Reducing the Overestimation of *p* in Problem Discovery Usability Studies: Normalization, Regression, and a Combination Normalization/Good-Turing Approach

TR 29.3361

James R. Lewis

Speech Product Design and Usability

West Palm Beach, Florida

Abstract

Overestimation of the likelihood of problem discovery (*p*) in usability studies is serious because it leads to underestimation of required sample sizes. The current experiments demonstrated that both normalization and multiple regression can significantly reduce this overestimation. Normalization proved to be as accurate as discounting and multiple regression methods, but tended to produce underestimates rather than overestimates of *p*. Normalization is the preferred method because it does not require empirical estimation of regression weights. A combination of normalization and Good-Turing discounting appears to provide very accurate sample size projection, even with *p* estimated from as few as two participants.

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*.

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 (2000b) 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. A second follow-on Monte Carlo study (Lewis, 2000d) illustrated the use of the Good-Turing estimator (Jelinek, 1997; Manning and Schutze, 1999) to discount (adjust to a lower value) the original estimate of *p*, resulting in much more accurate projection of required sample sizes. Estimates of *p* adjusted with the Good-Turing method still tended to slightly overestimate *p*, 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. 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. It isn't clear, though, how serious the underestimation would typically be, so submitting this procedure to evaluation via a Monte Carlo experiment would be reasonable.

Using Multiple Regression to Model *p*

Another approach for the estimation of true *p* from a small sample study would be to develop one or more multiple regression models (Cliff, 1987; Draper & Smith, 1966; Pedhazur, 1982). The goal of the models would be to predict true *p* from information available in the output of a small sample usability problem discovery study, such as the original estimate of *p*, a normalized estimate of *p*, and the sample size.

Purpose of these Studies

The purpose of the current Monte Carlo experiments was to investigate the extent to which different multiple regression models and the normalization procedure compensate for the overestimation of *p*. To do this, it was necessary to:

- Develop regression models for the prediction of true *p*
- Test the accuracy of prediction for the regression models and the normalization procedure

Experiment 1: Development of Regression Models

Purpose

The purpose of this experiment was to generate data for the development of different multiple regression models for predicting the true value of *p*.

Method

I wrote a BASIC program that produced the following measurements for each of 1000 Monte Carlo iterations from problem discovery databases:

- unadjusted estimate of *p*
- normalized estimate of *p*
- the sample size
- the known, true value of *p*

I ran the program on a Micron Millennia[1] computer (Windows[2] 95, 64 MB memory) to generate this data from the following problem discovery databases (see Lewis, 2000a, 2000b) for sample sizes ranging from two to six participants:

- MACERR10
- MACERR25
- MACERR50
- MACERR73

These databases were all subsets of the MACERR database (Lewis, 1994; Lewis, Henry, & Mack, 1990), specifically developed to have true discovery rates (*p*) of .10, .25, .50 and .73 respectively. (These databases are available in Lewis, 2000b). Their use ensured the presence of training data for the regression equations with a range of values for true *p*.

The Monte Carlo simulation produced 20,000 cases of data (1,000 cases for each of the combinations of the five sample sizes and four databases). I used SYSTAT[3] Version 5 to create three simple regression models (predicting true *p* with the initial estimate of *p* only, with the normalized estimate of *p* only, and with the sample size only) and three multiple regression models (predicting true *p* with a combination of the initial estimate of *p* and the sample size, the normalized estimate of *p* and the sample size, and both the initial and normalized estimates of *p* and the sample size).

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

Results

Table 4 contains the resulting regression equations, the percentage of variance explained by the regression (R^2), and the overall statistical significance of the regression (osl).

Table 4. Regression equations

Number	Equation	R^2	osl
1	*truep* = -.109 + 1.017**estp*	0.699	0.000
2	*truep* = .16 + .823**normp*	0.785	0.000
3	*truep* = .396 + 0**n*	0.000	1.000
4	*truep* = -.387 + 1.145**estp* + .054**n*	0.786	0.000
5	*truep* = .210 + .829**normp* - .013**n*	0.791	0.000
6	*truep* = -.064 + .520**estp* + .463**normp* + .017**n*	0.799	0.000

In Table 4, *truep* is the true value of *p* as predicted by the equation, *estp* is the unadjusted estimate of *p* from the sample, *normp* is the normalized estimate of *p* from the sample, and *n* is the sample size. All regressions except for Regression 3 (using only *n*) were significant. For all significant regressions, *t*-tests for the elements of the equations (constants and beta weights) were all statistically significant ($p<.0001$). The percentage of explained variance was highest for Regressions 2, 4, 5, and 6. Because previous Monte Carlo studies with this type of data (Lewis, 2000a, 2000d) have shown that the sample size plays an important role when estimating *p*, Regressions 4, 5, and 6 received further evaluation in Experiment 2.

Experiment 2: Evaluation Estimation Accuracy for Regression and Normalization

Purpose

The purpose of this experiment was to assess the accuracy of the three multiple regression models and the normalization procedure regarding their estimation of true *p*.

Method

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

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

The program produced this set of statistics for the unadjusted estimate of *p* and the following adjusted estimates:

- Regression formula 4
- Regression formula 5
- Regression formula 6
- Good-Turing estimation (Lewis, 2000d)
- Normalization

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

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

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

Results

Analysis of Variance for Root Mean Square Error

The root mean square error (RMS error) is the average squared deviation of estimates of *p* from the known true value of *p* in these databases. Thus, the RMS error is an excellent measure of accuracy to use to assess regression and normalization procedures.

I conducted an analysis of variance using RMS error as the dependent variable, and treating databases as subjects in a within-subjects design. The independent variables were sample size (from two to six) and adjustment method (None, Norm, Reg4, Reg5, Reg6, and GT). The analysis indicated significant main effects of sample size ($F(4,12)=84.0$, $p=.00000009$) and adjustment method ($F(5,15)=32.7$, $p=.0000002$), and a significant interaction between these effects ($F(20,60)=20.6$, $p=.000008$). Table 5 and Figure 1 illustrate this interaction.

Table 5. The sample size by adjustment method interaction: Experiment 2

Sample	**None**	**Norm**	**Reg4**	**Reg5**	**Reg6**	**GT**	**Average**
N=2	0.360	0.105	0.180	0.157	0.116	0.113	0.172
N=3	0.239	0.073	0.100	0.131	0.059	0.076	0.113
N=4	0.177	0.063	0.083	0.112	0.062	0.060	0.093
N=5	0.138	0.056	0.087	0.094	0.082	0.048	0.084
N=6	0.112	0.050	0.108	0.079	0.104	0.043	0.082
Average	0.205	0.070	0.111	0.114	0.084	0.068	

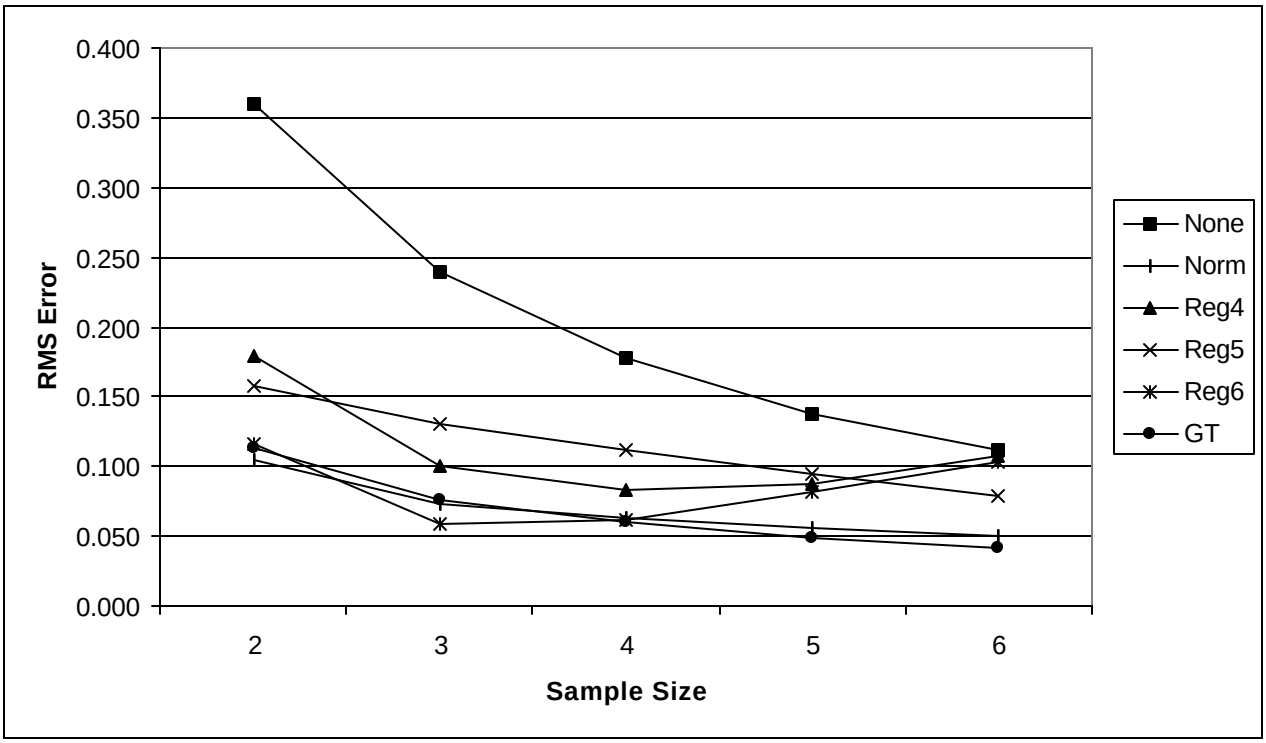

Figure 1. The sample size by adjustment method interaction: Experiment 2

Figure 1 shows that as the sample size increased, accuracy generally increased for all estimation procedures (the main effect of sample size). Overall, the lines for Norm and Good-Turing (GT) seemed to indicate the most accurate performance. I used *t*-tests to compare, at each level of sample size, the significance of difference between no adjustment and each adjustment method (Table 6) and the significance of difference between Good-Turing estimation (which was the best method of those evaluated in Lewis, 2000d) and the other procedures (Table 7). (In the tables, a bold entry for osl indicates $p<.05$ and an italicized entry indicates $p<.10$. All tests in Tables 6 and 7 had 3 degrees of freedom.) Table 6 shows that for sample sizes of 2, 3, and 4, all adjustment methods improved estimation accuracy of *p* relative to no adjustment. At a sample size of 5, the accuracy of Regression 6 was no longer significantly more accurate than no adjustment, and at a sample size of six, both Regressions 4 and 6 (which included the unadjusted estimate of *p* as an element in the equation) failed to be more accurate than the unadjusted estimate.

Table 6. Comparisons of accuracy between adjustment methods and no adjustment

Sample	***Statistic***	**Norm**	**Reg4**	**Reg5**	**Reg6**	**GT**
2	*t*	8.203	29.199	8.210	12.893	32.674
	osl	**0.0038**	**0.0001**	**0.0038**	**0.001**	**0.0001**
3	*t*	8.729	17.705	9.581	9.785	20.039
	osl	**0.0032**	**0.0004**	**0.0024**	**0.0023**	**0.0003**
4	*t*	6.617	11.105	11.446	4.768	9.634
	osl	**0.007**	**0.0016**	**0.0014**	**0.0175**	**0.0024**
5	*t*	5.143	6.433	17.719	2.025	6.175
	osl	**0.0142**	**0.0076**	**0.0004**	0.136	**0.0085**
6	*t*	3.612	0.616	38.682	0.295	4.469
	osl	**0.0364**	0.5816	**0.0000**	0.7872	**0.0209**

As shown in Table 7, Good-Turing estimation was always more accurate than either Regression 4 or 5. Regression 6 was competitive with Good-Turing at sample sizes 2, 3, and 4, but its accuracy seemed to degrade at sample sizes of 5 and 6. When the sample size was 6, Good-Turing was significantly more accurate than Regression 6. Only the normalization procedure was competitive with (not significantly less accurate than) Good-Turing at all five levels of sample size. For this reason, the remaining analyses focus on the effect of adjusting *p* with Good-Turing and normalization procedures.

Table 7. Comparison of Good-Turing and other adjustment procedures

Sample	***Statistic***	**Norm**	**Reg4**	**Reg5**	**Reg6**
2	*t*	0.318	-26.745	-2.488	-0.229
	osl	0.7713	**0.0001**	*0.0887*	0.8333
3	*t*	0.293	-6.905	-8.793	1.636
	osl	0.7888	**0.0062**	**0.0031**	0.2004
4	*t*	-0.621	-2.810	-5.359	-0.123
	osl	0.5785	*0.0673*	**0.0127**	0.9102
5	*t*	-1.344	-3.186	-3.489	-1.790
	osl	0.2714	**0.0499**	**0.0398**	0.1715
6	*t*	-1.139	-4.496	-2.471	-3.301
	osl	0.3376	**0.0205**	*0.09*	**0.0457**

Accuracy of Projected Sample Sizes for Problem Discovery Studies: Effect of Good-Turing and Normalization Procedures

Improved estimation of true *p* should lead to more accurate estimation of required sample sizes for problem discovery usability studies. The following analyses show (Tables 8-11, Figures 2-5), for each database and for estimates based on sample sizes from two to six participants, the difference in projected sample sizes for studies having the goal of uncovering 90% and 95% of the usability problems in a product. 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.

As was the case in Lewis (2000d), application of Good-Turing estimation appeared to improve the accuracy of sample size projection, but because Good-Turing estimation, for three of these four databases (MACERR, MANTEL, SAVE) still resulted in some residual overestimation of *p*, the projected sample sizes tended to slightly underestimate the truly required sample sizes. Interestingly, application of the normalization procedure also improved the accuracy of sample size projection, but because normalization tended to underestimate *p*, projected sample sizes based on the normalized estimate tended to slightly overestimate the truly required sample sizes.

This is clear by examination of the data in Table 12. The cells of Table 12 provide the underestimate of the required sample size for each type of estimation procedure for each combination of sample size and problem discovery database (a negative value in a cell indicates overestimation of the required sample size). Averaging the cells across estimation method appeared to improve the accuracy of the sample size projection.

Table 8. Projected sample sizes for MACERR: Experiment 2

N	Norm2	GT2	Norm3	GT3	Norm4	GT4	Norm5	GT5	Norm6	GT6	TrueP
1	0.130	0.303	0.133	0.238	0.129	0.203	0.125	0.180	0.124	0.165	0.160
2	0.243	0.514	0.248	0.419	0.241	0.365	0.234	0.328	0.233	0.303	0.294
3	0.341	0.661	0.348	0.558	0.339	0.494	0.330	0.449	0.328	0.418	0.407
4	0.427	0.764	0.435	0.663	0.424	0.597	0.414	0.548	0.411	0.514	0.502
5	0.502	0.836	0.510	0.743	0.499	0.678	0.487	0.629	0.484	0.594	0.582
6	0.566	0.885	0.575	0.804	0.563	0.744	0.551	0.696	0.548	0.661	0.649
7	0.623	**0.920**	0.632	0.851	0.620	0.796	0.607	0.751	0.604	0.717	0.705
8	0.672	0.944	0.681	0.886	0.669	0.837	0.656	0.796	0.653	0.764	0.752
9	0.714	***0.961***	0.723	**0.913**	0.711	0.870	0.699	0.832	0.696	0.803	0.792
10	0.752	0.973	0.760	0.934	0.749	0.897	0.737	0.863	0.734	0.835	0.825
11	0.784	0.981	0.792	***0.950***	0.781	**0.918**	0.770	0.887	0.767	0.862	0.853
12	0.812	0.987	0.820	0.962	0.809	0.934	0.799	**0.908**	0.796	0.885	0.877
13	0.836	0.991	0.844	0.971	0.834	0.948	0.824	0.924	0.821	**0.904**	0.896
14	0.858	0.994	0.864	0.978	0.855	***0.958***	0.846	0.938	0.843	0.920	**0.913**
15	0.876	0.996	0.882	0.983	0.874	0.967	0.865	0.949	0.863	0.933	0.927
16	0.892	0.997	0.898	0.987	0.890	0.973	0.882	***0.958***	0.880	0.944	0.939
17	**0.906**	0.998	**0.912**	0.990	**0.904**	0.979	0.897	0.966	0.895	***0.953***	0.948
18	0.918	0.998	0.923	0.992	0.917	0.983	**0.910**	0.972	**0.908**	0.961	***0.957***
19	0.929	0.999	0.934	0.994	0.927	0.987	0.921	0.977	0.919	0.967	0.964
20	0.938	0.999	0.942	0.996	0.937	0.989	0.931	0.981	0.929	0.973	0.969
21	0.946	0.999	***0.950***	0.997	0.945	0.991	0.939	0.985	0.938	0.977	0.974
22	***0.953***	1.000	0.957	0.997	***0.952***	0.993	0.947	0.987	0.946	0.981	0.978
23	0.959	1.000	0.962	0.998	0.958	0.995	***0.954***	0.990	***0.952***	0.984	0.982
24	0.965	1.000	0.967	0.999	0.964	0.996	0.959	0.991	0.958	0.987	0.985
25	0.969	1.000	0.972	0.999	0.968	0.997	0.965	0.993	0.963	0.989	0.987

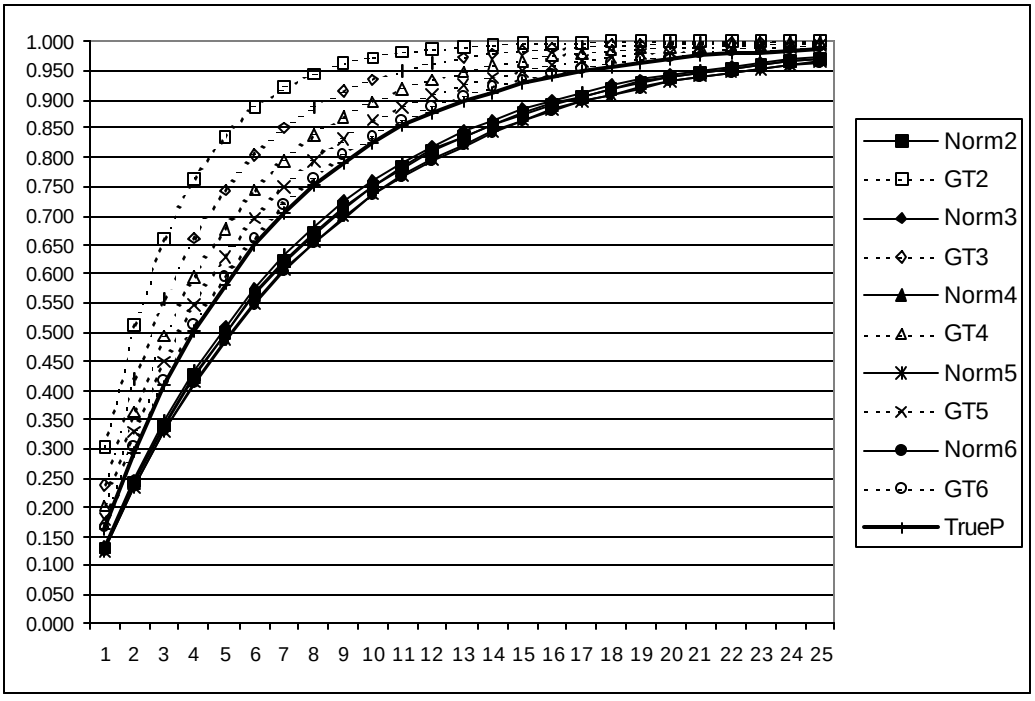

Figure 2. Projected problem discovery curves for MACERR: Experiment 2

Note: Norm = normalized *p*; GT = *p* adjusted with Good-Turing estimation

Table 9. Projected sample sizes for VIRZI90: Experiment 2

N	Norm2	GT2	Norm3	GT3	Norm4	GT4	Norm5	GT5	Norm6	GT6	TrueP
1	0.317	0.394	0.32	0.361	0.313	0.34	0.311	0.331	0.31	0.325	0.359
2	0.534	0.633	0.538	0.592	0.528	0.564	0.525	0.552	0.524	0.544	0.589
3	0.681	0.777	0.686	0.739	0.676	0.713	0.673	0.701	0.671	0.692	0.737
4	0.782	0.865	0.786	0.833	0.777	0.810	0.775	0.800	0.773	0.792	0.831
5	0.851	**0.918**	0.855	0.893	0.847	0.875	0.845	0.866	0.844	0.860	0.892
6	0.898	***0.950***	**0.901**	**0.932**	0.895	**0.917**	0.893	**0.910**	0.892	**0.905**	**0.931**
7	**0.931**	0.970	0.933	***0.956***	**0.928**	0.945	**0.926**	0.940	**0.926**	0.936	***0.956***
8	***0.953***	0.982	***0.954***	0.972	***0.950***	***0.964***	0.949	***0.960***	0.949	***0.957***	0.971
9	0.968	0.989	0.969	0.982	0.966	0.976	***0.965***	0.973	***0.965***	0.971	0.982
10	0.978	0.993	0.979	0.989	0.977	0.984	0.976	0.982	0.976	0.980	0.988
11	0.985	0.996	0.986	0.993	0.984	0.990	0.983	0.988	0.983	0.987	0.992
12	0.990	0.998	0.990	0.995	0.989	0.993	0.989	0.992	0.988	0.991	0.995
13	0.993	0.999	0.993	0.997	0.992	0.995	0.992	0.995	0.992	0.994	0.997
14	0.995	0.999	0.995	0.998	0.995	0.997	0.995	0.996	0.994	0.996	0.998
15	0.997	0.999	0.997	0.999	0.996	0.998	0.996	0.998	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	0.999
17	0.998	1.000	0.999	1.000	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	1.000
19	0.999	1.000	0.999	1.000	0.999	1.000	0.999	1.000	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	1.000

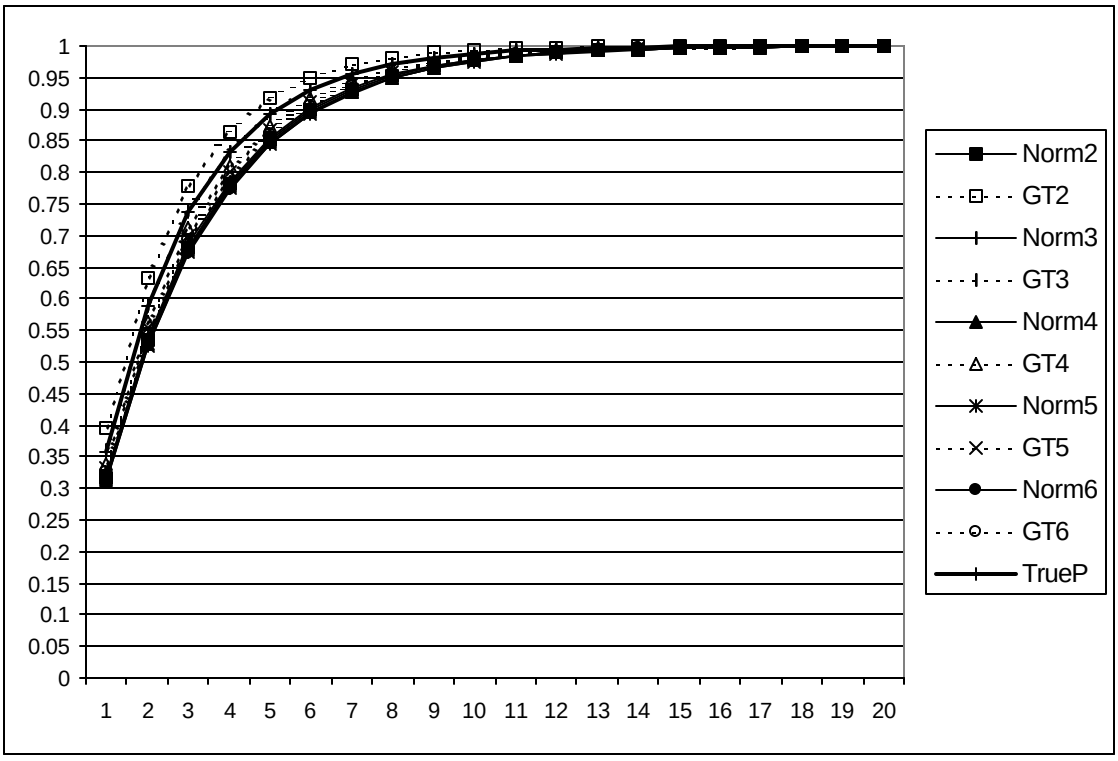

Figure 3. Projected problem discovery curves for VIRZI90: Experiment 2

Note: Norm = normalized *p*; GT = *p* adjusted with Good-Turing estimation

Table 10. Projected sample sizes for MANTEL: Experiment 2

N	Norm2	GT2	Norm3	GT3	Norm4	GT4	Norm5	GT5	Norm6	GT6	TrueP
1	0.451	0.474	0.434	0.446	0.429	0.43	0.421	0.416	0.41	0.399	0.375
2	0.699	0.723	0.680	0.693	0.674	0.675	0.665	0.659	0.652	0.639	0.609
3	0.835	0.854	0.819	0.830	0.814	0.815	0.806	0.801	0.795	0.783	0.756
4	**0.909**	**0.923**	0.897	**0.906**	0.894	0.894	0.888	0.884	0.879	0.870	0.847
5	***0.950***	***0.960***	**0.942**	0.948	**0.939**	**0.940**	**0.935**	**0.932**	**0.929**	**0.922**	**0.905**
6	0.973	0.979	***0.967***	***0.971***	***0.965***	***0.966***	***0.962***	***0.960***	***0.958***	***0.953***	0.940
7	0.985	0.989	0.981	0.984	0.980	0.980	0.978	0.977	0.975	0.972	***0.963***
8	0.992	0.994	0.989	0.991	0.989	0.989	0.987	0.986	0.985	0.983	0.977
9	0.995	0.997	0.994	0.995	0.994	0.994	0.993	0.992	0.991	0.990	0.985
10	0.998	0.998	0.997	0.997	0.996	0.996	0.996	0.995	0.995	0.994	0.991
11	0.999	0.999	0.998	0.998	0.998	0.998	0.998	0.997	0.997	0.996	0.994
12	0.999	1.000	0.999	0.999	0.999	0.999	0.999	0.998	0.998	0.998	0.996
13	1.000	1.000	0.999	1.000	0.999	0.999	0.999	0.999	0.999	0.999	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
15	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	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
18	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
20	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000

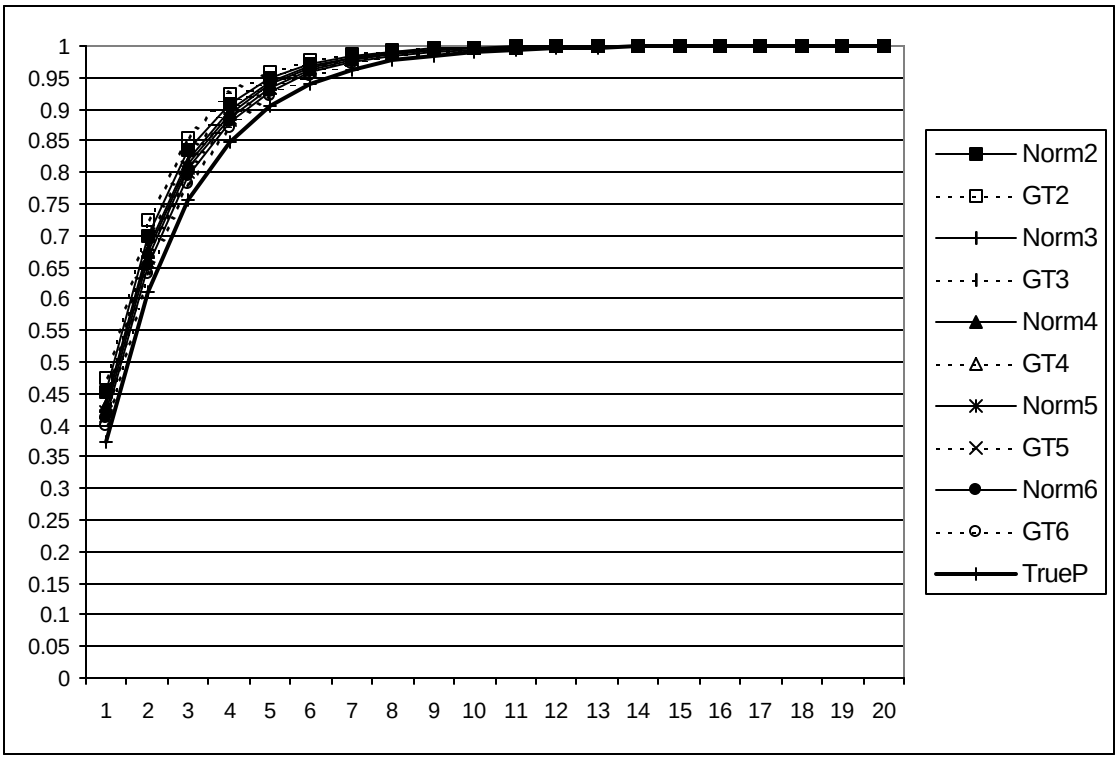

Figure 4. Projected problem discovery curves for MANTEL: Experiment 2

Note: Norm = normalized *p*; GT = *p* adjusted with Good-Turing estimation

Table 11. Projected sample sizes for SAVE: Experiment 2

N	Norm2	GT2	Norm3	GT3	Norm4	GT4	Norm5	GT5	Norm6	GT6	TrueP
1	0.253	0.362	0.254	0.318	0.256	0.298	0.253	0.284	0.257	0.28	0.256
2	0.442	0.593	0.443	0.535	0.446	0.507	0.442	0.487	0.448	0.482	0.446
3	0.583	0.740	0.585	0.683	0.588	0.654	0.583	0.633	0.590	0.627	0.588
4	0.689	0.834	0.690	0.784	0.694	0.757	0.689	0.737	0.695	0.731	0.694
5	0.767	0.894	0.769	0.852	0.772	0.830	0.767	0.812	0.774	0.807	0.772
6	0.826	**0.933**	0.828	0.899	0.830	0.880	0.826	0.865	0.832	0.861	0.830
7	0.870	***0.957***	0.871	**0.931**	0.874	**0.916**	0.870	**0.904**	0.875	**0.900**	0.874
8	**0.903**	0.973	**0.904**	***0.953***	**0.906**	0.941	**0.903**	0.931	**0.907**	0.928	**0.906**
9	0.928	0.982	0.928	0.968	0.930	***0.959***	0.928	***0.951***	0.931	0.948	0.930
10	0.946	0.989	0.947	0.978	0.948	0.971	0.946	0.965	0.949	***0.963***	0.948
11	***0.960***	0.993	***0.960***	0.985	***0.961***	0.980	***0.960***	0.975	***0.962***	0.973	***0.961***
12	0.970	0.995	0.970	0.990	0.971	0.986	0.970	0.982	0.972	0.981	0.971
13	0.977	0.997	0.978	0.993	0.979	0.990	0.977	0.987	0.979	0.986	0.979
14	0.983	0.998	0.983	0.995	0.984	0.993	0.983	0.991	0.984	0.990	0.984
15	0.987	0.999	0.988	0.997	0.988	0.995	0.987	0.993	0.988	0.993	0.988
16	0.991	0.999	0.991	0.998	0.991	0.997	0.991	0.995	0.991	0.995	0.991
17	0.993	1.000	0.993	0.999	0.993	0.998	0.993	0.997	0.994	0.996	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
19	0.996	1.000	0.996	0.999	0.996	0.999	0.996	0.998	0.996	0.998	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

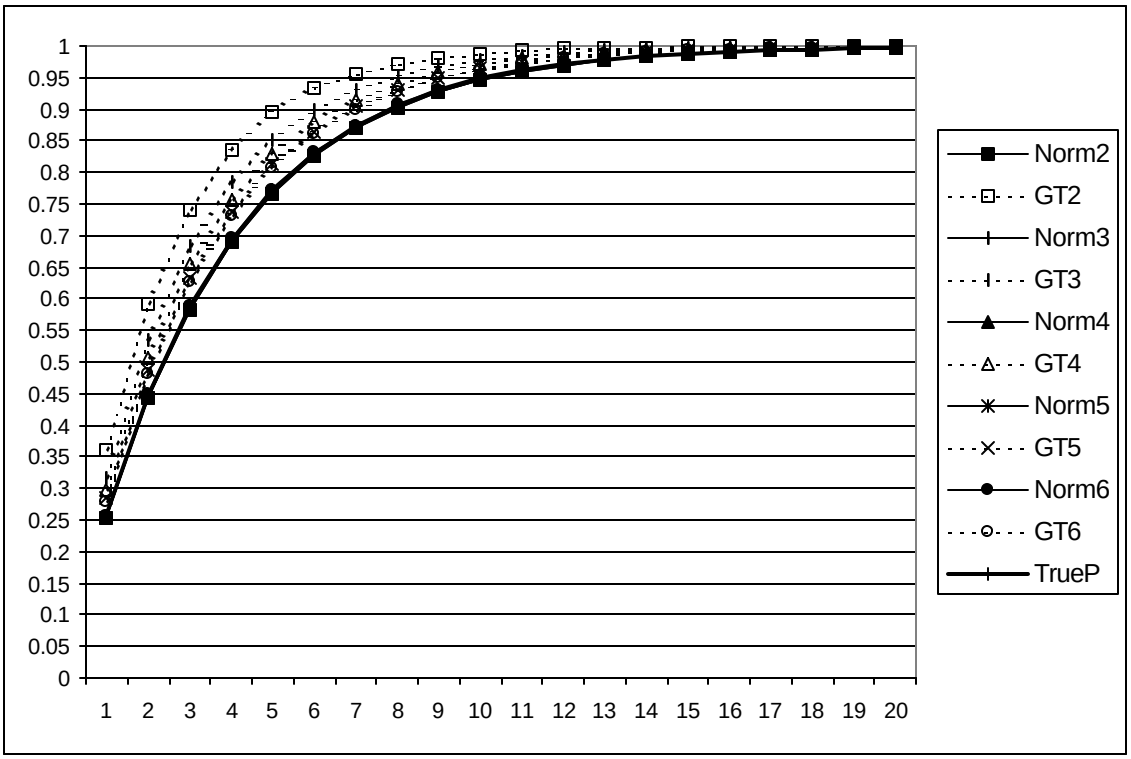

Figure 5. Projected problem discovery curves for SAVE: Experiment 2

Note: Norm = normalized *p*; GT = *p* adjusted with Good-Turing estimation

Table 12. Magnitude of underestimates of required sample sizes: Experiment 2

Database	***N***	**Norm90**	**GT90**	**Norm95**	**GT95**	**Ave90**	**Ave95**
MACERR	*2*	-3	7	-4	9	2.0	2.5
(true p=.16)	*3*	-3	5	-3	7	1.0	2.0
	4	-3	3	-4	4	0.0	0.0
	5	-4	2	-5	2	-1.0	-1.5
	6	-4	1	-5	1	-1.5	-2.0
VIRZI90	***N***	**Norm90**	**GT90**	**Norm95**	**GT95**	**Ave90**	**Ave95**
(true p=.36)	*2*	-1	1	-1	1	0.0	0.0
	3	0	0	-1	0	0.0	-0.5
	4	0	0	-1	-1	0.0	-1.0
	5	-1	0	-2	-1	-0.5	-1.5
	6	-1	0	-2	-1	-0.5	-1.5
MANTEL	***N***	**Norm90**	**GT90**	**Norm95**	**GT95**	**Ave90**	**Ave95**
(true p=.38)	*2*	1	1	2	2	1.0	2.0
	3	0	1	1	1	0.5	1.0
	4	0	0	1	1	0.0	1.0
	5	0	0	1	1	0.0	1.0
	6	0	0	1	1	0.0	1.0
SAVE	***N***	**Norm90**	**GT90**	**Norm95**	**GT95**	**Ave90**	**Ave95**
(true p=.26)	*2*	0	2	0	4	1.0	2.0
	3	0	1	0	3	0.5	1.5
	4	0	1	0	2	0.5	1.0
	5	0	1	0	2	0.5	1.0
	6	0	1	0	1	0.5	0.5
AVERAGE	***N***	**Norm90**	**GT90**	**Norm95**	**GT95**	**Ave90**	**Ave95**
	2	-0.8	2.8	-0.8	4.0	1.0	1.6
	3	-0.8	1.8	-0.8	2.8	0.5	1.0
	4	-0.8	1.0	-1.0	1.5	0.1	0.3
	5	-1.3	0.8	-1.5	1.0	-0.3	-0.3
	6	-1.3	0.5	-1.5	0.5	-0.4	-0.5

Note: The values in the cells are the difference between the truly required sample size and the projected sample size requirement. A positive number indicates underestimation of the truly required sample size, a negative number indicates overestimation. Norm indicates application of the normalization procedure; GT indicates Good-Turing discounting. 90 and 95 indicate the hypothetical problem discovery goals.

Experiment 3: Validation of Improved Accuracy Through Combination of Normalization and Good-Turing Estimation

Purpose

The purpose of this experiment was to assess the improvement in accuracy regarding the estimation of true *p* obtained by combining estimates of *p* calculated with the normalization and Good-Turing methods. This is important because even though the results of Experiment 2 indicated that this should work well, the data in Experiment 2 came from the averaging of average estimates – something that a practitioner conducting a single usability study would not be able to do. To account for this in the current experiment, the combination of estimates took place for each case generated via Monte Carlo simulation.

Method

I adapted the BASIC program used in Experiment 2 to produce the following set of statistics for the unadjusted estimate of *p*, normalized *p*, Good-Turing discounted *p*, and the average of the normalized and discounted estimates of *p*:

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

I ran the program on a Micron Millennia computer (Windows 95, 64 MB memory) to evaluate the following published problem discovery databases for sample sizes ranging from two to six participants (exactly as in Experiment 2):

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

Results

Analysis of Variance for Root Mean Square Error

I conducted an analysis of variance using RMS error as the dependent variable, and treating databases as subjects in a within-subjects design. The independent variables were sample size (from two to six) and adjustment method (None, Norm, GT and the combined Norm/GT average). The analysis indicated significant main effects of sample size ($F(4,12)=94.3$, $p=.00000003$) and adjustment method ($F(3,9)=64.4$, $p=.000002$), and a significant interaction between these effects ($F(12,36)=36.5$, $p=.0000002$). Table 13 and Figure 6 illustrate this interaction.

Table 13. The sample size by adjustment method interaction: Experiment 3

Sample	**None**	**Norm**	**GT**	**Comb**
2	0.360	0.103	0.113	0.092
3	0.239	0.075	0.077	0.067
4	0.177	0.063	0.060	0.055
5	0.138	0.057	0.049	0.048
6	0.113	0.053	0.043	0.046

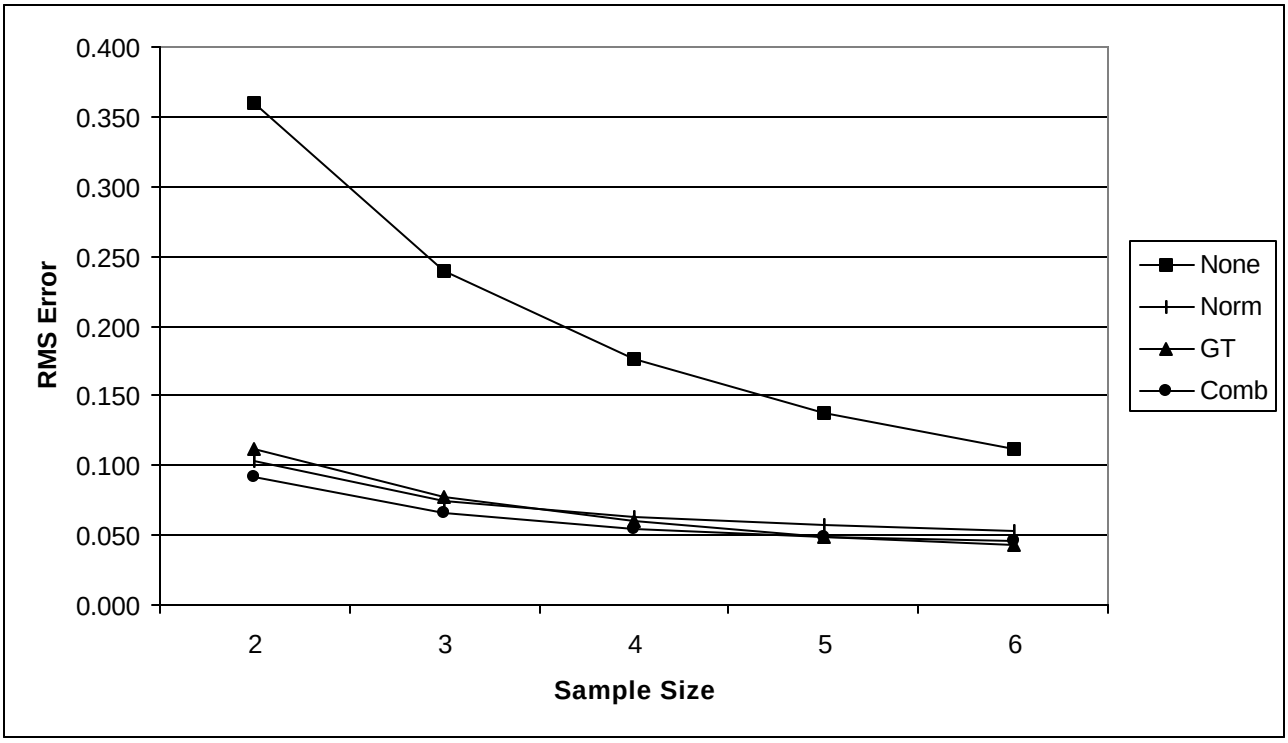

Figure 6. The sample size by adjustment method interaction: Experiment 3

Figure 1 shows that as the sample size increased, accuracy generally increased for all estimation procedures (the main effect of sample size). The lines for Norm, Good-Turing (GT), and their combination (Comb) almost overlaid each other, with the combination having slightly less RMS error than either GT or Norm. The only new line in this graph is the one for the combination of normalization and Good-Turing. A set of *t*-tests showed that estimates based on this combination resulted in significantly lower RMS error than unadjusted estimates for all sample sizes (all $p<.02$). A similar set of t-tests showed that none of the RMS error differences among Good-Turing, normalization, or their combination were significant (all $p>.10$). In this analysis, the source of the significance of the main effect of adjustment type was solely due to the difference between unadjusted and adjusted estimates of *p*.

Accuracy of Projected Sample Sizes for Problem Discovery Studies Using a Combination of Good-Turing and Normalization Procedures

The following analyses show (Tables 14-17, Figures 7-10), for each database and for estimates based on sample sizes from two to six 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 Good-Turing and normalization procedures. 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. The purpose of these analyses was to determine if the combination of Good-Turing and normalization produced, as expected, highly accurate sample size projections for 90% and 95% problem discovery goals.

The data in Table 18 are consistent with this expectation. The cells of Table 18 contain the underestimate of the required sample size for that case. The estimates are quite accurate, with the least accurate estimations for the MACERR database (*truep* = .16) for projections based on estimating *p* from a sample size of two participants. For those cases, the overestimation was 4 and 5 participants for 90% and 95% problem discovery, respectively. Averaged across all four databases, the underestimation of required sample size never exceeded 2.3 and, for estimates of *p* based on at least three participants, never exceeded 1.3.

I conducted an analysis of variance on the required sample size underestimation data for unadjusted *p*, Good-Turing discounting, normalization, and the averaging of Good-Turing and normalization, treating databases as subjects in a within-subjects design with independent variables of sample size, adjustment, and discovery goal (with levels of 90% and 95%). The main effects of sample size ($F(4,12)=5.6$, $p=.009$) and adjustment ($F(3,9)=4.9$, $p=.03$) were significant, as were the discovery goal by adjustment type interaction ($F(3,9)=3.9$, $p=.05$, see Table 19 and Figure 11) and the sample size by adjustment type interaction ($F(12,36)=2.9$, $p=.006$, see Table 20 and Figure 12). In the discovery goal by adjustment type interaction, the required sample size underestimation for 95% was generally greater than for 90%, except for the normalization adjustment type, which had equal underestimation for both levels of discovery goal. The sample size by adjustment type interaction indicated a general decline in the magnitude of underestimation as a function of the sample size used to estimate *p*. This trend seemed strong for estimates based on unadjusted *p*, Good-Turing estimates, and the combination estimate, but not for estimates based on normalized *p*.

As expected, across all sample sizes the Good-Turing estimate tended to underestimate the required sample size and the normalized estimate tended to overestimate the required sample size. For the combination estimate of *p* based on sample sizes of 4, 5, and 6 participants, the estimates of required sample sizes had almost no deviation from the truly required sample sizes. For sample sizes of 2 and 3 participants, the mean underestimation from projecting the combination *p* was 2 and 1 participant(s), respectively.

Table 14. Projected sample sizes for MACERR: Experiment 3

N	None2	Comb2	None3	Comb3	None4	Comb4	None5	Comb5	None6	Comb6	TrueP
1	0.566	0.218	0.421	0.185	0.346	0.165	0.301	0.154	0.269	0.143	0.160
2	0.812	0.388	0.665	0.336	0.572	0.303	0.511	0.284	0.466	0.266	0.294
3	**0.918**	0.522	0.806	0.459	0.720	0.418	0.658	0.395	0.609	0.371	0.407
4	***0.965***	0.626	0.888	0.559	0.817	0.514	0.761	0.488	0.714	0.461	0.502
5	0.985	0.708	**0.935**	0.640	0.880	0.594	0.833	0.567	0.791	0.538	0.582
6	0.993	0.771	***0.962***	0.707	**0.922**	0.661	0.883	0.633	0.847	0.604	0.649
7	0.997	0.821	0.978	0.761	0.949	0.717	**0.918**	0.690	0.888	0.660	0.705
8	0.999	0.860	0.987	0.805	***0.967***	0.764	0.943	0.738	**0.918**	0.709	0.752
9	0.999	0.891	0.993	0.841	0.978	0.803	***0.960***	0.778	0.940	0.751	0.792
10	1.000	**0.914**	0.996	0.871	0.986	0.835	0.972	0.812	***0.956***	0.786	0.825
11	1.000	0.933	0.998	0.895	0.991	0.862	0.981	0.841	0.968	0.817	0.853
12	1.000	0.948	0.999	**0.914**	0.994	0.885	0.986	0.866	0.977	0.843	0.877
13	1.000	***0.959***	0.999	0.930	0.996	**0.904**	0.990	0.886	0.983	0.865	0.896
14	1.000	0.968	1.000	0.943	0.997	0.920	0.993	**0.904**	0.988	0.885	**0.913**
15	1.000	0.975	1.000	***0.954***	0.998	0.933	0.995	0.919	0.991	**0.901**	0.927
16	1.000	0.980	1.000	0.962	0.999	0.944	0.997	0.931	0.993	0.915	0.939
17	1.000	0.985	1.000	0.969	0.999	***0.953***	0.998	0.942	0.995	0.927	0.948
18	1.000	0.988	1.000	0.975	1.000	0.961	0.998	***0.951***	0.996	0.938	***0.957***
19	1.000	0.991	1.000	0.979	1.000	0.967	0.999	0.958	0.997	0.947	0.964
20	1.000	0.993	1.000	0.983	1.000	0.973	0.999	0.965	0.998	***0.954***	0.969

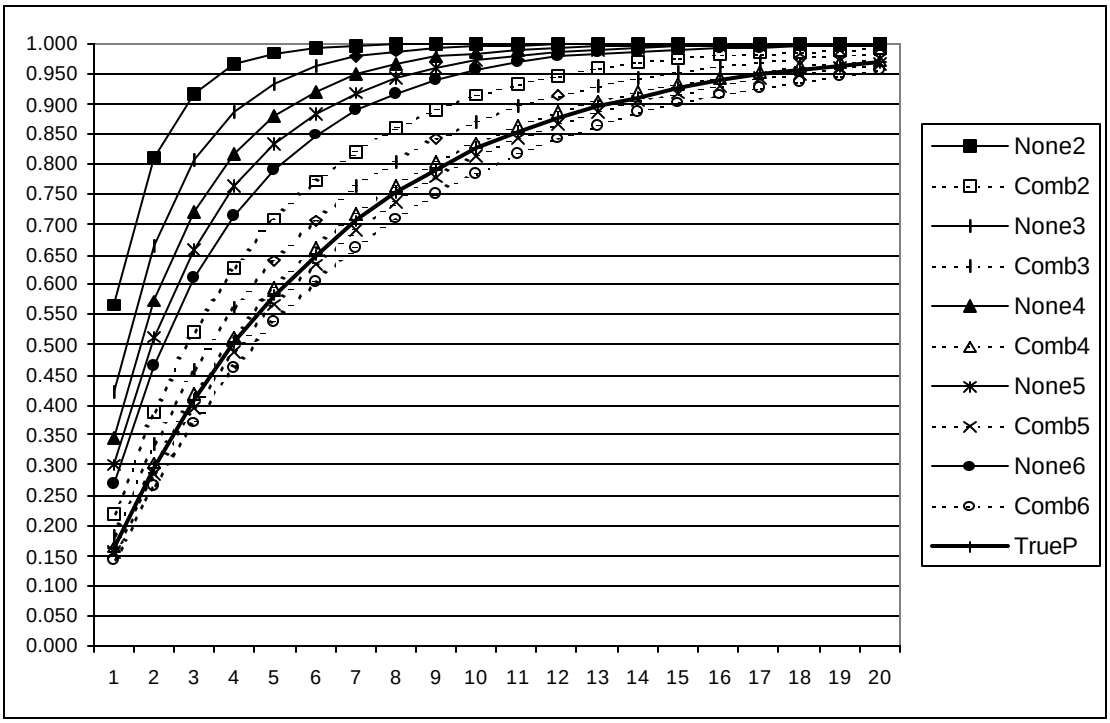

Figure 7. Projected problem discovery curves for MACERR: Experiment 3

Notes: None = unadjusted *p*
Comb = *p* adjusted with combination of Good-Turing estimation and normalization

Table 15. Projected sample sizes for VIRZI90: Experiment 3

N	None2	Comb2	None3	Comb3	None4	Comb4	None5	Comb5	None6	Comb6	TrueP
1	0.660	0.358	0.543	0.336	0.483	0.325	0.447	0.319	0.425	0.318	0.359
2	0.884	0.588	0.791	0.559	0.733	0.544	0.694	0.536	0.669	0.535	0.589
3	***0.961***	0.735	**0.905**	0.707	0.862	0.692	0.831	0.684	0.810	0.683	0.737
4	0.987	0.830	***0.956***	0.806	**0.929**	0.792	**0.906**	0.785	0.891	0.784	0.831
5	0.995	0.891	0.980	0.871	***0.963***	0.860	0.948	0.854	**0.937**	0.852	0.892
6	0.998	**0.930**	0.991	**0.914**	0.981	**0.905**	***0.971***	**0.900**	***0.964***	0.899	**0.931**
7	0.999	***0.955***	0.996	0.943	0.990	0.936	0.984	0.932	0.979	**0.931**	***0.956***
8	1.000	0.971	0.998	***0.962***	0.995	***0.957***	0.991	***0.954***	0.988	***0.953***	0.971
9	1.000	0.981	0.999	0.975	0.997	0.971	0.995	0.968	0.993	0.968	0.982
10	1.000	0.988	1.000	0.983	0.999	0.980	0.997	0.979	0.996	0.978	0.988
11	1.000	0.992	1.000	0.989	0.999	0.987	0.999	0.985	0.998	0.985	0.992
12	1.000	0.995	1.000	0.993	1.000	0.991	0.999	0.990	0.999	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	0.997
14	1.000	0.998	1.000	0.997	1.000	0.996	1.000	0.995	1.000	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	0.999
16	1.000	0.999	1.000	0.999	1.000	0.998	1.000	0.998	1.000	0.998	0.999
17	1.000	0.999	1.000	0.999	1.000	0.999	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
19	1.000	1.000	1.000	1.000	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

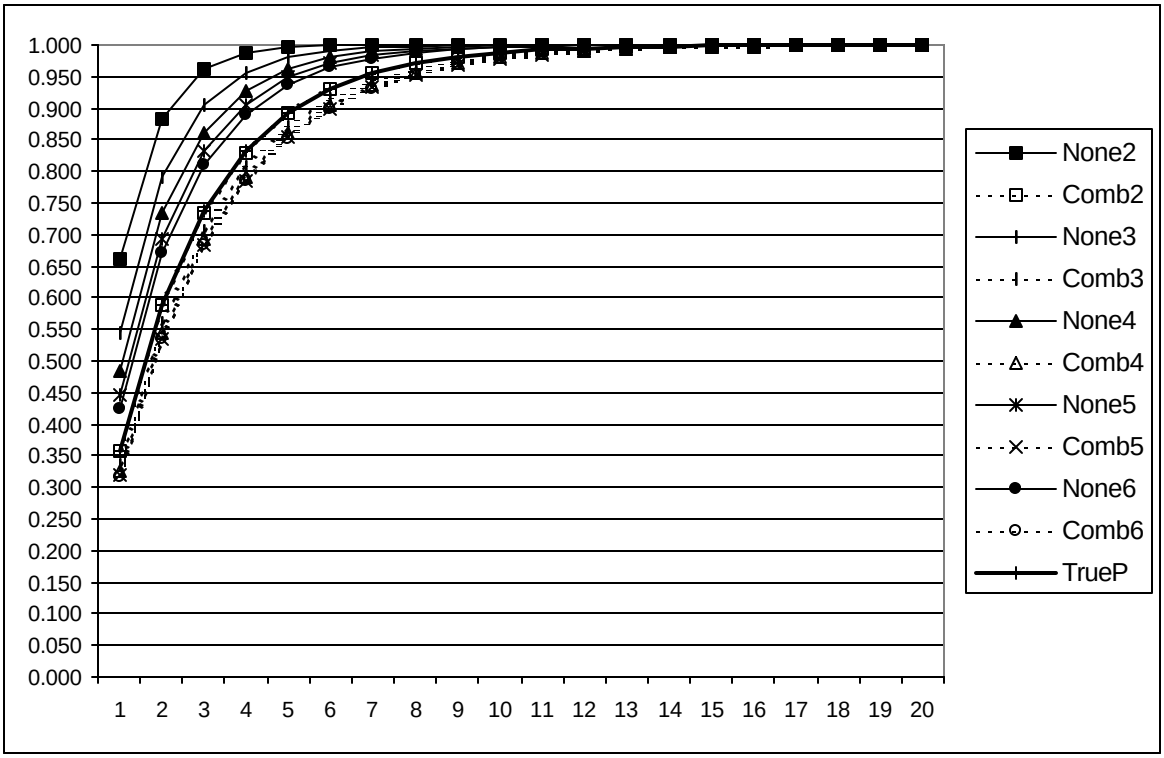

Figure 8. Projected problem discovery curves for VIRZI90: Experiment 3

Notes: None = unadjusted *p*
Comb = *p* adjusted with combination of Good-Turing estimation and normalization

Table 16. Projected sample sizes for MANTEL: Experiment 3

N	None2	Comb2	None3	Comb3	None4	Comb4	None5	Comb5	None6	Comb6	TrueP
1	0.723	0.459	0.626	0.444	0.571	0.429	0.536	0.417	0.514	0.412	0.375
2	**0.923**	0.707	0.860	0.691	0.816	0.674	0.785	0.660	0.764	0.654	0.609
3	***0.979***	0.842	**0.948**	0.828	**0.921**	0.814	**0.900**	0.802	0.885	0.797	0.756
4	0.994	**0.914**	***0.980***	**0.904**	***0.966***	0.894	***0.954***	0.884	**0.944**	0.880	0.847
5	0.998	***0.954***	0.993	0.947	0.985	**0.939**	0.978	**0.933**	***0.973***	**0.930**	**0.905**
6	1.000	0.975	0.997	***0.970***	0.994	***0.965***	0.990	***0.961***	0.987	***0.959***	0.940
7	1.000	0.986	0.999	0.984	0.997	0.980	0.995	0.977	0.994	0.976	***0.963***
8	1.000	0.993	1.000	0.991	0.999	0.989	0.998	0.987	0.997	0.986	0.977
9	1.000	0.996	1.000	0.995	1.000	0.994	0.999	0.992	0.998	0.992	0.985
10	1.000	0.998	1.000	0.997	1.000	0.996	1.000	0.995	0.999	0.995	0.991
11	1.000	0.999	1.000	0.998	1.000	0.998	1.000	0.997	1.000	0.997	0.994
12	1.000	0.999	1.000	0.999	1.000	0.999	1.000	0.998	1.000	0.998	0.996
13	1.000	1.000	1.000	1.000	1.000	0.999	1.000	0.999	1.000	0.999	0.998
14	1.000	1.000	1.000	1.000	1.000	1.000	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	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
17	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
19	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

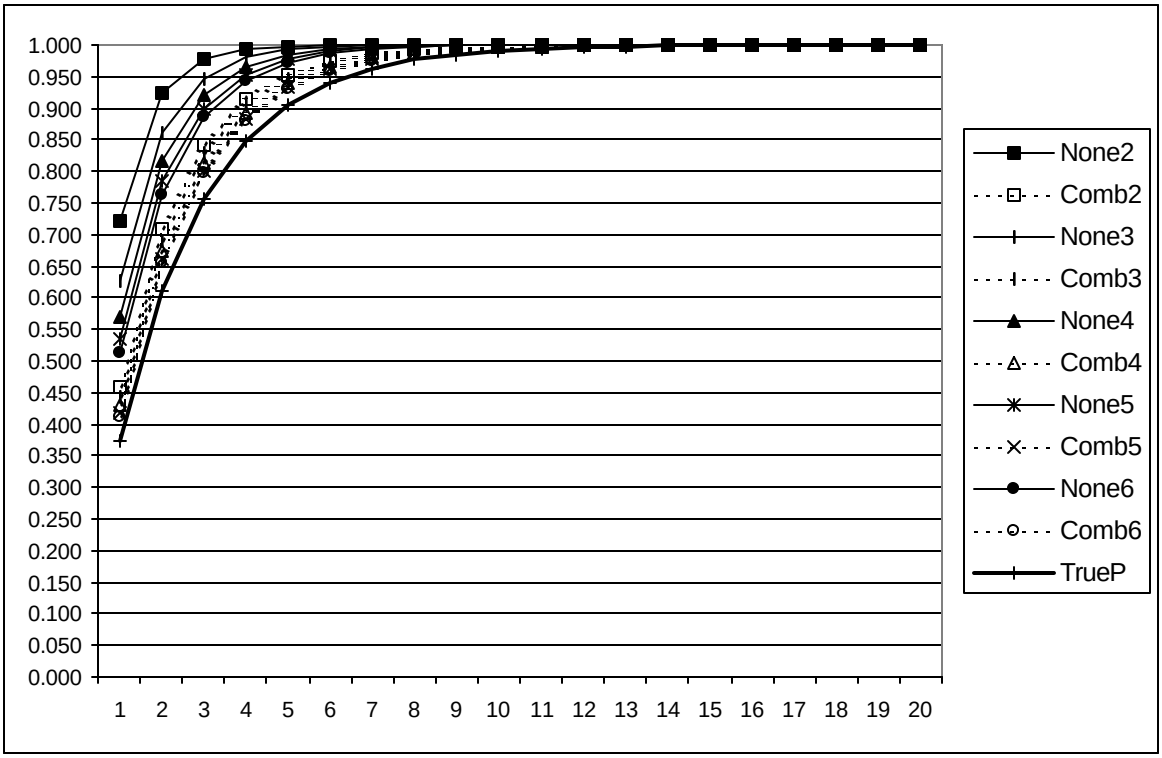

Figure 9. Projected problem discovery curves for MANTEL: Experiment 3

Notes: None = unadjusted *p*
Comb = *p* adjusted with combination of Good-Turing estimation and normalization

Table 17. Projected sample sizes for SAVE: Experiment 3

N	None2	Comb2	None3	Comb3	None4	Comb4	None5	Comb5	None6	Comb6	TrueP
1	0.627	0.308	0.503	0.287	0.442	0.277	0.404	0.271	0.378	0.265	0.256
2	0.861	0.521	0.753	0.492	0.689	0.477	0.645	0.469	0.613	0.460	0.446
3	**0.948**	0.669	0.877	0.638	0.826	0.622	0.788	0.613	0.759	0.603	0.588
4	***0.981***	0.771	**0.939**	0.742	**0.903**	0.727	0.874	0.718	0.850	0.708	0.694
5	0.993	0.841	***0.970***	0.816	0.946	0.802	**0.925**	0.794	**0.907**	0.785	0.772
6	0.997	0.890	0.985	0.869	***0.970***	0.857	***0.955***	0.850	0.942	0.842	0.830
7	0.999	**0.924**	0.993	**0.906**	0.983	0.897	0.973	0.891	***0.964***	0.884	0.874
8	1.000	0.947	0.996	0.933	0.991	**0.925**	0.984	**0.920**	0.978	**0.915**	**0.906**
9	1.000	***0.964***	0.998	***0.952***	0.995	0.946	0.991	0.942	0.986	0.937	0.930
10	1.000	0.975	0.999	0.966	0.997	***0.961***	0.994	***0.958***	0.991	***0.954***	0.948
11	1.000	0.983	1.000	0.976	0.998	0.972	0.997	0.969	0.995	0.966	***0.961***
12	1.000	0.988	1.000	0.983	0.999	0.980	0.998	0.977	0.997	0.975	0.971
13	1.000	0.992	1.000	0.988	0.999	0.985	0.999	0.984	0.998	0.982	0.979
14	1.000	0.994	1.000	0.991	1.000	0.989	0.999	0.988	0.999	0.987	0.984
15	1.000	0.996	1.000	0.994	1.000	0.992	1.000	0.991	0.999	0.990	0.988
16	1.000	0.997	1.000	0.996	1.000	0.994	1.000	0.994	0.999	0.993	0.991
17	1.000	0.998	1.000	0.997	1.000	0.996	1.000	0.995	1.000	0.995	0.993
18	1.000	0.999	1.000	0.998	1.000	0.997	1.000	0.997	1.000	0.996	0.995
19	1.000	0.999	1.000	0.998	1.000	0.998	1.000	0.998	1.000	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	0.997

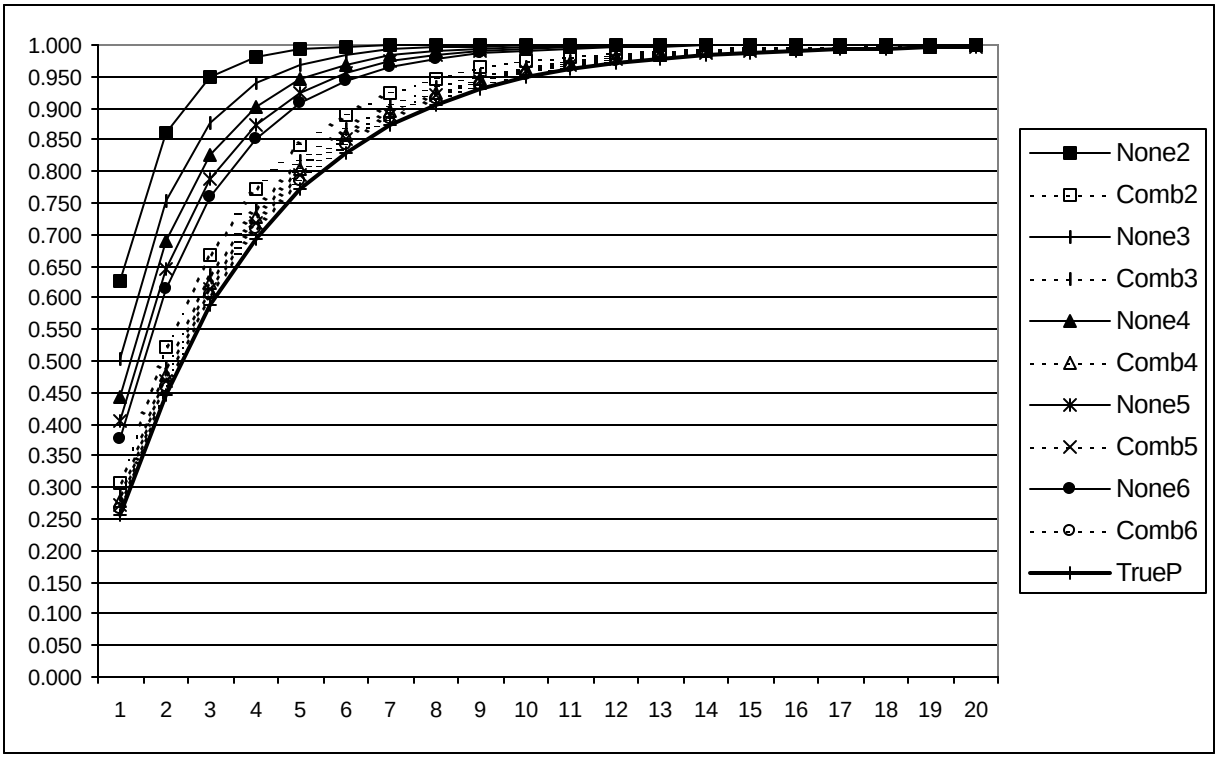

Figure 10. Projected problem discovery curves for SAVE: Experiment 3

Notes: None = unadjusted *p*
Comb = *p* adjusted with combination of Good-Turing estimation and normalization

Table 18. Magnitude of underestimates of required sample sizes: Experiment 3

Database	***N***	**None90**	**Comb90**	**None95**	**Comb95**
MACERR	*2*	11	4	14	5
(true p=.16)	*3*	9	2	12	3
	4	8	1	10	1
	5	7	0	9	0
	6	6	-1	8	-2
VIRZI90	***N***	**None90**	**Comb90**	**None95**	**Comb95**
(true p=.36)	*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
MANTEL	***N***	**None90**	**Comb90**	**None95**	**Comb95**
(true p=.38)	*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
SAVE	***N***	**None90**	**Comb90**	**None95**	**Comb95**
(true p=.26)	*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
AVERAGE	***N***	**None90**	**Comb90**	**None95**	**Comb95**
	2	5.5	1.5	7.3	2.3
	3	4.5	1.0	6.0	1.3
	4	4.0	0.3	5.0	0.5
	5	3.5	0.0	4.5	0.3
	6	2.8	-0.5	3.8	-0.3

Note: The values in the cells are the difference between the truly required sample size and the projected sample size requirement. A positive number indicates underestimation of the truly required sample size, a negative number indicates overestimation. Norm indicates application of the normalization procedure; GT indicates Good-Turing discounting. 90 and 95 indicate the hypothetical problem discovery goals.

Table 19. Underestimation of n as a function of discovery goal and adjustment type

Goal	None	GT	Norm	Comb
95 Percent	5.30	1.95	-1.10	0.80
90 Percent	4.05	1.35	-0.95	0.45

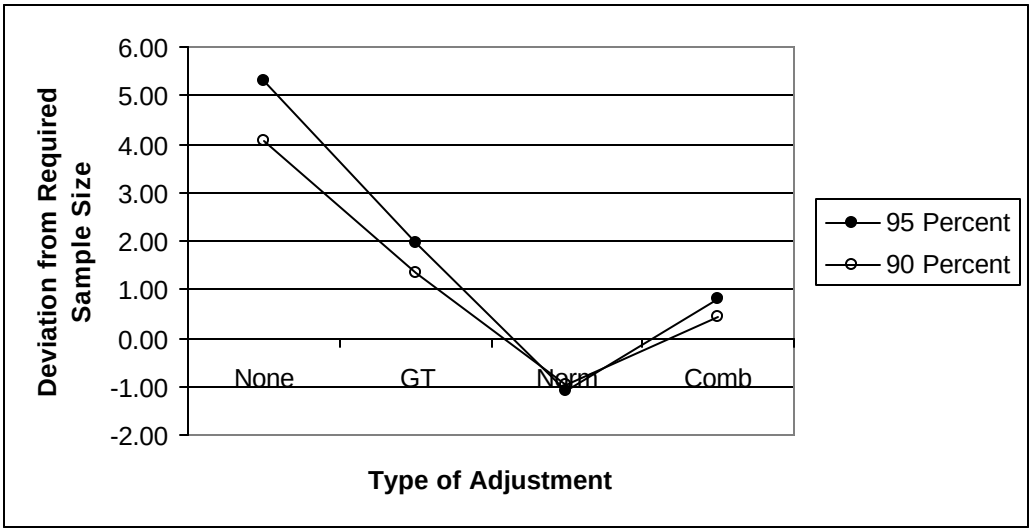

Figure 11. Discovery goal by adjustment type interaction

Table 20. Underestimation of n as a function of sample size and adjustment type

Sample	None	GT	Norm	Comb
N=2	6.38	3.38	-0.75	1.88
N=3	5.25	2.25	-0.75	1.13
N=4	4.50	1.25	-0.88	0.38
N=5	4.00	0.88	-1.38	0.13
N=6	3.25	0.50	-1.38	-0.38

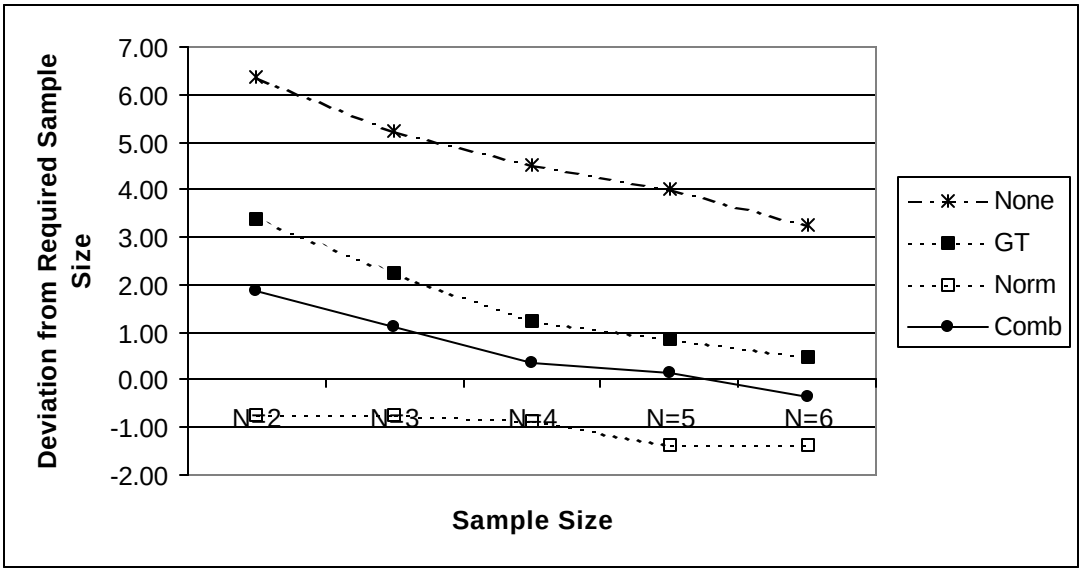

Figure 12. Sample size by adjustment type interaction

Distributions of Normalized and Combination-Adjusted *p*

The preceding analyses have focused on means of various distributions. The analyses in this section address the broader distribution of *p*, both normalized and adjusted with the normalization/Good-Turing combination procedure. These analyses will help practitioners understand the variability of these distributions and to take this variability into account when planning problem-discovery usability studies.

The cells in Table 21 are the averages collapsed across databases and sample sizes. The cells in Table 22 are the averages across databases, still organized by sample size. Tables 23-26 provide this information by each database for sample sizes from two to six participants. Tables 21 and 22 contain overestimation ratios and deviations from true *p* for estimates of *p* adjusted by normalization (NormOvr and NormDev respectively) and by the normalization/Good-Turing combination procedure (CombOvr and CombDev respectively). Tables 23-26 contain these same indicators of accuracy plus the values for *p* as adjusted with these procedures. All tables provide key percentiles (1st, 5th, 10th, 25th, 50th, 75th, 90th, 95th, and 99th) for the distribution of these measures. The cells for the median row (50th percentile) are bold. The cells for the 25th and 75th percentiles (which define the interquartile range – the boundaries that contain the central 50% of the distribution) are bold italic. The cells for the 5th and 95th percentiles (the boundaries that contain the central 90% of the distribution) are in italics.

The data in Table 21 show that, in general, the variability of these measures is not too extreme. The interquartile range for the overestimation ratio for normalized *p* is 0.81 to 1.09 and for combination-adjusted *p* is 0.93 to 1.17. The corresponding deviations from true *p* are –.05 to .03 and -.02 to .05. The overestimation ratio boundaries that contain the central 90% of the distribution range from 0.63 to 1.29 for normalized *p* and range from 0.79 to 1.35 for combination-adjusted *p*. The corresponding deviations from true *p* are -.10 to .09 and -.06 to .10. The data in Table 22 show that these patterns are similar across the various sample sizes used to estimate *p*, with the intervals being wider at the smaller sample sizes and narrower at the larger sample sizes. For each measure, the interval width for a sample size of six participants is about half the interval width for a sample size of two participants. The interval width data as a function of sample size appears in Table 27.

Table 21. Distribution of normalized and combination-adjusted p: Grand averages

Percentile	**NormOvr**	**CombOvr**	**NormDev**	**CombDev**
1st	0.51	0.69	-0.13	-0.09
5th	*0.63*	*0.79*	*-0.10*	*-0.06*
10th	0.69	0.84	-0.08	-0.05
25th	***0.81***	***0.93***	***-0.05***	***-0.02***
50th	**0.94**	**1.05**	**-0.01**	**0.01**
75th	***1.09***	***1.17***	***0.03***	***0.05***
90th	1.21	1.28	0.07	0.08
95th	*1.29*	*1.35*	*0.09*	*0.10*
99th	1.44	1.50	0.13	0.14

Table 22. Distribution of normalized and combination-adjusted p: Averages

Case	***Statistic***	**NormOvr**	**CombOvr**	**NormDev**	**CombDev**
Average	1st %ile	0.25	0.66	-0.20	-0.10
n=2	*5th %ile*	*0.46*	*0.81*	*-0.15*	*-0.06*
	10th %ile	0.54	0.87	-0.12	-0.04
	25th %ile	***0.72***	***1.00***	***-0.07***	***0.00***
	50th %ile	**0.96**	**1.17**	**0.00**	**0.04**
	75th %ile	***1.20***	***1.36***	***0.06***	***0.10***
	90th %ile	1.40	1.52	0.12	0.14
	95th %ile	*1.52*	*1.62*	*0.15*	*0.17*
	99th %ile	1.85	1.89	0.24	0.24
Average	1st %ile	0.47	0.69	-0.14	-0.09
n=3	*5th %ile*	*0.59*	*0.78*	*-0.11*	*-0.06*
	10th %ile	0.68	0.85	-0.08	-0.04
	25th %ile	***0.80***	***0.96***	***-0.05***	***-0.01***
	50th %ile	**0.95**	**1.08**	**-0.01**	**0.02**
	75th %ile	***1.12***	***1.22***	***0.04***	***0.06***
	90th %ile	1.26	1.35	0.08	0.10
	95th %ile	*1.34*	*1.43*	*0.10*	*0.12*
	99th %ile	1.49	1.56	0.14	0.16
Average	1st %ile	0.56	0.69	-0.12	-0.09
n=4	*5th %ile*	*0.66*	*0.78*	*-0.09*	*-0.06*
	10th %ile	0.72	0.83	-0.07	-0.05
	25th %ile	***0.83***	***0.92***	***-0.04***	***-0.02***
	50th %ile	**0.95**	**1.03**	**-0.01**	**0.01**
	75th %ile	***1.06***	***1.14***	***0.02***	***0.04***
	90th %ile	1.18	1.25	0.06	0.07
	95th %ile	*1.24*	*1.31*	*0.08*	*0.09*
	99th %ile	1.36	1.41	0.11	0.12
Average	1st %ile	0.60	0.69	-0.11	-0.09
n=5	*5th %ile*	*0.71*	*0.78*	*-0.08*	*-0.06*
	10th %ile	0.75	0.83	-0.06	-0.05
	25th %ile	***0.84***	***0.91***	***-0.04***	***-0.03***
	50th %ile	**0.93**	**0.99**	**-0.01**	**0.00**
	75th %ile	***1.04***	***1.09***	***0.02***	***0.03***
	90th %ile	1.13	1.18	0.04	0.05
	95th %ile	*1.19*	*1.23*	*0.06*	*0.07*
	99th %ile	1.28	1.33	0.09	0.10
Average	1st %ile	0.65	0.71	-0.09	-0.08
n=6	*5th %ile*	*0.73*	*0.79*	*-0.07*	*-0.06*
	10th %ile	0.77	0.82	-0.06	-0.05
	25th %ile	***0.84***	***0.89***	***-0.04***	***-0.03***
	50th %ile	**0.92**	**0.97**	**-0.01**	**-0.01**
	75th %ile	***1.01***	***1.05***	***0.01***	***0.02***
	90th %ile	1.09	1.13	0.03	0.04
	95th %ile	*1.14*	*1.18*	*0.05*	*0.06*
	99th %ile	1.23	1.29	0.08	0.09

Table 23. Distribution of normalized and combination-adjusted p: MACERR

Case	***Statistic***	**Norm-*p***	**Comb-*p***	**NormOvr**	**CombOvr**	**NormDev**	**CombDev**
MACERR	1st %ile	0.000	0.125	0.00	0.77	-0.163	-0.038
n=2	*5th %ile*	*0.050*	*0.160*	*0.31*	*0.98*	*-0.113*	*-0.003*
truep=.16	10th %ile	0.056	0.163	0.34	1.00	-0.107	0.000
	25th %ile	***0.077***	***0.178***	***0.47***	***1.09***	***-0.086***	***0.015***
	50th %ile	**0.129**	**0.215**	**0.79**	**1.32**	**-0.034**	**0.052**
	75th %ile	***0.178***	***0.250***	***1.09***	***1.53***	***0.015***	***0.087***
	90th %ile	0.214	0.277	1.31	1.70	0.051	0.114
	95th %ile	*0.226*	*0.286*	*1.39*	*1.75*	*0.063*	*0.123*
	99th %ile	0.293	0.336	1.80	2.06	0.130	0.173
MACERR	1st %ile	0.043	0.116	0.26	0.71	-0.120	-0.047
n=3	*5th %ile*	*0.061*	*0.128*	*0.37*	*0.79*	*-0.102*	*-0.035*
truep=.16	10th %ile	0.080	0.144	0.49	0.88	-0.083	-0.019
	25th %ile	***0.105***	***0.163***	***0.64***	***1.00***	***-0.058***	***0.000***
	50th %ile	**0.130**	**0.184**	**0.80**	**1.13**	**-0.033**	**0.021**
	75th %ile	***0.162***	***0.209***	***0.99***	***1.28***	***-0.001***	***0.046***
	90th %ile	0.183	0.227	1.12	1.39	0.020	0.064
	95th %ile	*0.194*	*0.236*	*1.19*	*1.45*	*0.031*	*0.073*
	99th %ile	0.213	0.247	1.31	1.52	0.050	0.084
MACERR	1st %ile	0.063	0.112	0.39	0.69	-0.100	-0.051
n=4	*5th %ile*	*0.078*	*0.125*	*0.48*	*0.77*	*-0.085*	*-0.038*
truep=.16	10th %ile	0.090	0.133	0.55	0.82	-0.073	-0.030
	25th %ile	***0.108***	***0.148***	***0.66***	***0.91***	***-0.055***	***-0.015***
	50th %ile	**0.128**	**0.165**	**0.79**	**1.01**	**-0.035**	**0.002**
	75th %ile	***0.148***	***0.181***	***0.91***	***1.11***	***-0.015***	***0.018***
	90th %ile	0.167	0.198	1.02	1.21	0.004	0.035
	95th %ile	*0.177*	*0.206*	*1.09*	*1.26*	*0.014*	*0.043*
	99th %ile	0.196	0.219	1.20	1.34	0.033	0.056
MACERR	1st %ile	0.069	0.107	0.42	0.66	-0.094	-0.056
n=5	*5th %ile*	*0.089*	*0.121*	*0.55*	*0.74*	*-0.074*	*-0.042*
truep=.16	10th %ile	0.096	0.128	0.59	0.79	-0.067	-0.035
	25th %ile	***0.109***	***0.139***	***0.67***	***0.85***	***-0.054***	***-0.024***
	50th %ile	**0.127**	**0.153**	**0.78**	**0.94**	**-0.036**	**-0.010**
	75th %ile	***0.144***	***0.168***	***0.88***	***1.03***	***-0.019***	***0.005***
	90th %ile	0.158	0.181	0.97	1.11	-0.005	0.018
	95th %ile	*0.167*	*0.188*	*1.02*	*1.15*	*0.004*	*0.025*
	99th %ile	0.178	0.197	1.09	1.21	0.015	0.034
MACERR	1st %ile	0.084	0.110	0.52	0.67	-0.079	-0.053
n=6	*5th %ile*	*0.094*	*0.119*	*0.58*	*0.73*	*-0.069*	*-0.044*
truep=.16	10th %ile	0.100	0.123	0.61	0.75	-0.063	-0.040
	25th %ile	***0.110***	***0.133***	***0.67***	***0.82***	***-0.053***	***-0.030***
	50th %ile	**0.122**	**0.142**	**0.75**	**0.87**	**-0.041**	**-0.021**
	75th %ile	***0.133***	***0.153***	***0.82***	***0.94***	***-0.030***	***-0.010***
	90th %ile	0.145	0.163	0.89	1.00	-0.018	0.000
	95th %ile	*0.152*	*0.169*	*0.93*	*1.04*	*-0.011*	*0.006*
	99th %ile	0.164	0.179	1.01	1.10	0.001	0.016

Table 24. Distribution of normalized and combination-adjusted p: VIRZI90

Case	Statistic	Norm-*p*	Comb-*p*	NormOvr	CombOvr	NormDev	CombDev
VIRZI90	1st %ile	0.125	0.213	0.35	0.59	-0.234	-0.146
n=2	*5th %ile*	*0.160*	*0.238*	*0.45*	*0.66*	*-0.199*	*-0.121*
truep=.36	10th %ile	0.185	0.256	0.52	0.71	-0.174	-0.103
	25th %ile	***0.263***	***0.313***	***0.73***	***0.87***	***-0.096***	***-0.046***
	50th %ile	**0.318**	**0.355**	**0.89**	**0.99**	**-0.041**	**-0.004**
	75th %ile	***0.385***	***0.407***	***1.07***	***1.13***	***0.026***	***0.048***
	90th %ile	0.435	0.447	1.21	1.25	0.076	0.088
	95th %ile	*0.474*	*0.478*	*1.32*	*1.33*	*0.115*	*0.119*
	99th %ile	0.478	0.482	1.33	1.34	0.119	0.123
VIRZI90	1st %ile	0.196	0.237	0.55	0.66	-0.163	-0.122
n=3	*5th %ile*	*0.222*	*0.256*	*0.62*	*0.71*	*-0.137*	*-0.103*
truep=.36	10th %ile	0.250	0.280	0.70	0.78	-0.109	-0.079
	25th %ile	***0.271***	***0.299***	***0.75***	***0.83***	***-0.088***	***-0.060***
	50th %ile	**0.310**	**0.329**	**0.86**	**0.92**	**-0.049**	**-0.030**
	75th %ile	***0.352***	***0.369***	***0.98***	***1.03***	***-0.007***	***0.010***
	90th %ile	0.396	0.408	1.10	1.14	0.037	0.049
	95th %ile	*0.417*	*0.426*	*1.16*	*1.19*	*0.058*	*0.067*
	99th %ile	0.457	0.461	1.27	1.28	0.098	0.102
VIRZI90	1st %ile	0.215	0.238	0.60	0.66	-0.144	-0.121
n=4	*5th %ile*	*0.237*	*0.257*	*0.66*	*0.72*	*-0.122*	*-0.102*
truep=.36	10th %ile	0.256	0.274	0.71	0.76	-0.103	-0.085
	25th %ile	***0.281***	***0.295***	***0.78***	***0.82***	***-0.078***	***-0.064***
	50th %ile	**0.310**	**0.322**	**0.86**	**0.90**	**-0.049**	**-0.037**
	75th %ile	***0.333***	***0.352***	***0.93***	***0.98***	***-0.026***	***-0.007***
	90th %ile	0.370	0.383	1.03	1.07	0.011	0.024
	95th %ile	*0.389*	*0.401*	*1.08*	*1.12*	*0.030*	*0.042*
	99th %ile	0.420	0.430	1.17	1.20	0.061	0.071
VIRZI90	1st %ile	0.227	0.245	0.63	0.68	-0.132	-0.114
n=5	*5th %ile*	*0.257*	*0.265*	*0.72*	*0.74*	*-0.102*	*-0.094*
truep=.36	10th %ile	0.265	0.275	0.74	0.77	-0.094	-0.084
	25th %ile	***0.283***	***0.293***	***0.79***	***0.82***	***-0.076***	***-0.066***
	50th %ile	**0.307**	**0.317**	**0.86**	**0.88**	**-0.052**	**-0.042**
	75th %ile	***0.333***	***0.342***	***0.93***	***0.95***	***-0.026***	***-0.017***
	90th %ile	0.353	0.364	0.98	1.01	-0.006	0.005
	95th %ile	*0.370*	*0.380*	*1.03*	*1.06*	*0.011*	*0.021*
	99th %ile	0.394	0.410	1.10	1.14	0.035	0.051
VIRZI90	1st %ile	0.242	0.250	0.67	0.70	-0.117	-0.109
n=6	*5th %ile*	*0.261*	*0.267*	*0.73*	*0.74*	*-0.098*	*-0.092*
truep=.36	10th %ile	0.271	0.277	0.75	0.77	-0.088	-0.082
	25th %ile	***0.288***	***0.295***	***0.80***	***0.82***	***-0.071***	***-0.064***
	50th %ile	**0.309**	**0.317**	**0.86**	**0.88**	**-0.050**	**-0.042**
	75th %ile	***0.331***	***0.338***	***0.92***	***0.94***	***-0.028***	***-0.021***
	90th %ile	0.352	0.359	0.98	1.00	-0.007	0.000
	95th %ile	*0.364*	*0.374*	*1.01*	*1.04*	*0.005*	*0.015*
	99th %ile	0.394	0.402	1.10	1.12	0.035	0.043

Table 25. Distribution of normalized and combination-adjusted p: MANTEL

Case	***Statistic***	**Norm-*p***	**Comb-*p***	**NormOvr**	**CombOvr**	**NormDev**	**CombDev**
MANTEL	1st %ile	0.167	0.242	0.45	0.65	-0.208	-0.133
n=2	*5th %ile*	*0.250*	*0.304*	*0.67*	*0.81*	*-0.125*	*-0.071*
truep=.38	10th %ile	0.294	0.337	0.78	0.90	-0.081	-0.038
	25th %ile	***0.357***	***0.385***	***0.95***	***1.03***	***-0.018***	***0.010***
	50th %ile	**0.438**	**0.449**	**1.17**	**1.20**	**0.063**	**0.074**
	75th %ile	***0.529***	***0.525***	***1.41***	***1.40***	***0.154***	***0.150***
	90th %ile	0.600	0.586	1.60	1.56	0.225	0.211
	95th %ile	*0.647*	*0.628*	*1.73*	*1.67*	*0.272*	*0.253*
	99th %ile	0.750	0.725	2.00	1.93	0.375	0.350
MANTEL	1st %ile	0.233	0.263	0.62	0.70	-0.142	-0.112
n=3	*5th %ile*	*0.294*	*0.314*	*0.78*	*0.84*	*-0.081*	*-0.061*
truep=.38	10th %ile	0.325	0.342	0.87	0.91	-0.050	-0.033
	25th %ile	***0.382***	***0.391***	***1.02***	***1.04***	***0.007***	***0.016***
	50th %ile	**0.438**	**0.443**	**1.17**	**1.18**	**0.063**	**0.068**
	75th %ile	***0.500***	***0.496***	***1.33***	***1.32***	***0.125***	***0.121***
	90th %ile	0.553	0.544	1.47	1.45	0.178	0.169
	95th %ile	*0.583*	*0.580*	*1.55*	*1.55*	*0.208*	*0.205*
	99th %ile	0.619	0.629	1.65	1.68	0.244	0.254
MANTEL	1st %ile	0.275	0.284	0.73	0.76	-0.100	-0.091
n=4	*5th %ile*	*0.316*	*0.323*	*0.84*	*0.86*	*-0.059*	*-0.052*
truep=.38	10th %ile	0.333	0.344	0.89	0.92	-0.042	-0.031
	25th %ile	***0.379***	***0.381***	***1.01***	***1.02***	***0.004***	***0.006***
	50th %ile	**0.429**	**0.430**	**1.14**	**1.15**	**0.054**	**0.055**
	75th %ile	***0.476***	***0.475***	***1.27***	***1.27***	***0.101***	***0.100***
	90th %ile	0.515	0.513	1.37	1.37	0.140	0.138
	95th %ile	*0.533*	*0.535*	*1.42*	*1.43*	*0.158*	*0.160*
	99th %ile	0.583	0.579	1.55	1.54	0.208	0.204
MANTEL	1st %ile	0.274	0.274	0.73	0.73	-0.101	-0.101
n=5+A102	*5th %ile*	*0.321*	*0.322*	*0.86*	*0.86*	*-0.054*	*-0.053*
truep=.38	10th %ile	0.344	0.344	0.92	0.92	-0.031	-0.031
	25th %ile	***0.381***	***0.380***	***1.02***	***1.01***	***0.006***	***0.005***
	50th %ile	**0.417**	**0.414**	**1.11**	**1.10**	**0.042**	**0.039**
	75th %ile	***0.460***	***0.456***	***1.23***	***1.22***	***0.085***	***0.081***
	90th %ile	0.500	0.495	1.33	1.32	0.125	0.120
	95th %ile	*0.523*	*0.516*	*1.39*	*1.38*	*0.148*	*0.141*
	99th %ile	0.563	0.562	1.50	1.50	0.188	0.187
MANTEL	1st %ile	0.288	0.286	0.77	0.76	-0.087	-0.089
n=6	*5th %ile*	*0.330*	*0.327*	*0.88*	*0.87*	*-0.045*	*-0.048*
truep=.38	10th %ile	0.348	0.343	0.93	0.91	-0.027	-0.032
	25th %ile	***0.380***	***0.373***	***1.01***	***0.99***	***0.005***	***-0.002***
	50th %ile	**0.417**	**0.410**	**1.11**	**1.09**	**0.042**	**0.035**
	75th %ile	***0.453***	***0.449***	***1.21***	***1.20***	***0.078***	***0.074***
	90th %ile	0.484	0.481	1.29	1.28	0.109	0.106
	95th %ile	*0.504*	*0.502*	*1.34*	*1.34*	*0.129*	*0.127*
	99th %ile	0.542	0.548	1.45	1.46	0.167	0.173

Table 26. Distribution of normalized and combination-adjusted p: SAVE

Case	***Statistic***	**Norm-*p***	**Comb-*p***	**NormOvr**	**CombOvr**	**NormDev**	**CombDev**
SAVE	1st %ile	0.048	0.158	0.19	0.62	-0.208	-0.098
n=2	*5th %ile*	*0.111*	*0.203*	*0.43*	*0.79*	*-0.145*	*-0.053*
truep=.26	10th %ile	0.133	0.218	0.52	0.85	-0.123	-0.038
	25th %ile	***0.188***	***0.258***	***0.73***	***1.01***	***-0.068***	***0.002***
	50th %ile	**0.250**	**0.304**	**0.98**	**1.19**	**-0.006**	**0.048**
	75th %ile	***0.316***	***0.353***	***1.23***	***1.38***	***0.060***	***0.097***
	90th %ile	0.375	0.399	1.46	1.56	0.119	0.143
	95th %ile	*0.421*	*0.436*	*1.64*	*1.70*	*0.165*	*0.180*
	99th %ile	0.579	0.567	2.26	2.21	0.323	0.311
SAVE	1st %ile	0.119	0.174	0.46	0.68	-0.137	-0.082
n=3	*5th %ile*	*0.150*	*0.198*	*0.59*	*0.77*	*-0.106*	*-0.058*
truep=.26	10th %ile	0.173	0.216	0.68	0.84	-0.083	-0.040
	25th %ile	***0.205***	***0.246***	***0.80***	***0.96***	***-0.051***	***-0.010***
	50th %ile	**0.250**	**0.281**	**0.98**	**1.10**	**-0.006**	**0.025**
	75th %ile	***0.300***	***0.322***	***1.17***	***1.26***	***0.044***	***0.066***
	90th %ile	0.342	0.361	1.34	1.41	0.086	0.105
	95th %ile	*0.370*	*0.390*	*1.45*	*1.52*	*0.114*	*0.134*
	99th %ile	0.440	0.450	1.72	1.76	0.184	0.194
SAVE	1st %ile	0.138	0.171	0.54	0.67	-0.118	-0.085
n=4+A209	*5th %ile*	*0.167*	*0.198*	*0.65*	*0.77*	*-0.089*	*-0.058*
truep=.26	10th %ile	0.188	0.215	0.73	0.84	-0.068	-0.041
	25th %ile	***0.218***	***0.240***	***0.85***	***0.94***	***-0.038***	***-0.016***
	50th %ile	**0.253**	**0.274**	**0.99**	**1.07**	**-0.003**	**0.018**
	75th %ile	***0.293***	***0.309***	***1.14***	***1.21***	***0.037***	***0.053***
	90th %ile	0.333	0.345	1.30	1.35	0.077	0.089
	95th %ile	*0.354*	*0.367*	*1.38*	*1.43*	*0.098*	*0.111*
	99th %ile	0.391	0.403	1.53	1.57	0.135	0.147
SAVE	1st %ile	0.157	0.181	0.61	0.71	-0.099	-0.075
n=5	*5th %ile*	*0.185*	*0.205*	*0.72*	*0.80*	*-0.071*	*-0.051*
truep=.26	10th %ile	0.198	0.216	0.77	0.84	-0.058	-0.040
	25th %ile	***0.225***	***0.241***	***0.88***	***0.94***	***-0.031***	***-0.015***
	50th %ile	**0.250**	**0.269**	**0.98**	**1.05**	**-0.006**	**0.013**
	75th %ile	***0.286***	***0.298***	***1.12***	***1.16***	***0.030***	***0.042***
	90th %ile	0.312	0.326	1.22	1.27	0.056	0.070
	95th %ile	*0.333*	*0.345*	*1.30*	*1.35*	*0.077*	*0.089*
	99th %ile	0.367	0.379	1.43	1.48	0.111	0.123
SAVE	1st %ile	0.168	0.180	0.66	0.70	-0.088	-0.076
n=6	*5th %ile*	*0.193*	*0.206*	*0.75*	*0.80*	*-0.063*	*-0.050*
truep=.26	10th %ile	0.200	0.216	0.78	0.84	-0.056	-0.040
	25th %ile	***0.224***	***0.235***	***0.88***	***0.92***	***-0.032***	***-0.021***
	50th %ile	**0.250**	**0.262**	**0.98**	**1.02**	**-0.006**	**0.006**
	75th %ile	***0.280***	***0.291***	***1.09***	***1.14***	***0.024***	***0.035***
	90th %ile	0.307	0.317	1.20	1.24	0.051	0.061
	95th %ile	*0.323*	*0.334*	*1.26*	*1.30*	*0.067*	*0.078*
	99th %ile	0.356	0.375	1.39	1.46	0.100	0.119

Table 27. Interval width as a function of sample size

Sample	***Statistic***	**NormOvr**	**CombOvr**	**NormDev**	**CombDev**
2	*Interquartile Range*	0.48	0.36	0.13	0.10
	90% Range	1.06	0.80	0.30	0.23
3	*Interquartile Range*	0.32	0.26	0.09	0.07
	90% Range	0.75	0.65	0.21	0.18
4	*Interquartile Range*	0.24	0.22	0.07	0.06
	90% Range	0.58	0.53	0.16	0.15
5	*Interquartile Range*	0.20	0.18	0.06	0.05
	90% Range	0.48	0.45	0.14	0.13
6	*Interquartile Range*	0.17	0.17	0.05	0.05
	90% Range	0.48	0.45	0.14	0.13

General Discussion

Solution to the Problem of Overestimation of *p* for Small Sample Sizes

The overestimation of small sample estimation of problem discovery *p* is clearly a problem (Hertzum & Jacobsen, in press; Lewis, 2000a). Various methods for discounting (reducing) this overestimation appear to work, with Good-Turing estimation providing a much more accurate estimate of the true value of *p* than unadjusted estimation (Lewis, 2000d). Good-Turing estimation, though, still generally leaves the estimate of *p* slightly inflated, leading to some underestimation of required total sample sizes when projecting from the initial sample. Estimating *p* with a normalization procedure based on the lower limit of possible *p* in a small-sample study has the same accuracy as Good-Turing estimation, but tends to underestimate true *p*, leading to some overestimation of required sample sizes when projecting from an initial sample. Averaging the Good-Turing and normalization estimates provides a highly accurate estimate of true *p* from very small samples, which in turn leads to highly accurate estimates of required sample sizes for specified problem discovery goals. These estimates are accurate enough that a practitioner should be able to make an initial projection from a combination estimate of *p* using a sample with as few as two participants, and will generally not underestimate the required sample size by much. A more conservative approach would be to use the normalized estimate of *p* when projecting from sample sizes with two or three participants (which should generally overestimate the required sample size slightly), and to use the combination estimate of *p* when projecting from sample sizes with four to six participants (which should usually be very accurate). The increased variation of *p* when estimated with a small sample also supports the use of the conservative approach. Using these techniques, usability practitioners can adjust small sample estimates of *p* when planning usability studies. As a study continues, practitioners can re-estimate *p* and project the revised sample size requirement.

The formula for the normalization estimate is:

[1] $truep = (estp - 1/n)(1 - 1/n)$

where *truep* is the adjusted estimate of *p* calculated from estimate of p derived from the participant by problem matrix (*estp*), and *n* is the sample size used to compute the initial estimate of *p*.

The formula for the combination estimate is:

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

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 (see Lewis, 2000d).

Surprisingly, the attempts to develop multiple regression equations for the prediction of true *p* did not fare as well as the non-regression approaches of Good-Turing and normalization. Even if the regression-based approaches had been as accurate (as measured by RMS error) as the non-regression approaches, the non-regression approaches would be preferable because they do not rely on statistically estimated parameters, making them solutions that have potentially greater generalizability.

Generalization Issues

Estimation of *p* from Other Problem Discovery Databases

Because the usability problem databases used to evaluate the different adjustment procedures described in this report 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). The following descriptions of the studies that were the source for the databases used in this report illustrate their broad diversity.

- 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.

Estimation of p from Larger Sample Sizes

The apparent trends in Figure 12 suggest that it might not be wise to use the combination or normalization approaches when the sample size exceeds six participants. At six participants, normalization continues to underestimate p, and the combination approach has begun to slightly underestimate p. 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 current data at what sample size a practitioner should abandon Good-Turing and move to the unadjusted estimate of p (a good topic for future research).

Conclusion

The use of these approaches for controlling the overestimation of p from small sample usability studies with six or fewer participants should work well. By adjusting the observed estimate of p, practitioners will be able to project required sample sizes more accurately, and will be able to get a better idea of the percentage of problems they have actually discovered from the set of problems available to discover.

References

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

Cliff, N. (1987). *Analyzing multivariate data.* San Diego, CA: Harcourt Brace Jovanovich.

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

Draper, N. R., & Smith, H. (1966). *Applied regression analysis*. New York, NY: John Wiley.

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

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

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

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

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

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

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

Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction*. Ft. Worth, TX: Harcourt Brace.

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.