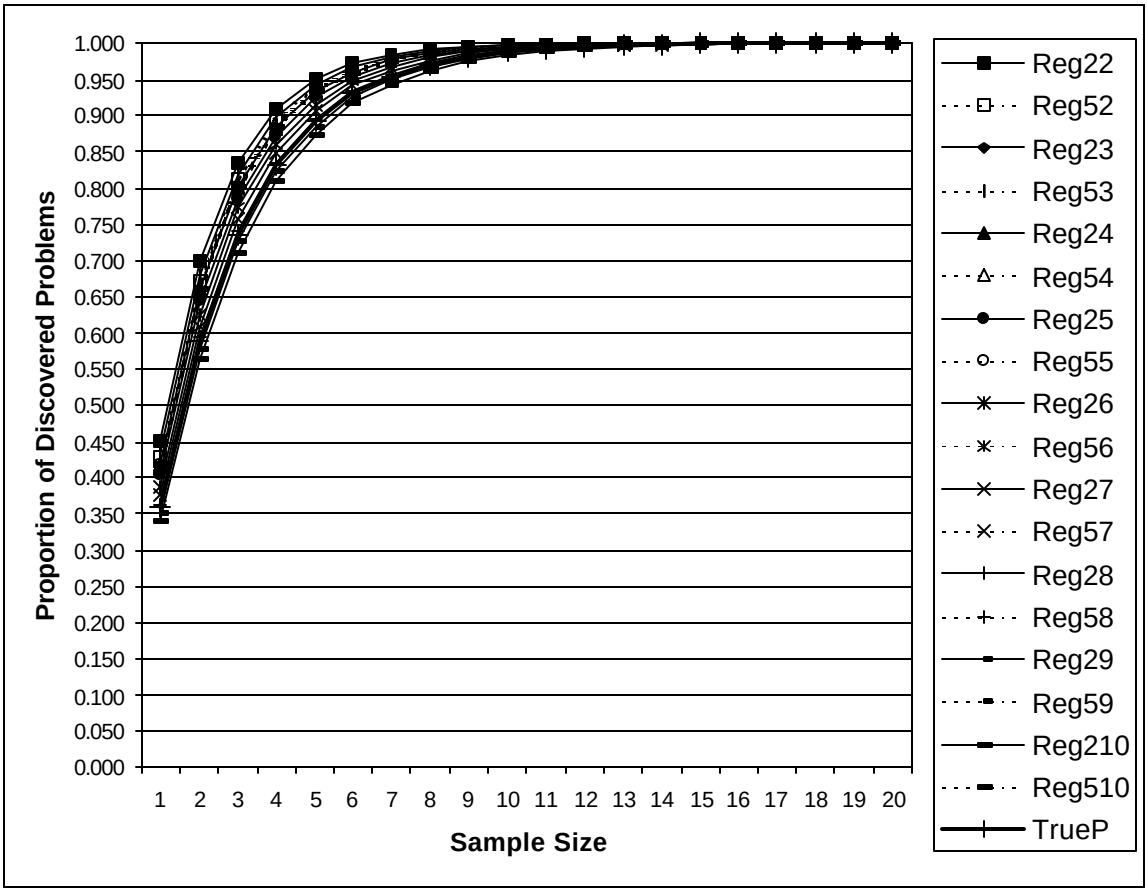

Figure 17. VIRZI90 regression equation 2 and 5 problem discovery projections

Table 39. Regression equation 2 and 5 estimation deviation from required sample size for VIRZI90

N	Reg2-90	Reg5-90	Reg2-95	Reg5-95
2	2	1	2	1
3	1	1	1	1
4	1	1	1	1
5	1	1	1	1
6	1	1	1	1
7	1	1	1	1
8	0	1	1	1
9	0	1	1	1
10	0	1	1	1

Note: A positive number indicates underestimation of the required sample size.

Table 40. Regression equation 2 and 5 estimation deviation from problem discovery goal for VIRZI90

N	Reg2-90	Reg5-90	Reg2-95	Reg5-95
2	-0.069	-0.008	-0.058	-0.019
3	-0.008	-0.008	-0.019	-0.019
4	-0.008	-0.008	-0.019	-0.019
5	-0.008	-0.008	-0.019	-0.019
6	-0.008	-0.008	-0.019	-0.019
7	-0.008	-0.008	-0.019	-0.019
8	0.031	-0.008	-0.019	-0.019
9	0.031	-0.008	-0.019	-0.019
10	0.031	-0.008	-0.019	-0.019

Note: A positive number indicates overachievement of the problem discovery goal

Table 41. Regression equation 2 and 5 problem discovery projections for MANTEL

N	Reg 2 2	Reg 5 2	Reg 2 3	Reg 5 3	Reg 2 4	Reg 5 4	Reg 2 5	Reg 5 5	Reg 2 6	Reg 5 6	Reg 2 7	Reg 5 7	Reg 2 8	Reg 5 8	Reg 2 9	Reg 5 9	Reg 2 10	Reg 5 10	True *p*
1	0.556	0.530	0.535	0.521	0.513	0.512	0.496	0.508	0.474	0.499	0.457	0.495	0.439	0.491	0.423	0.488	0.407	0.485	0.375
2	0.803	0.779	0.784	0.771	0.763	0.762	0.746	0.758	0.723	0.749	0.705	0.745	0.685	0.741	0.667	0.738	0.648	0.735	0.609
3	**0.912**	0.896	0.899	0.890	0.884	0.884	0.872	0.881	0.854	0.874	0.840	0.871	0.823	0.868	0.808	0.866	0.791	0.863	0.756
4	***0.961***	***0.951***	***0.953***	**0.947**	**0.944**	**0.943**	**0.935**	**0.941**	**0.923**	**0.937**	**0.913**	**0.935**	**0.901**	**0.933**	0.889	**0.931**	0.876	**0.930**	0.847
5	0.983	0.977	0.978	***0.975***	***0.973***	***0.972***	***0.967***	***0.971***	***0.960***	***0.968***	***0.953***	***0.967***	0.944	***0.966***	**0.936**	***0.965***	**0.927**	***0.964***	**0.905**
6	0.992	0.989	0.990	0.988	0.987	0.986	0.984	0.986	0.979	0.984	0.974	0.983	***0.969***	0.983	***0.963***	0.982	***0.957***	0.981	0.940
7	0.997	0.995	0.995	0.994	0.994	0.993	0.992	0.993	0.989	0.992	0.986	0.992	0.983	0.991	0.979	0.991	0.974	0.990	***0.963***
8	0.998	0.998	0.998	0.997	0.997	0.997	0.996	0.997	0.994	0.996	0.992	0.996	0.990	0.995	0.988	0.995	0.985	0.995	0.977
9	0.999	0.999	0.999	0.999	0.998	0.998	0.998	0.998	0.997	0.998	0.996	0.998	0.994	0.998	0.993	0.998	0.991	0.997	0.985
10	1.000	0.999	1.000	0.999	0.999	0.999	0.999	0.999	0.998	0.999	0.998	0.999	0.997	0.999	0.996	0.999	0.995	0.999	0.991
11	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.999	0.999	0.998	0.999	0.998	0.999	0.997	0.999	0.994
12	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.998	1.000	0.996
13	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.998
14	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999
15	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
16	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999
17	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
19	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

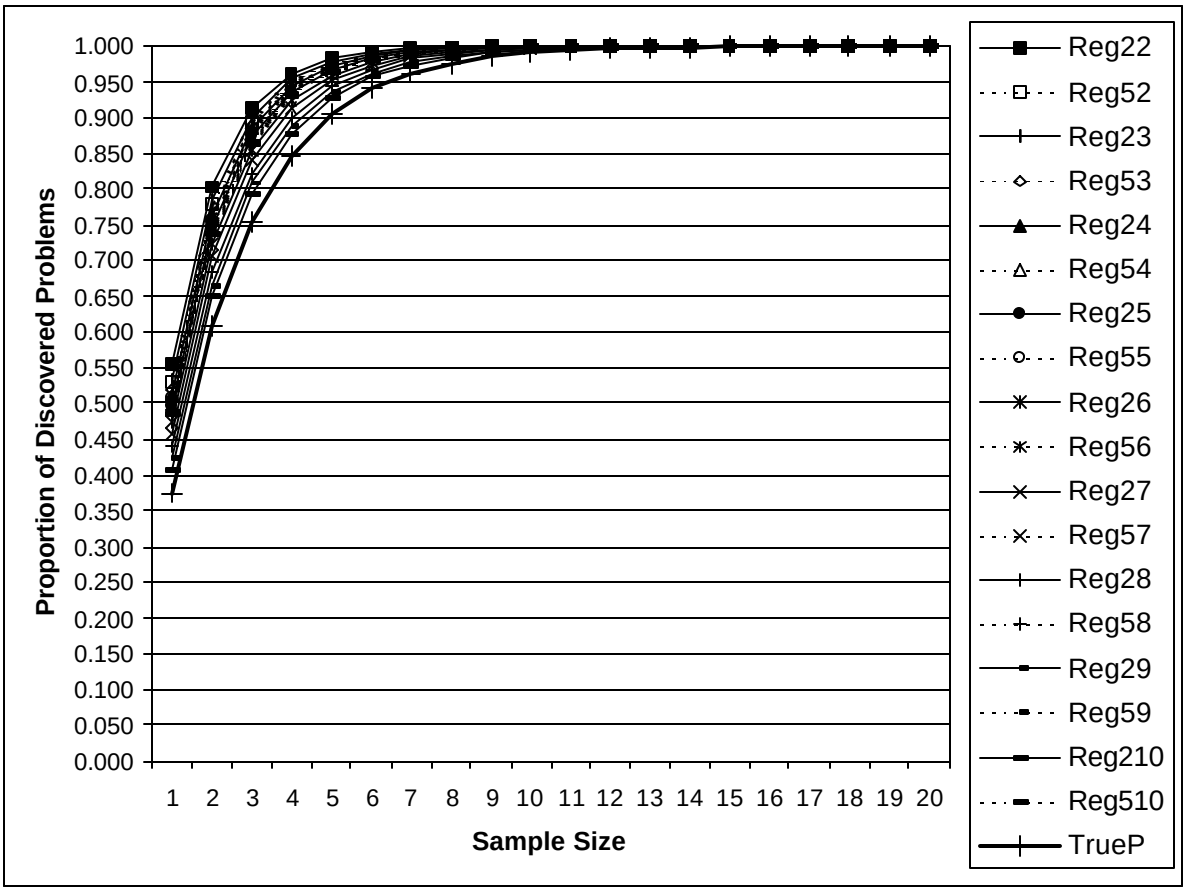

Figure 18. MANTEL regression equation 2 and 5 problem discovery projections

Table 42. Regression equation 2 and 5 estimation deviation from required sample size for MANTEL

N	Reg2-90	Reg5-90	Reg2-95	Reg5-95
2	2	1	3	3
3	1	0	2	2
4	1	0	2	2
5	1	0	2	2
6	1	0	2	2
7	1	0	2	2
8	1	0	1	2
9	0	0	1	2
10	0	0	1	2

Note: A positive number indicates underestimation of the required sample size.

Table 43. Regression equation 2 and 5 estimation deviation from problem discovery goal for MANTEL

N	Reg2-90	Reg5-90	Reg2-95	Reg5-95
2	-0.144	-0.053	-0.103	-0.103
3	-0.053	0.005	-0.045	-0.045
4	-0.053	0.005	-0.045	-0.045
5	-0.053	0.005	-0.045	-0.045
6	-0.053	0.005	-0.045	-0.045
7	-0.053	0.005	-0.045	-0.045
8	-0.053	0.005	-0.010	-0.045
9	0.005	0.005	-0.010	-0.045
10	0.005	0.005	-0.010	-0.045

Note: A positive number indicates overachievement of the problem discovery goal

Table 44. Regression equation 2 and 5 problem discovery projections for SAVE

N	Reg 2 2	Reg 5 2	Reg 2 3	Reg 5 3	Reg 2 4	Reg 5 4	Reg 2 5	Reg 5 5	Reg 2 6	Reg 5 6	Reg 2 7	Reg 5 7	Reg 2 8	Reg 5 8	Reg 2 9	Reg 5 9	Reg 2 10	Reg 5 10	True *p*
1	0.398	0.373	0.384	0.371	0.370	0.370	0.359	0.372	0.344	0.370	0.331	0.371	0.318	0.370	0.303	0.369	0.291	0.369	0.256
2	0.638	0.607	0.621	0.604	0.603	0.603	0.589	0.606	0.570	0.603	0.552	0.604	0.535	0.603	0.514	0.602	0.497	0.602	0.446
3	0.782	0.754	0.766	0.751	0.750	0.750	0.737	0.752	0.718	0.750	0.701	0.751	0.683	0.750	0.661	0.749	0.644	0.749	0.588
4	0.869	0.845	0.856	0.843	0.842	0.842	0.831	0.844	0.815	0.842	0.800	0.843	0.784	0.842	0.764	0.841	0.747	0.841	0.694
5	**0.921**	**0.903**	**0.911**	**0.902**	**0.901**	**0.901**	0.892	**0.902**	0.879	**0.901**	0.866	**0.902**	0.852	**0.901**	0.836	**0.900**	0.821	**0.900**	0.772
6	***0.952***	0.939	0.945	0.938	0.937	0.937	**0.931**	0.939	**0.920**	0.937	**0.910**	0.938	0.899	0.937	0.885	0.937	0.873	0.937	0.830
7	0.971	***0.962***	***0.966***	***0.961***	***0.961***	***0.961***	***0.956***	***0.961***	0.948	***0.961***	0.940	***0.961***	**0.931**	***0.961***	**0.920**	***0.960***	**0.910**	***0.960***	0.874
8	0.983	0.976	0.979	0.975	0.975	0.975	0.971	0.976	***0.966***	0.975	***0.960***	0.975	***0.953***	0.975	0.944	0.975	0.936	0.975	**0.906**
9	0.990	0.985	0.987	0.985	0.984	0.984	0.982	0.985	0.978	0.984	0.973	0.985	0.968	0.984	***0.961***	0.984	***0.955***	0.984	0.930
10	0.994	0.991	0.992	0.990	0.990	0.990	0.988	0.990	0.985	0.990	0.982	0.990	0.978	0.990	0.973	0.990	0.968	0.990	0.948
11	0.996	0.994	0.995	0.994	0.994	0.994	0.992	0.994	0.990	0.994	0.988	0.994	0.985	0.994	0.981	0.994	0.977	0.994	***0.961***
12	0.998	0.996	0.997	0.996	0.996	0.996	0.995	0.996	0.994	0.996	0.992	0.996	0.990	0.996	0.987	0.996	0.984	0.996	0.971
13	0.999	0.998	0.998	0.998	0.998	0.998	0.997	0.998	0.996	0.998	0.995	0.998	0.993	0.998	0.991	0.997	0.989	0.997	0.979
14	0.999	0.999	0.999	0.998	0.998	0.998	0.998	0.999	0.997	0.998	0.996	0.998	0.995	0.998	0.994	0.998	0.992	0.998	0.984
15	1.000	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.998	0.999	0.998	0.999	0.997	0.999	0.996	0.999	0.994	0.999	0.988
16	1.000	0.999	1.000	0.999	0.999	0.999	0.999	0.999	0.999	0.999	0.998	0.999	0.998	0.999	0.997	0.999	0.996	0.999	0.991
17	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.998	1.000	0.997	1.000	0.993
18	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.998	1.000	0.998	1.000	0.995
19	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.996
20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.997

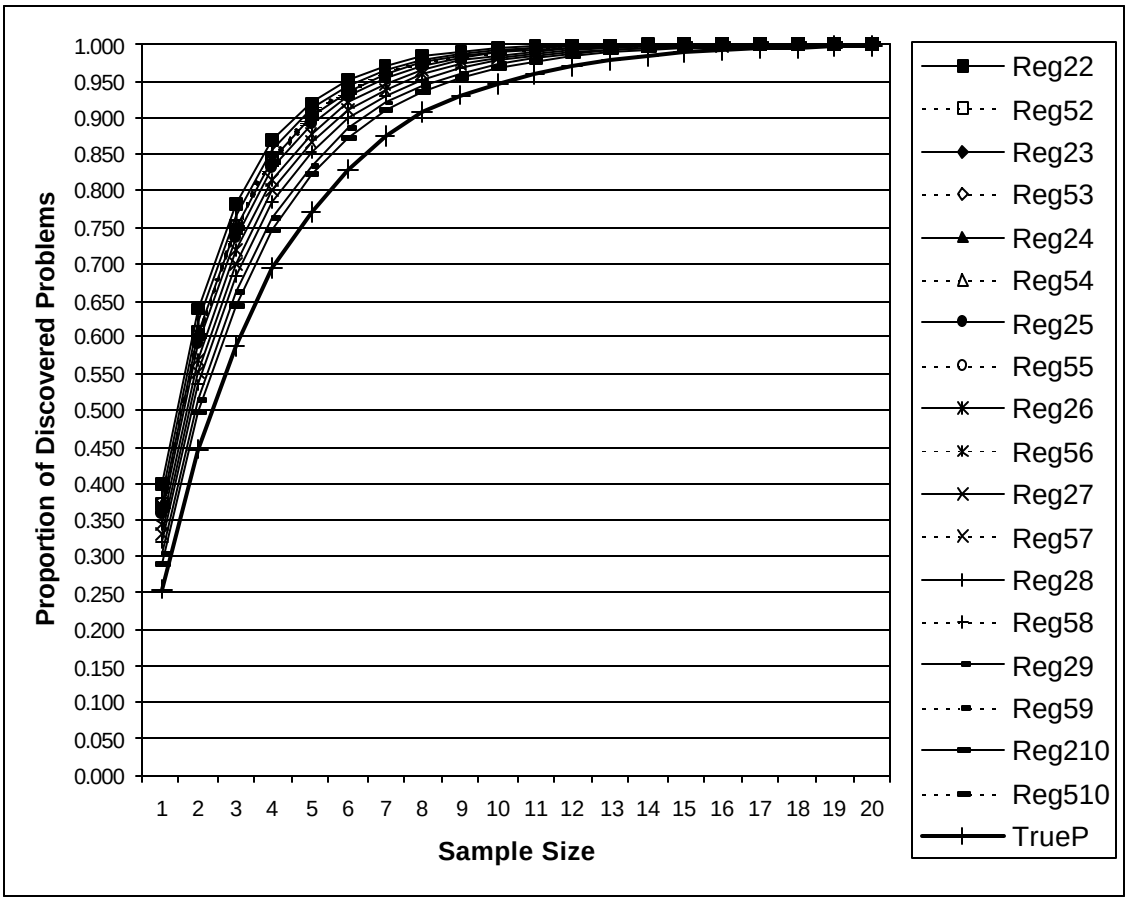

Figure 19. SAVE regression equation 2 and 5 problem discovery projections

Table 45. Regression equation 2 and 5 estimation deviation from required sample size for SAVE

N	Reg2-90	Reg5-90	Reg2-95	Reg5-95
2	3	3	5	4
3	3	3	4	4
4	3	3	4	4
5	2	3	4	4
6	2	3	3	4
7	2	3	3	4
8	1	3	2	4
9	1	3	2	4
10	1	3	2	4

Note: A positive number indicates underestimation of the required sample size.

Table 46. Regression equation 2 and 5 estimation deviation from problem discovery goal for SAVE

N	Reg2-90	Reg5-90	Reg2-95	Reg5-95
2	-0.128	-0.128	-0.120	-0.076
3	-0.128	-0.128	-0.076	-0.076
4	-0.128	-0.128	-0.076	-0.076
5	-0.070	-0.128	-0.076	-0.076
6	-0.070	-0.128	-0.044	-0.076
7	-0.070	-0.128	-0.044	-0.076
8	-0.026	-0.128	-0.020	-0.076
9	-0.026	-0.128	-0.020	-0.076
10	-0.026	-0.128	-0.020	-0.076

Note: A positive number indicates overachievement of the problem discovery goal

Deviation from Required Sample Sizes for 90% and 95% Problem Discovery

I conducted a within-subjects analysis of variance on the deviations from required sample sizes, treating problem discovery databases as subjects. The independent variables were adjustment method (None, Norm, Reg2, Reg5, GT, Comb), sample size used to estimate p (2, 3, 4, 5, 6, 7, 8, 9, 10), and problem discovery goal (90%, 95%). The analysis indicated the following significant effects:

- main effect of adjustment method ($F(5,15)=4.1$, $p=.015$)
- main effect of sample size ($F(8,24)=3.8$, $p=.005$)
- adjustment method by discovery goal interaction ($F(5,15)=3.9$, $p=.019$)
- adjustment method by sample size interaction ($F(40,120)=3.2$, $p=.0000006$)
- adjustment method by sample size by problem discovery goal interaction ($F(40,120)=2.0$, $p=.002$)

Tables 47-50 and Figures 20-24 illustrate these effects.

Table 47. Main effect of adjustment method

Type	**None**	**Norm**	**Reg2**	**Reg5**	**GT**	**Combo**
Underestimation	3.7	-1.3	2.5	3.1	0.7	-0.1

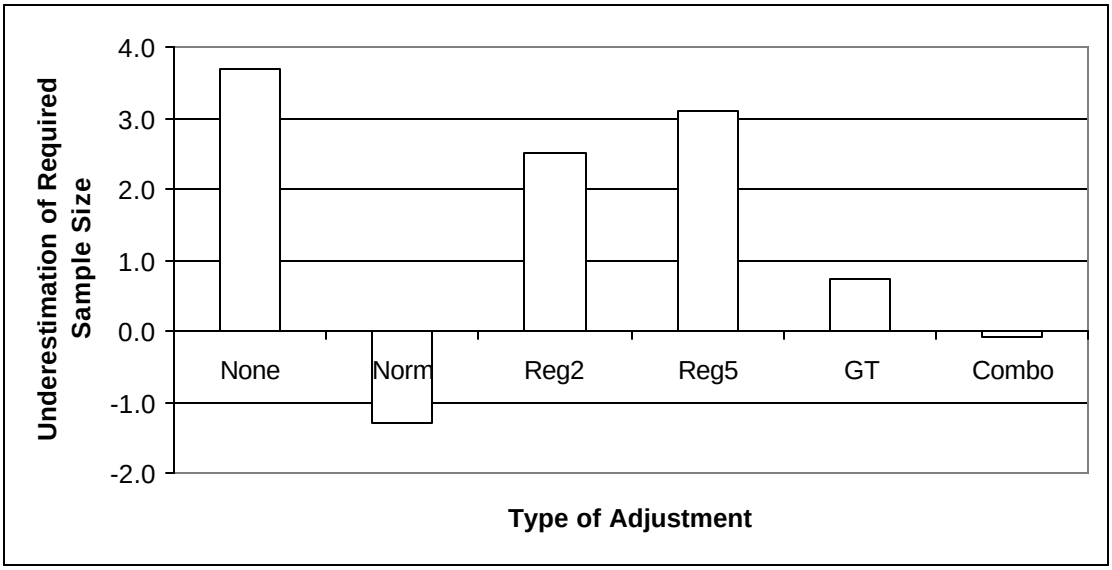

Figure 20. Sample size underestimation as a function of adjustment method

Table 48. Main effect of sample size used to estimate p

Sample	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Underestimation	3.1	2.4	1.9	1.7	1.3	1.0	0.8	0.5	0.3

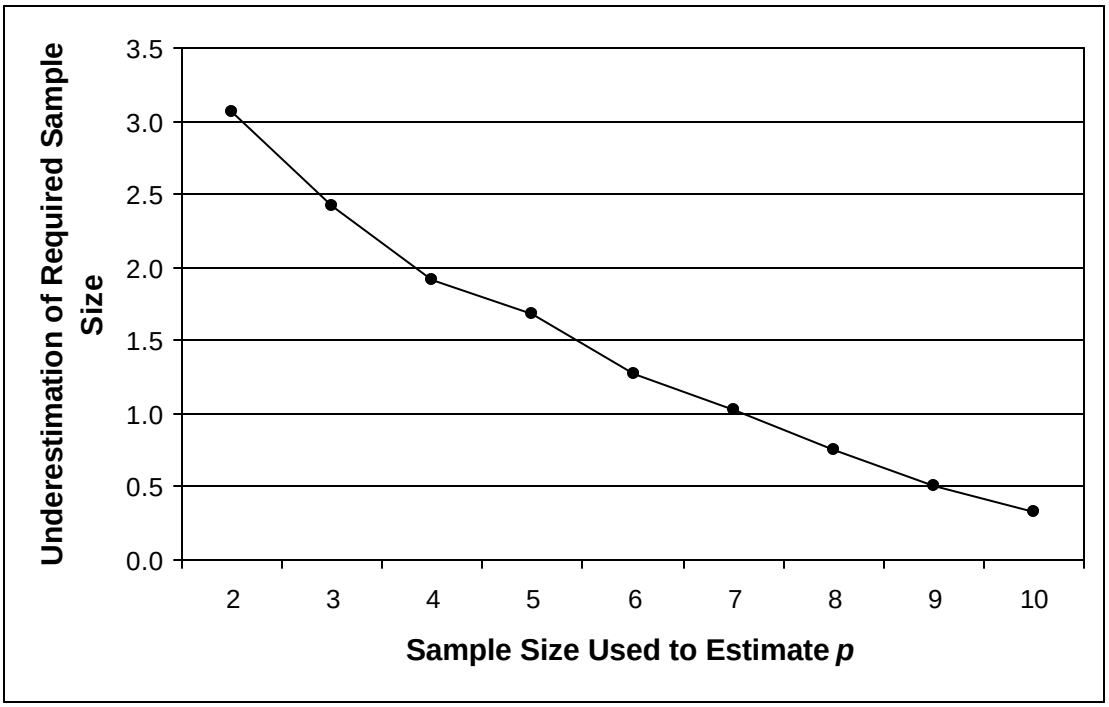

Figure 21. Sample size underestimation as a function of sample size used to estimate p

Table 49. Adjustment type by discovery goal interaction

Goal	None	Norm	Reg2	Reg5	GT	Combo
90%	3.2	-1.2	2.1	2.5	0.6	-0.2
95%	4.2	-1.4	3.0	3.7	0.9	0.0

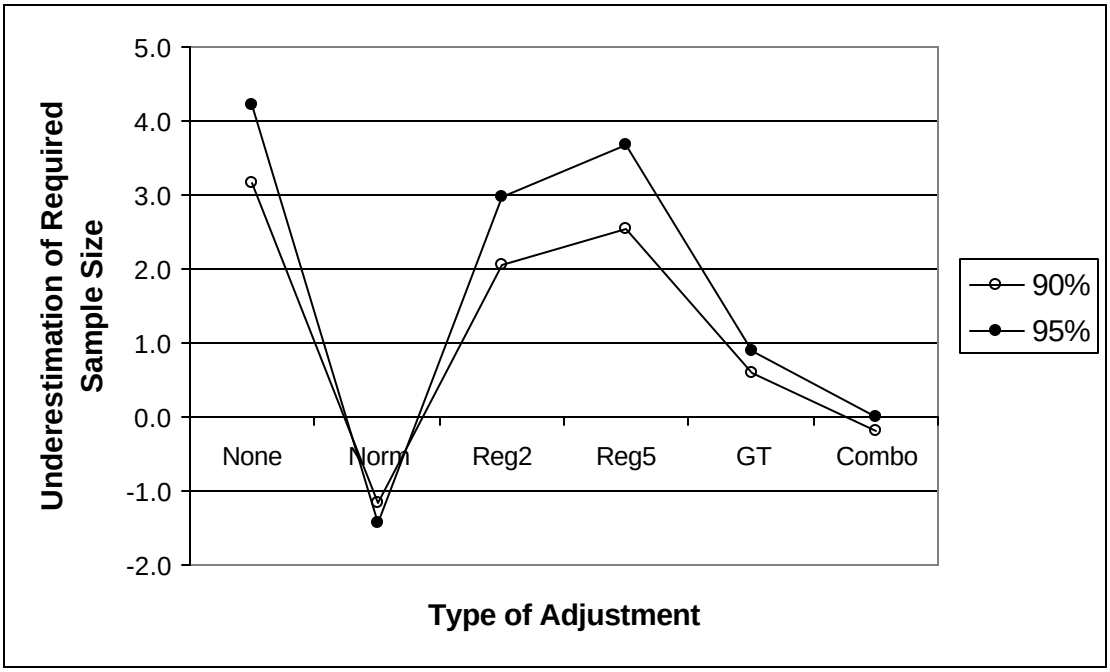

Figure 22. Sample size underestimation as a function of adjustment type and discovery goal

Table 49. Interaction of adjustment type and sample size used to estimate p

Sample	**None**	**Norm**	**Reg2**	**Reg5**	**GT**	**Combo**
2	6.4	-0.8	4.1	3.4	3.4	1.9
3	5.3	-0.9	3.4	3.1	2.5	1.1
4	4.5	-1.0	3.3	3.1	1.3	0.4
5	4.0	-1.1	3.0	3.1	1.0	0.1
6	3.3	-1.4	2.6	3.1	0.4	-0.4
7	2.9	-1.5	2.4	3.0	0.0	-0.6
8	2.8	-1.6	1.6	3.0	-0.4	-0.9
9	2.3	-1.6	1.3	3.0	-0.6	-1.3
10	2.0	-1.9	1.0	3.0	-0.9	-1.3

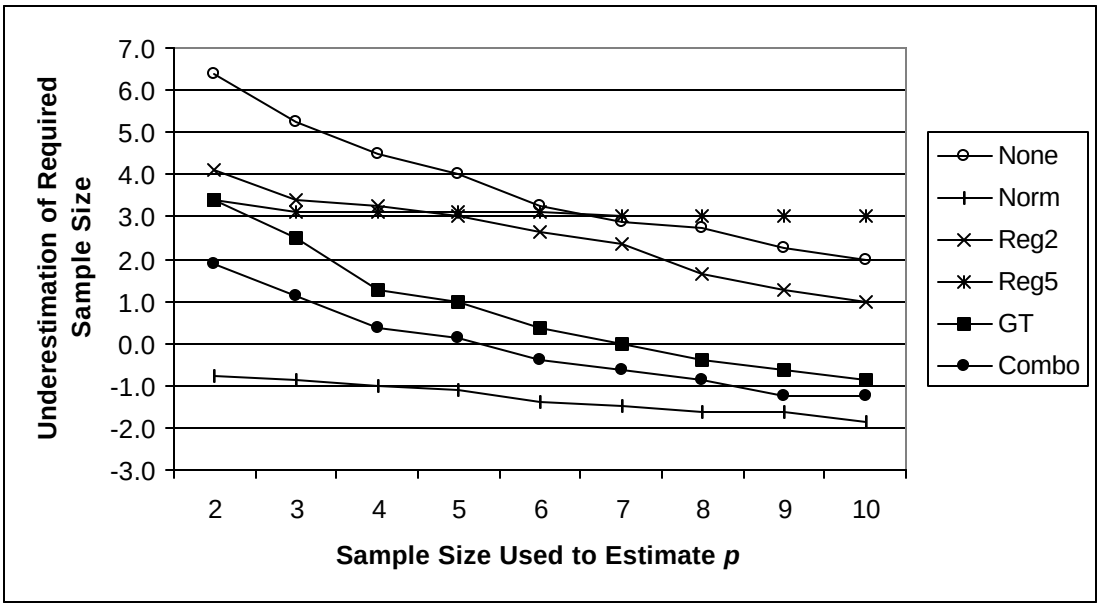

Figure 23. Sample size underestimation as a function of adjustment type and sample size used to estimate p

Table 50. Interaction of adjustment type, problem discovery goal, and sample size used to estimate p

Sample	**None-90**	**None-95**	**Norm-90**	**Norm-95**	**Reg2-90**	**Reg2-95**
2	5.5	7.3	-0.5	-1.0	3.5	4.8
3	4.5	6.0	-0.8	-1.0	3.0	3.8
4	4.0	5.0	-1.0	-1.0	2.8	3.8
5	3.5	4.5	-1.0	-1.3	2.5	3.5
6	2.8	3.8	-1.3	-1.5	2.3	3.0
7	2.3	3.5	-1.3	-1.8	2.0	2.8
8	2.3	3.3	-1.5	-1.8	1.3	2.0
9	2.0	2.5	-1.5	-1.8	0.8	1.8
10	1.8	2.3	-1.8	-2.0	0.5	1.5
Sample	**Reg5-90**	**Reg5-95**	**GT-90**	**GT-95**	**Comb-90**	**Comb-95**
2	2.8	4.0	2.8	4.0	1.5	2.3
3	2.5	3.8	2.0	3.0	1.0	1.3
4	2.5	3.8	1.0	1.5	0.3	0.5
5	2.5	3.8	0.8	1.3	0.0	0.3
6	2.5	3.8	0.3	0.5	-0.5	-0.3
7	2.5	3.5	0.0	0.0	-0.8	-0.5
8	2.5	3.5	-0.3	-0.5	-1.0	-0.8
9	2.5	3.5	-0.5	-0.8	-1.3	-1.3
10	2.5	3.5	-0.8	-1.0	-1.0	-1.5

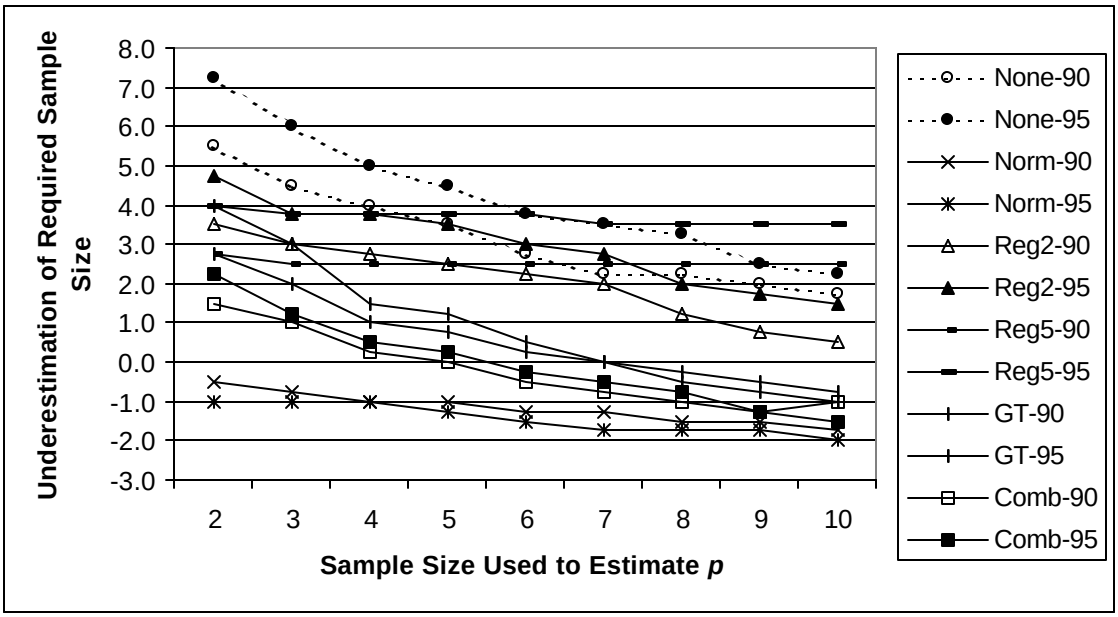

Figure 24. Sample size underestimation as a function of adjustment type, problem discovery goal, and sample size used to estimate p

Deviation from 90% and 95% Problem Discovery Goals

Because the fundamental issue in underestimating a required sample size is the failure to achieve a specified problem discovery goal, I also conducted a within-subjects analysis of variance on the deviations from specified problem discovery goals of 90% and 95% discovery, treating problem discovery databases as subjects. The independent variables were adjustment method (None, Norm, Reg2, Reg5, GT, Comb), sample size used to estimate p (2, 3, 4, 5, 6, 7, 8, 9, 10), and problem discovery goal (90%, 95%). The analysis indicated the following significant effects:

- main effect of adjustment method ($F(5,15)=13.6$, $p=.00004$)
- main effect of sample size ($F(8,24)=15.5$, $p=.0000001$)
- adjustment method by sample size interaction ($F(40,120)=12.5$, $p=.005$)
- sample size by problem discovery goal interaction ($F(8,24)=3.8$, $p=.006$)
- adjustment method by sample size by discovery goal interaction ($F(40,120)=1.4$, $p=.07$)

Tables 51-55 and Figures 25-29 illustrate these effects.

Table 51. Main effect of adjustment method

Type	**None**	**Norm**	**Reg2**	**Reg5**	**GT**	**Combo**
Underestimation	3.7	-1.3	2.5	3.1	0.7	-0.1

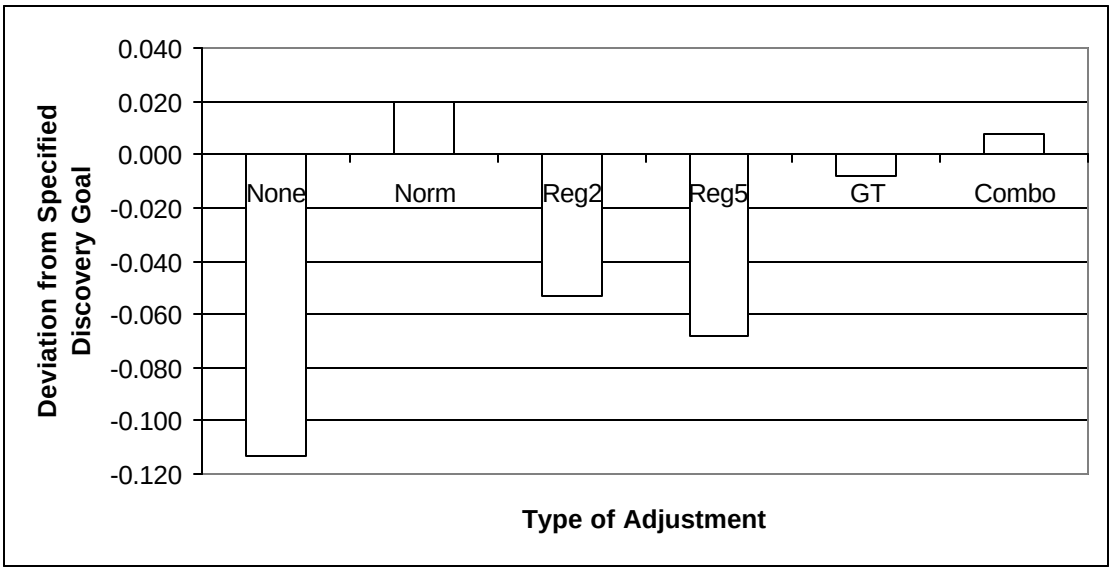

Figure 25. Discovery goal deviation as a function of adjustment method

Table 52. Main effect of sample size used to estimate p

Sample	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Underestimation	-0.100	-0.065	-0.045	-0.037	-0.025	-0.020	-0.014	-0.010	-0.008

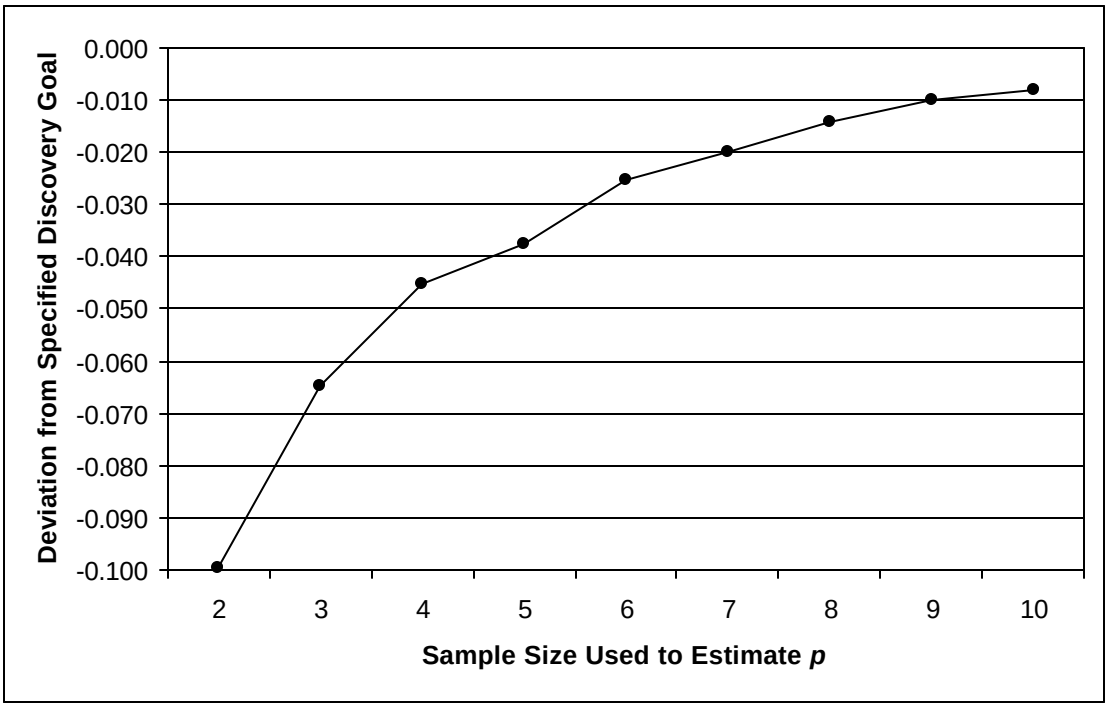

Figure 26. Discovery goal deviation as a function of sample size used to estimate p

Table 53. Interaction of adjustment type and sample size used to estimate p

Sample	**None**	**Norm**	**Reg2**	**Reg5**	**GT**	**Combo**
2	-0.296	0.011	-0.122	-0.083	-0.078	-0.030
3	-0.192	0.014	-0.081	-0.068	-0.048	-0.013
4	-0.144	0.021	-0.075	-0.068	-0.010	0.006
5	-0.117	0.022	-0.065	-0.068	-0.006	0.009
6	-0.075	0.024	-0.053	-0.068	0.006	0.015
7	-0.060	0.024	-0.046	-0.065	0.010	0.017
8	-0.057	0.025	-0.023	-0.065	0.015	0.019
9	-0.043	0.018	-0.011	-0.065	0.017	0.023
10	-0.035	0.019	-0.007	-0.065	0.019	0.021

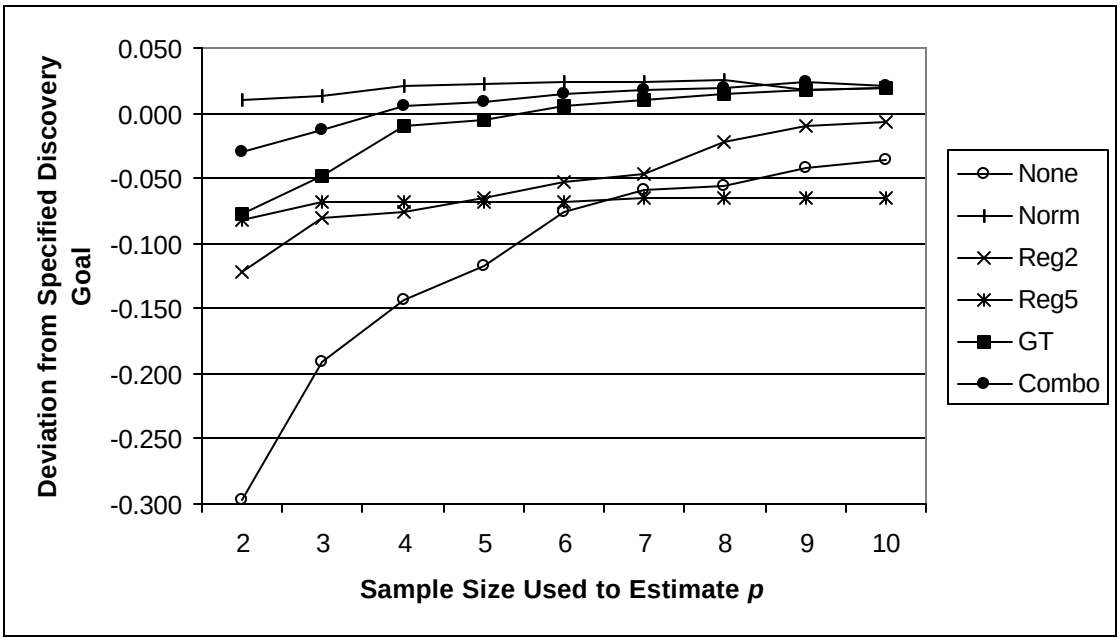

Figure 27. Discovery goal deviation as a function of adjustment type and sample size used to estimate p

Table 54. Interaction of discovery goal and sample size used to estimate p

Sample	Disc90%	Disc95%
2	-0.106	-0.093
3	-0.071	-0.058
4	-0.049	-0.041
5	-0.039	-0.036
6	-0.025	-0.025
7	-0.018	-0.021
8	-0.012	-0.016
9	-0.007	-0.013
10	-0.005	-0.011

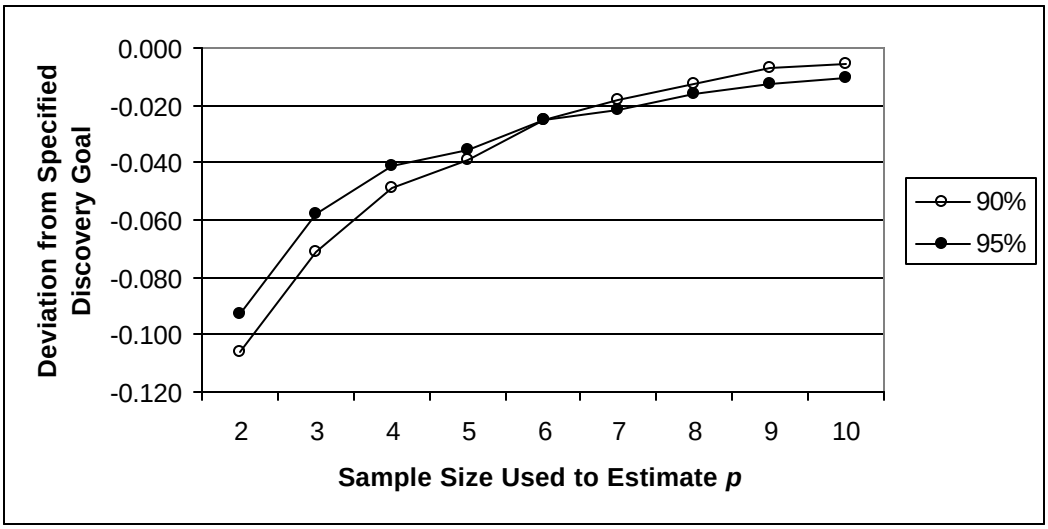

Figure 28. Discovery goal deviation as a function of discovery goal and sample size used to estimate p

Table 55. Interaction of adjustment type, problem discovery goal, and sample size used to estimate p

Sample	**None-90**	**None-95**	**Norm-90**	**Norm-95**	**Reg2-90**	**Reg2-95**
2	-0.315	-0.278	0.008	0.013	-0.134	-0.110
3	-0.208	-0.175	0.014	0.013	-0.096	-0.066
4	-0.168	-0.120	0.029	0.013	-0.084	-0.066
5	-0.134	-0.100	0.029	0.015	-0.070	-0.059
6	-0.084	-0.066	0.031	0.016	-0.060	-0.045
7	-0.060	-0.059	0.031	0.017	-0.052	-0.040
8	-0.060	-0.054	0.032	0.017	-0.024	-0.021
9	-0.051	-0.034	0.032	0.004	-0.003	-0.018
10	-0.044	-0.026	0.034	0.005	0.002	-0.015
Sample	**Reg5-90**	**Reg5-95**	**GT-90**	**GT-95**	**Comb-90**	**Comb-95**
2	-0.084	-0.081	-0.081	-0.075	-0.031	-0.028
3	-0.070	-0.066	-0.050	-0.045	-0.018	-0.008
4	-0.070	-0.066	-0.009	-0.011	0.009	0.002
5	-0.070	-0.066	-0.004	-0.008	0.014	0.004
6	-0.070	-0.066	0.009	0.002	0.023	0.007
7	-0.070	-0.059	0.014	0.006	0.026	0.009
8	-0.070	-0.059	0.017	0.013	0.029	0.010
9	-0.070	-0.059	0.020	0.014	0.031	0.016
10	-0.070	-0.059	0.022	0.015	0.025	0.017

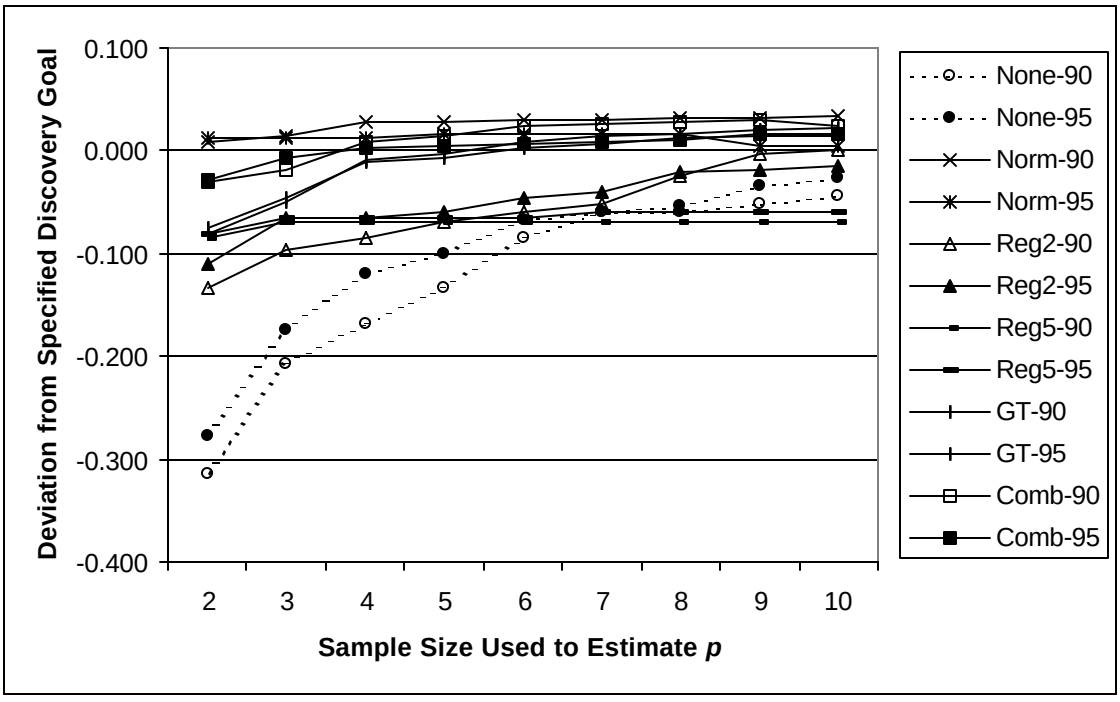

Figure 29. Sample size underestimation as a function of adjustment type, problem discovery goal, and sample size used to estimate p

Discussion

General Superiority of Combined Estimator for Adjusting *p*

It is clearly important to adjust small-sample estimates of *p* to compensate for the overestimation bias identified by Hertzum and Jacobsen (in press). The decision about which adjustment procedure(s) to recommend should take into account both the central tendency and variability of distributions created by applying the adjustment procedure(s). A measure that produces mean estimates close in value to true *p* will generally be more accurate in the long run. A measure with low variability is less likely to produce an extreme outlier in any single study. Usability practitioners do not typically conduct large studies, however, so it is important to balance benefits associated with the statistical long run with the benefits associated with reducing the risk of encountering an extreme outlier.

The regression equations (Reg2 and Reg5) tended to be less variable than other adjustment procedures, but their accuracy was very poor relative to all other adjustment procedures. The accuracy of Reg2 improved as a function of the sample size used to estimate *p*, but the accuracy of Reg5 did not. Their relatively poor performance removes these regression equations from consideration as a recommended method for adjusting *p*.

The remaining procedures (normalization, Good-Turing, combined normalization/Good-Turing) show similar patterns for overestimation ratios and deviations from true *p* (Figures 3 and 4). When estimating *p* with smaller sample sizes (2-4 participants), the curves for these three measures showed some separation, with the differences diminishing and the curves converging as the size of the sample used to estimate *p* increased. For these measures, especially at small sample sizes, the normalization procedure appeared to produce the best results and the combined estimator produced the second-best results.

The measures of variance (interquartile range, 90% range – Figures 6 and 7), however, indicated that the estimates produced by the normalization procedure were much more variable than those produced by the Good-Turing or the combined estimator. The results for the root mean square (rms) error (which take both central-tendency accuracy and variability into account) showed that all three measures had essentially equal accuracy at all levels of sample size from two to ten. As expected, the variability of all measures decreased as a function of the sample size used to estimate *p*.

Considering all this information, the combined estimator seems to provide the best balance between central-tendency accuracy and lower variability, making it the preferred adjustment procedure. What really matters, though, is the extent to which the adjustment procedure leads to accurate sample size estimation and achievement of specified problem discovery goals.

The analyses of underestimation of required sample sizes and deviation from problem discovery goals also support the use of the combined estimator. As shown in Figure 20, after averaging across all problem discovery databases (MACERR, VIRZI90, MANTEL, SAVE), both

problem discovery goals (90%, 95%) and sample sizes from two to ten, the accuracy of sample size estimation with the combined estimator was almost perfect, deviating from the required sample size by only –0.1 participant on average. The results were similar for mean deviation from problem discovery goal (Figure 25). That figure shows the magnitude of deviation to be about the same for the combined estimator and the Good-Turing estimator. On average, however, the combined estimator tended to slightly overachieve the discovery goal while the Good-Turing estimator tended to slightly underachieve the discovery goal.

With one exception, the patterns of results for the significant interactions support the unqualified use of the combined estimator. The exception is the interaction between adjustment method and the sample size used to estimate *p*. As shown in Figure 23, at sample sizes of 2 and 3 the combined estimator tends to underestimate the required sample size while the normalization procedure tends to overestimate it. The same interaction for the deviation from the discovery goal, however, indicates that the consequence of this underestimation of the required sample size is slight, even when the sample size used to estimate *p* is only two participants (in which case the underachievement is, on average, 3%). This does suggest some need on the part of practitioners to balance the cost of additional participants against their need to achieve a specific problem discovery goal. If the former is more important, then the practitioner should use the combined estimator. If the latter is more important and the practitioner is estimating *p* from a very small sample size, then it would be reasonable to use the more conservative normalization procedure.

Using the Combined Estimator for Usability Study Sample Size Estimation

One practical application for the use of the combined estimator is to allow usability practitioners to estimate their final sample size requirement from their first few participants. To do this, practitioners must keep a careful record of which participants experienced which usability problems so they will be able to make an unadjusted initial estimate of *p* that they then adjust using the combined estimator.

Suppose, for example, that a practitioner is conducting a study on a product with problem discovery characteristics similar to MACERR (true *p* of .16), and the practitioner has set a goal of discovering 90% of the problems. Referring to Table 11, the initial unadjusted estimate of *p* calculated from the first two participants would be, on average, .566. Adjusting this initial estimate with the combined procedure results in a value for *p* of .218. Using $(1-(1-p)^n)$ to project the sample size until the estimated proportion of discovery exceeds .900 yields a preliminary sample size requirement of 10 participants. After the practitioner runs two more participants toward the goal of 10 participants, he or she would recalculate the adjusted value of *p* to be .165, and would project the final required sample size to be 13 participants. As shown in the "True *p*" column of the table, the sample size actually required to exceed a problem discovery proportion of .900 is 14 participants. With 13 participants, the true proportion of problem discovery is .896, which misses the specified problem discovery target goal by only .004 – less than half a percent.

Table 56 shows the outcome of repeating this exercise for each database and for each investigated problem discovery goal. These outcomes show that following this procedure leads to very accurate estimates of sample sizes and very little deviation from problem discovery goals – a remarkable outcome given the differences in the usability studies that produced these problem discovery databases.

Table 56. Problem discovery outcomes achieved by using the combined estimator at sample sizes of two and four participants

Database	***True-p***	***Goal***	**Est-*p* | *n*=2**	**Comb-*p* | *n*=2**	***N* | *n*=2**	**Est-*p* | *n*=4**	**Comb-*p* | *n*=4**	***N* | *n*=4**	**True-*N***	**Deviation from Goal**
MACERR	*.16*	*90%*	.566	.218	10	.346	.165	13	14	-.004
MACERR	*.16*	*95%*	.566	.218	13	.346	.165	17	18	-.002
VIRZI90	*.36*	*90%*	.662	.361	6	.485	.328	6	6	.031
VIRZI90	*.36*	*95%*	.662	.361	7	.485	.328	8	7	.021
MANTEL	*.38*	*90%*	.725	.462	4	.571	.429	5	5	.005
MANTEL	*.38*	*95%*	.725	.462	5	.571	.429	6	7	-.010
SAVE	*.26*	*90%*	.629	.311	7	.442	.277	8	8	.006
SAVE	*.26*	*95%*	.629	.311	9	.442	.277	10	11	-.002

Notes: Est-*p* | *n*=2 is the unadjusted estimate of *p* given a sample size of 2 participants.
Comb-*p* | *n*=2 is the adjusted estimate of *p* using the combined estimator given a sample size of 2.
N | *n*=2 is the projected sample size requirement given *p* estimated from a sample size of 2.
Est-*p* | *n*=4 is the unadjusted estimate of *p* given a sample size of 4 participants.
Comb-*p* | *n*=4 is the adjusted estimate of *p* using the combined estimator given a sample size of 4.
N | *n*=4 is the projected sample size requirement given *p* estimated from a sample size of 4.
True-*p* is the value of *p* estimated from the entire database.
True-*N* is the sample size requirement projected from True-*p*.

Conclusions

- The overestimation of *p* from small-sample usability studies is a real problem with potentially troubling consequences for usability practitioners.
- It is possible to compensate for the overestimation bias of *p* calculated from small-sample usability studies.
- The combined normalization/Good-Turing estimator is the best procedure for adjusting initial estimates of *p* calculated from small samples (two to ten participants).
- If (1) the cost of additional participants is low, (2) the sample size used to estimate *p* is very small (two or three participants), and (3) it is very important to achieve or exceed specified problem discovery goals, then practitioners should use the normalization procedure to adjust the initial estimate of *p*.
- Practitioners can obtain highly accurate sample size estimates for 90% and 95% problem discovery goals 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.

References

Chapanis, A. (1988). Some generalizations about generalization. *Human Factors*, *30*, 253-267.

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. (1982). Testing small-system customer set-up. In *Proceedings of the Human Factors Society 26th Annual Meeting* (pp. 718-720). Santa Monica, CA: Human Factors Society.

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). *Overestimation of p in problem discovery usability studies: How serious is the problem?* (Tech Report in press). Raleigh, NC: International Business Machines Corp.

Lewis, J. R. (2000b). *Reducing the overestimation of p in problem discovery usability studies: Normalization, regression, and a combination normalization/Good-Turing approach* (Tech. Report in press). Raleigh, NC: International Business Machines Corp.

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

Lewis, J. R. (2000d). *Using discounting methods to reduce overestimation of p in problem discovery usability studies* (Tech. Report in press). Raleigh, NC: International Business Machines Corp.

Lewis, J. R. (2000e). *Validation of Monte Carlo estimation of problem discovery likelihood* (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.

Nielsen, J., and Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Conference Proceedings on Human Factors in Computing Systems – CHI93* (pp. 206-213). New York, NY: ACM.

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.