# Using Discounting Methods to Reduce Overestimation of *p* in Problem Discovery Usability Studies

TR 29.3359

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 a potentially serious problem because it leads to underestimation of required sample sizes. The current Monte Carlo experiments demonstrate that discounting methods can significantly reduce this overestimation. Using the Good-Turing estimator, the mean extent of underestimation of the required sample size never exceeded one participant when estimating *p* from a six-participant sample. Thus, usability practitioners can use this technique to adjust small sample estimates of *p* when planning a usability study. As the study continues, they can re-estimate *p* and project revised required sample sizes.

## ITIRC Keywords

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

## 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.

**Discounting *p***

Given that small sample estimates of $p$ necessarily produce an overestimate, one way to reduce the magnitude of overestimation and increase the accuracy of the estimate is to apply 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).

The oldest discounting procedure is LaPlace's law of succession (Jelinek, 1997), sometimes referred to as the "add 1" method because you add one to the count for each observation. A common criticism of LaPlace's law is that it tends to assign too much of the probability space to unseen events, underestimating true $p$ (Manning and Schutze, 1999).

A widely used procedure that is more accurate than LaPlace's law is Good-Turing estimation (Jelinek, 1977; 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 $p$.

For problem discovery studies, there are other ways to systematically discount the estimate of $p$ by increasing the count in the denominator, such as adding the number of problems that occurred once (Add Ones), the total number of problems observed (Add Probs), or the total number of problem occurrences (Add Occs). Using the example in Table 3, this would result in estimates of .696 (16/(21+2)), .571 (16/(21+7)) and .432 (16/(21+16)) respectively.

Suppose one discount method consistently fails to reduce $p$ sufficiently, and a different one consistently reduces $p$ to too great an extent. It is then possible to use simple linear interpolation to arrive at the best estimate of $p$. In the examples used above, averaging the Add Occs estimation with the Add Probs estimation results in an estimate of .502 ((.571+.432)/2) – the closest estimate in this set of examples to the true $p$ of .500.

The purpose of the current Monte Carlo experiments was to investigate the extent to which different discounting methods compensate for the overestimation of $p$.

## 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 also generated this set of statistics for each of the following discount methods:

- Occs – add the number of problem occurrences to the denominator
- Probs – add the number of different types of problems observed to the denominator
- Ones – add the number of problems that occurred once to the denominator
- LinOp – compute the average of the estimates produced using the ADDOCC and ADDPROB procedures
- GT – adjust *p* using Good-Turing estimation

Good-Turing is in the set because it is a widely used discounting procedure. Its applicability to small sample estimates of probability is uncertain. The other discounting procedures use information easily derived from problem discovery studies, and provide a range of discounting from small to large (in order from least to greatest discounting, Ones, Probs, LinOp, Occs).

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

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

(For copies of the MACERR, VIRZI90, and MANTEL databases, see Lewis, 2000b. The SAVE database appears in Appendix A.)

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

## 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 discounting 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 discounting method (None, Ones, Probs, Occs, LinOp, and GT). The analysis indicated significant main effects of sample size ($F(4,12)=22.0$, $p=.00002$) and discounting method ($F(5,15)=30.6$, $p=.0000003$), and a significant interaction between these effects ($F(20,60)=29.0$, $p=.00035$). Table 4 and Figure 1 illustrate this interaction.

*Table 4. The sample size by discounting method interaction*

| ***Sample*** | **None** | **Ones** | **Probs** | **Occs** | **LinOp** | **GT** | **Average** |
|---|---|---|---|---|---|---|---|
| *N=2* | 0.360 | 0.201 | 0.146 | 0.105 | 0.120 | 0.113 | 0.174 |
| *N=3* | 0.239 | 0.162 | 0.110 | 0.064 | 0.080 | 0.077 | 0.122 |
| *N=4* | 0.177 | 0.132 | 0.088 | 0.053 | 0.059 | 0.060 | 0.095 |
| *N=5* | 0.138 | 0.108 | 0.071 | 0.047 | 0.048 | 0.048 | 0.077 |
| *N=6* | 0.113 | 0.092 | 0.060 | 0.045 | 0.042 | 0.043 | 0.066 |
| *Average* | 0.205 | 0.139 | 0.095 | 0.063 | 0.070 | 0.068 | |

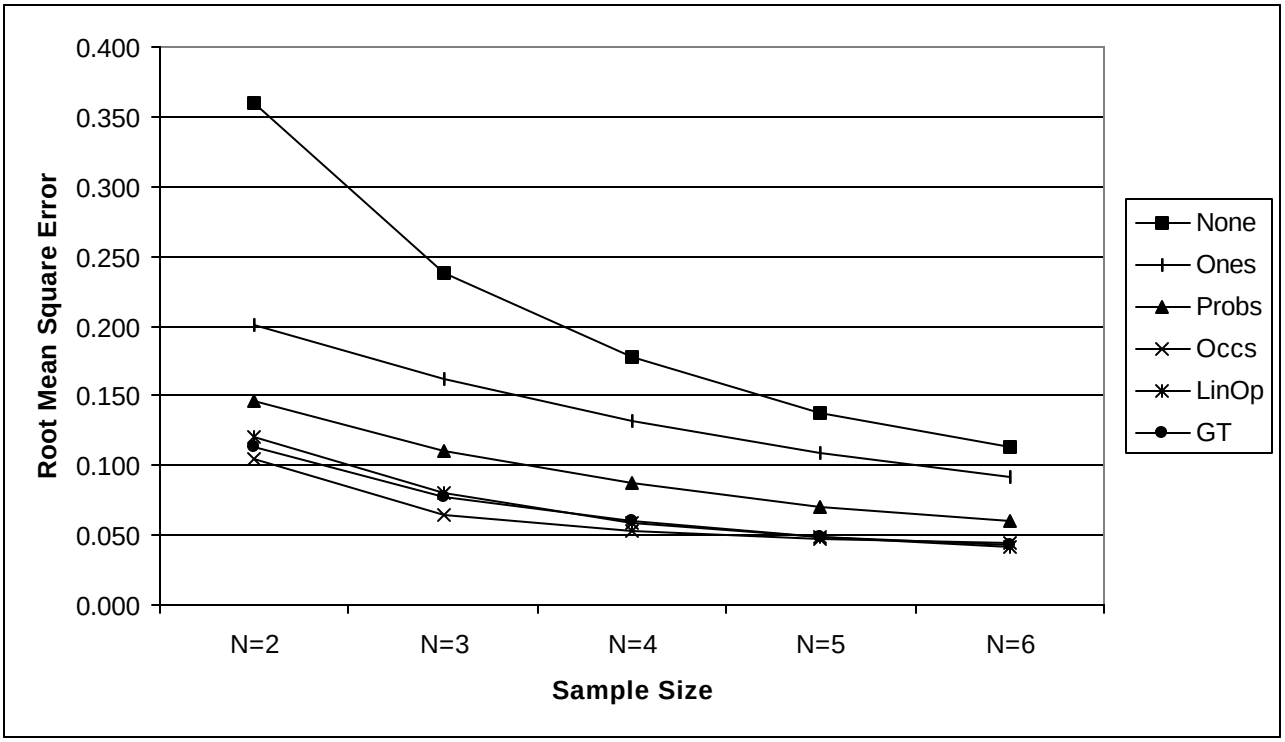


*Figure 1. The sample size by discounting method interaction*

Figure 1 shows that as the sample size increased, accuracy generally increased for all estimation procedures (the main effect of sample size). The lines show reasonably clear separation and relatively less accuracy for None, Ones, and Probs – no discounting and the two procedures that provided the least discounting. The RMS errors for Occs, LinOp and GT were almost identical, especially for sample sizes of four, five and six. The lines suggest potential convergence at some much larger sample size.

I used *t*-tests to compare, at each level of sample size, the significance of difference between no discounting and each discounting method and the significance of difference between Good-Turing estimation and the other procedures. Table 5 shows that all discounting methods improved estimation accuracy of *p* relative to no discounting. As shown in Table 6, Good-Turing estimation was more accurate than the Add Ones method. Because no other evaluated discounting procedure was significantly more accurate than Good-Turing, all additional analyses focus on that well-known estimator. All tests in Tables 5 and 6 had three degrees of freedom. Bold entries for a test's observed significance level (osl) indicate a statistically significant difference.

*Table 5. Comparisons of accuracy between discounting methods and no discounting*

| ***Sample*** | ***Statistic*** | **Ones** | **Probs** | **Occs** | **LinOp** | **GT** |
|---|---|---|---|---|---|---|
| *2* | *t* | 27.959 | 20.671 | 12.390 | 14.147 | 32.622 |
| | *osl* | **0.0001** | **0.0002** | **0.0011** | **0.0008** | **0.0001** |
| *3* | *t* | 20.231 | 13.287 | 8.123 | 10.381 | 17.928 |
| | *osl* | **0.0003** | **0.0009** | **0.0039** | **0.0019** | **0.0004** |
| *4* | *t* | 18.121 | 10.382 | 5.955 | 8.171 | 10.014 |
| | *osl* | **0.0004** | **0.0019** | **0.0095** | **0.0038** | **0.0021** |
| *5* | *t* | 15.488 | 9.064 | 4.341 | 6.014 | 6.115 |
| | *osl* | **0.0006** | **0.0028** | **0.0226** | **0.0092** | **0.0088** |
| *6* | *t* | 12.334 | 8.030 | 3.065 | 4.228 | 4.537 |
| | *osl* | **0.0011** | **0.004** | **0.0548** | **0.0242** | **0.02** |

*Table 6. Comparison of Good-Turing and other discounting procedures*

| ***Sample*** | ***Statistic*** | **Ones** | **Probs** | **Occs** | **LinOp** |
|---|---|---|---|---|---|
| *2* | *t* | -35.425 | -1.865 | 0.293 | -0.335 |
| | *osl* | **0.0000** | 0.159 | 0.7888 | 0.75966 |
| *3* | *t* | -13.230 | -1.852 | 0.458 | -0.158 |
| | *osl* | **0.0009** | 0.1611 | 0.6783 | 0.8844 |
| *4* | *t* | -7.036 | -1.601 | 0.299 | 0.003 |
| | *osl* | **0.0059** | 0.2076 | 0.7844 | 0.9977 |
| *5* | *t* | -4.513 | -1.270 | 0.037 | 0.001 |
| | *osl* | **0.0203** | 0.2936 | 0.9728 | 0.9993 |
| *6* | *t* | -3.488 | -1.064 | -0.177 | 0.051 |
| | *osl* | **0.0398** | 0.3654 | 0.871 | 0.9626 |

**Good-Turing Estimation and Projected Sample Sizes for Problem Discovery Studies**

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 7-10, 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.

Application of Good-Turing estimation did appear 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.

This is clear by examination of the data in Table 11. The cells of Table 11 contain the underestimate of the required sample size for that case. The value of true *p* seemed to affect the magnitude of unadjusted underestimation of required sample sizes, with lower values of *p* leading to greater underestimation. I conducted an analysis of variance on this data, treating databases as subjects in a within-subjects design with independent variables of sample size (with levels from two to six), discounting (with levels of None and GT), and discovery goal (with levels of 90% and 95%). The main effects of sample size ($F(4,12)=21.4$, $p=.004$) and discounting ($F(1,3)=15.3$, $p=.03$) were significant, as was the sample size by goal interaction ($F(4,12)=3.6$, $p=.04$, see Figure 6). As the size of the sample used to estimate *p* increased, the magnitude of underestimation in the projected sample size decreased. Good-Turing estimation reduced the magnitude of underestimation. Although the underestimation for 95% discovery consistently exceeded that for 90% discovery, as the sample size used to estimate *p* increased, the difference between the magnitude of underestimation for 90% and 95% discovery decreased.

*Table 7. Projected sample sizes for MACERR*

| *N* | NoAdj2 | GT2 | NoAdj3 | GT3 | NoAdj4 | GT4 | NoAdj5 | GT5 | NoAdj6 | GT6 | TrueP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *1* | 0.568 | 0.305 | 0.421 | 0.237 | 0.346 | 0.202 | 0.301 | 0.181 | 0.269 | 0.164 | 0.160 |
| *2* | 0.813 | 0.517 | 0.665 | 0.418 | 0.572 | 0.363 | 0.511 | 0.329 | 0.466 | 0.301 | 0.294 |
| *3* | **0.919** | 0.664 | 0.806 | 0.556 | 0.720 | 0.492 | 0.658 | 0.451 | 0.609 | 0.416 | 0.407 |
| *4* | ***0.965*** | 0.767 | 0.888 | 0.661 | 0.817 | 0.594 | 0.761 | 0.550 | 0.714 | 0.512 | 0.502 |
| *5* | 0.985 | 0.838 | **0.935** | 0.741 | 0.880 | 0.676 | 0.833 | 0.632 | 0.791 | 0.592 | 0.582 |
| *6* | 0.994 | 0.887 | ***0.962*** | 0.803 | **0.922** | 0.742 | 0.883 | 0.698 | 0.847 | 0.659 | 0.649 |
| *7* | 0.997 | **0.922** | 0.978 | 0.849 | 0.949 | 0.794 | **0.918** | 0.753 | 0.888 | 0.715 | 0.705 |
| *8* | 0.999 | 0.946 | 0.987 | 0.885 | ***0.967*** | 0.836 | 0.943 | 0.798 | **0.918** | 0.761 | 0.752 |
| *9* | 0.999 | ***0.962*** | 0.993 | **0.912** | 0.978 | 0.869 | ***0.960*** | 0.834 | 0.940 | 0.801 | 0.792 |
| *10* | 1.000 | 0.974 | 0.996 | 0.933 | 0.986 | 0.895 | 0.972 | 0.864 | ***0.956*** | 0.833 | 0.825 |
| *11* | 1.000 | 0.982 | 0.998 | 0.949 | 0.991 | **0.916** | 0.981 | 0.889 | 0.968 | 0.861 | 0.853 |
| *12* | 1.000 | 0.987 | 0.999 | ***0.961*** | 0.994 | 0.933 | 0.986 | **0.909** | 0.977 | 0.883 | 0.877 |
| *13* | 1.000 | 0.991 | 0.999 | 0.970 | 0.996 | 0.947 | 0.990 | 0.925 | 0.983 | **0.903** | 0.896 |
| *14* | 1.000 | 0.994 | 1.000 | 0.977 | 0.997 | ***0.958*** | 0.993 | 0.939 | 0.988 | 0.919 | **0.913** |
| *15* | 1.000 | 0.996 | 1.000 | 0.983 | 0.998 | 0.966 | 0.995 | ***0.950*** | 0.991 | 0.932 | 0.927 |
| *16* | 1.000 | 0.997 | 1.000 | 0.987 | 0.999 | 0.973 | 0.997 | 0.959 | 0.993 | 0.943 | 0.939 |
| *17* | 1.000 | 0.998 | 1.000 | 0.990 | 0.999 | 0.978 | 0.998 | 0.966 | 0.995 | ***0.952*** | 0.948 |
| *18* | 1.000 | 0.999 | 1.000 | 0.992 | 1.000 | 0.983 | 0.998 | 0.973 | 0.996 | 0.960 | ***0.957*** |
| *19* | 1.000 | 0.999 | 1.000 | 0.994 | 1.000 | 0.986 | 0.999 | 0.977 | 0.997 | 0.967 | 0.964 |
| *20* | 1.000 | 0.999 | 1.000 | 0.996 | 1.000 | 0.989 | 0.999 | 0.982 | 0.998 | 0.972 | 0.969 |

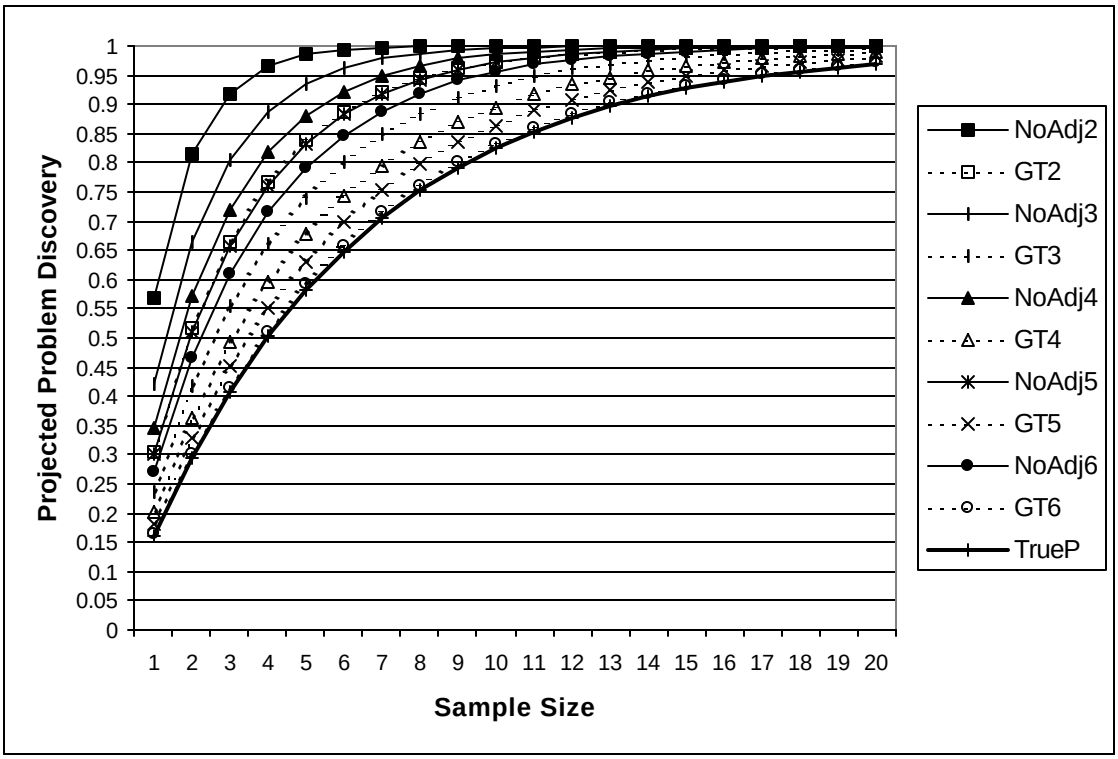


*Figure 2. Projected problem discovery curves for MACERR*

Note: NoAdj = no adjustment; GT = adjusted with Good-Turing estimation

*Table 8. Projected sample sizes for VIRZI90*

| *N* | NoAdj2 | GT2 | NoAdj3 | GT3 | NoAdj4 | GT4 | NoAdj5 | GT5 | NoAdj6 | GT6 | TrueP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *1* | 0.661 | 0.397 | 0.544 | 0.358 | 0.484 | 0.34 | 0.448 | 0.331 | 0.425 | 0.327 | 0.359 |
| *2* | 0.885 | 0.636 | 0.792 | 0.588 | 0.734 | 0.564 | 0.695 | 0.552 | 0.669 | 0.547 | 0.589 |
| *3* | ***0.961*** | 0.781 | **0.905** | 0.735 | 0.863 | 0.713 | 0.832 | 0.701 | 0.810 | 0.695 | 0.737 |
| *4* | 0.987 | 0.868 | ***0.957*** | 0.830 | **0.929** | 0.810 | **0.907** | 0.800 | 0.891 | 0.795 | 0.831 |
| *5* | 0.996 | **0.920** | 0.980 | 0.891 | ***0.963*** | 0.875 | 0.949 | 0.866 | **0.937** | 0.862 | 0.892 |
| *6* | 0.998 | ***0.952*** | 0.991 | **0.930** | 0.981 | **0.917** | ***0.972*** | **0.910** | ***0.964*** | **0.907** | **0.931** |
| *7* | 0.999 | 0.971 | 0.996 | ***0.955*** | 0.990 | 0.945 | 0.984 | 0.940 | 0.979 | 0.937 | ***0.956*** |
| *8* | 1.000 | 0.983 | 0.998 | 0.971 | 0.995 | ***0.964*** | 0.991 | ***0.960*** | 0.988 | ***0.958*** | 0.971 |
| *9* | 1.000 | 0.989 | 0.999 | 0.981 | 0.997 | 0.976 | 0.995 | 0.973 | 0.993 | 0.972 | 0.982 |
| *10* | 1.000 | 0.994 | 1.000 | 0.988 | 0.999 | 0.984 | 0.997 | 0.982 | 0.996 | 0.981 | 0.988 |
| *11* | 1.000 | 0.996 | 1.000 | 0.992 | 0.999 | 0.990 | 0.999 | 0.988 | 0.998 | 0.987 | 0.992 |
| *12* | 1.000 | 0.998 | 1.000 | 0.995 | 1.000 | 0.993 | 0.999 | 0.992 | 0.999 | 0.991 | 0.995 |
| *13* | 1.000 | 0.999 | 1.000 | 0.997 | 1.000 | 0.995 | 1.000 | 0.995 | 0.999 | 0.994 | 0.997 |
| *14* | 1.000 | 0.999 | 1.000 | 0.998 | 1.000 | 0.997 | 1.000 | 0.996 | 1.000 | 0.996 | 0.998 |
| *15* | 1.000 | 0.999 | 1.000 | 0.999 | 1.000 | 0.998 | 1.000 | 0.998 | 1.000 | 0.997 | 0.999 |
| *16* | 1.000 | 1.000 | 1.000 | 0.999 | 1.000 | 0.999 | 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 | 0.999 |
| *18* | 1.000 | 1.000 | 1.000 | 1.000 | 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 | 1.000 | 1.000 | 1.000 | 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 |

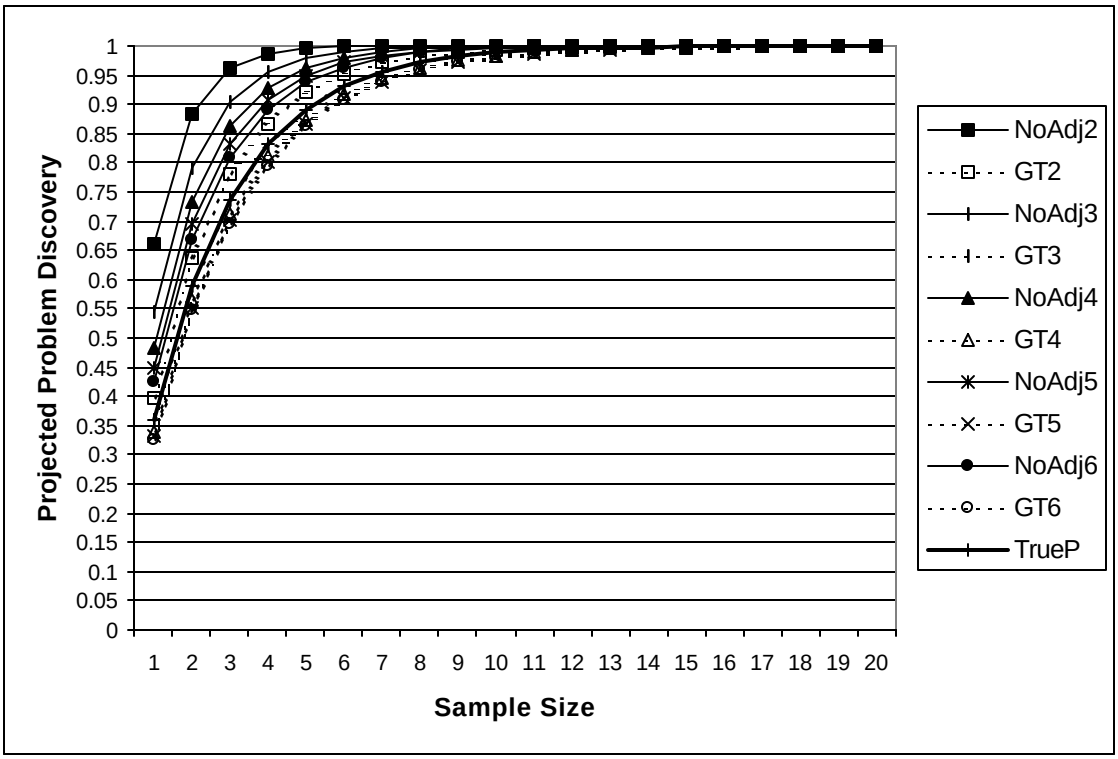


*Figure 3. Projected problem discovery curves for VIRZI90*

Note: NoAdj = no adjustment; GT = adjusted with Good-Turing estimation

*Table 9. Projected sample sizes for MANTEL*

| *N* | NoAdj2 | GT2 | NoAdj3 | GT3 | NoAdj4 | GT4 | NoAdj5 | GT5 | NoAdj6 | GT6 | TrueP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *1* | 0.724 | 0.473 | 0.622 | 0.446 | 0.572 | 0.431 | 0.536 | 0.415 | 0.511 | 0.403 | 0.375 |
| *2* | **0.924** | 0.722 | 0.857 | 0.693 | 0.817 | 0.676 | 0.785 | 0.658 | 0.761 | 0.644 | 0.609 |
| *3* | ***0.979*** | 0.854 | **0.946** | 0.830 | **0.922** | 0.816 | **0.900** | 0.800 | 0.883 | 0.787 | 0.756 |
| *4* | 0.994 | **0.923** | ***0.980*** | **0.906** | ***0.966*** | 0.895 | ***0.954*** | 0.883 | **0.943** | 0.873 | 0.847 |
| *5* | 0.998 | ***0.959*** | 0.992 | 0.948 | 0.986 | **0.940** | 0.978 | **0.931** | ***0.972*** | **0.924** | **0.905** |
| *6* | 1.000 | 0.979 | 0.997 | ***0.971*** | 0.994 | ***0.966*** | 0.990 | ***0.960*** | 0.986 | ***0.955*** | 0.940 |
| *7* | 1.000 | 0.989 | 0.999 | 0.984 | 0.997 | 0.981 | 0.995 | 0.977 | 0.993 | 0.973 | ***0.963*** |
| *8* | 1.000 | 0.994 | 1.000 | 0.991 | 0.999 | 0.989 | 0.998 | 0.986 | 0.997 | 0.984 | 0.977 |
| *9* | 1.000 | 0.997 | 1.000 | 0.995 | 1.000 | 0.994 | 0.999 | 0.992 | 0.998 | 0.990 | 0.985 |
| *10* | 1.000 | 0.998 | 1.000 | 0.997 | 1.000 | 0.996 | 1.000 | 0.995 | 0.999 | 0.994 | 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 | 1.000 | 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 |

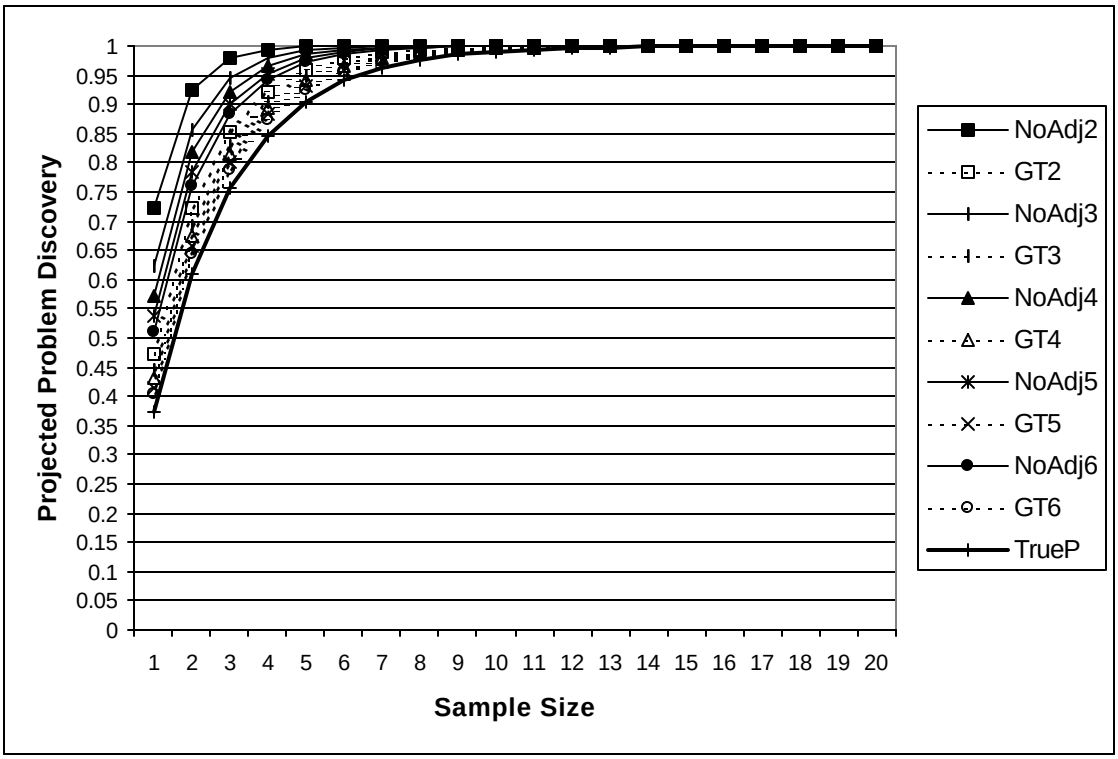


*Figure 4. Projected problem discovery curves for MANTEL*

Note: NoAdj = no adjustment; GT = adjusted with Good-Turing estimation

*Table 10. Projected sample sizes for SAVE*

| *N* | NoAdj2 | GT2 | NoAdj3 | GT3 | NoAdj4 | GT4 | NoAdj5 | GT5 | NoAdj6 | GT6 | TrueP |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *1* | 0.627 | 0.361 | 0.503 | 0.318 | 0.442 | 0.298 | 0.403 | 0.284 | 0.381 | 0.280 | 0.256 |
| *2* | 0.861 | 0.592 | 0.753 | 0.535 | 0.689 | 0.507 | 0.644 | 0.487 | 0.617 | 0.482 | 0.446 |
| *3* | **0.948** | 0.739 | 0.877 | 0.683 | 0.826 | 0.654 | 0.787 | 0.633 | 0.763 | 0.627 | 0.588 |
| *4* | ***0.981*** | 0.833 | **0.939** | 0.784 | **0.903** | 0.757 | 0.873 | 0.737 | 0.853 | 0.731 | 0.694 |
| *5* | 0.993 | 0.893 | ***0.970*** | 0.852 | 0.946 | 0.830 | **0.924** | 0.812 | **0.909** | 0.807 | 0.772 |
| *6* | 0.997 | **0.932** | 0.985 | 0.899 | ***0.970*** | 0.880 | ***0.955*** | 0.865 | 0.944 | 0.861 | 0.830 |
| *7* | 0.999 | ***0.956*** | 0.993 | **0.931** | 0.983 | **0.916** | 0.973 | **0.904** | ***0.965*** | **0.900** | 0.874 |
| *8* | 1.000 | 0.972 | 0.996 | ***0.953*** | 0.991 | 0.941 | 0.984 | 0.931 | 0.978 | 0.928 | **0.906** |
| *9* | 1.000 | 0.982 | 0.998 | 0.968 | 0.995 | ***0.959*** | 0.990 | ***0.951*** | 0.987 | 0.948 | 0.930 |
| *10* | 1.000 | 0.989 | 0.999 | 0.978 | 0.997 | 0.971 | 0.994 | 0.965 | 0.992 | ***0.963*** | 0.948 |
| *11* | 1.000 | 0.993 | 1.000 | 0.985 | 0.998 | 0.980 | 0.997 | 0.975 | 0.995 | 0.973 | ***0.961*** |
| *12* | 1.000 | 0.995 | 1.000 | 0.990 | 0.999 | 0.986 | 0.998 | 0.982 | 0.997 | 0.981 | 0.971 |
| *13* | 1.000 | 0.997 | 1.000 | 0.993 | 0.999 | 0.990 | 0.999 | 0.987 | 0.998 | 0.986 | 0.979 |
| *14* | 1.000 | 0.998 | 1.000 | 0.995 | 1.000 | 0.993 | 0.999 | 0.991 | 0.999 | 0.990 | 0.984 |
| *15* | 1.000 | 0.999 | 1.000 | 0.997 | 1.000 | 0.995 | 1.000 | 0.993 | 0.999 | 0.993 | 0.988 |
| *16* | 1.000 | 0.999 | 1.000 | 0.998 | 1.000 | 0.997 | 1.000 | 0.995 | 1.000 | 0.995 | 0.991 |
| *17* | 1.000 | 1.000 | 1.000 | 0.999 | 1.000 | 0.998 | 1.000 | 0.997 | 1.000 | 0.996 | 0.993 |
| *18* | 1.000 | 1.000 | 1.000 | 0.999 | 1.000 | 0.998 | 1.000 | 0.998 | 1.000 | 0.997 | 0.995 |
| *19* | 1.000 | 1.000 | 1.000 | 0.999 | 1.000 | 0.999 | 1.000 | 0.998 | 1.000 | 0.998 | 0.996 |
| *20* | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.999 | 1.000 | 0.999 | 1.000 | 0.999 | 0.997 |

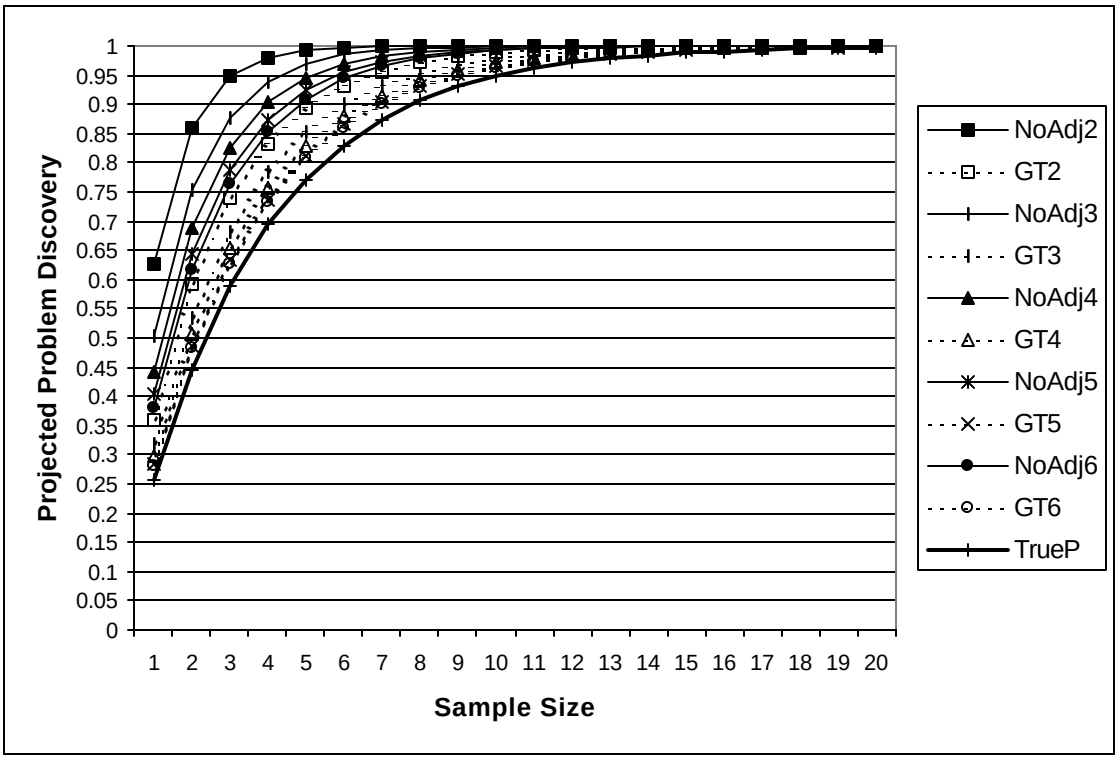


*Figure 5. Projected problem discovery curves for SAVE*

Note: NoAdj = no adjustment; GT = adjusted with Good-Turing estimation

*Table 11. Underestimates of required sample sizes*

| ***Database*** | ***N*** | **None90** | **GT90** | **None95** | **GT95** |
|---|---|---|---|---|---|
| *MACERR* | *2* | 11 | 7 | 14 | 9 |
| *(true p=.16)* | *3* | 9 | 5 | 12 | 6 |
| | *4* | 8 | 4 | 10 | 4 |
| | *5* | 7 | 2 | 9 | 3 |
| | *6* | 6 | 1 | 8 | 1 |
| *VIRZI90* | ***N*** | **None90** | **GT90** | **None95** | **GT95** |
| *(true p=.36)* | *2* | 3 | 1 | 4 | 1 |
| | *3* | 3 | 0 | 3 | 0 |
| | *4* | 2 | 0 | 2 | -1 |
| | *5* | 2 | 0 | 1 | -1 |
| | *6* | 1 | 0 | 1 | -1 |
| *MANTEL* | ***N*** | **None90** | **GT90** | **None95** | **GT95** |
| *(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** | **GT90** | **None95** | **GT95** |
| *(true p=.26)* | *2* | 5 | 2 | 7 | 4 |
| | *3* | 4 | 1 | 6 | 3 |
| | *4* | 4 | 1 | 5 | 2 |
| | *5* | 3 | 1 | 5 | 2 |
| | *6* | 3 | 1 | 4 | 1 |
| *AVERAGE* | ***N*** | **None90** | **GT90** | **None95** | **GT95** |
| | *2* | 5.5 | 2.8 | 7.3 | 4.0 |
| | *3* | 4.5 | 1.8 | 6.0 | 2.5 |
| | *4* | 4.0 | 1.3 | 5.0 | 1.5 |
| | *5* | 3.5 | 0.8 | 4.5 | 1.3 |
| | *6* | 2.8 | 0.5 | 3.8 | 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. None indicates no discounting; GT indicates Good-Turing discounting. 90 and 95 indicate the hypothetical problem discovery goals.

*Table 12. Underestimates of required sample sizes: Sample size by discovery goal interaction*

| ***Sample*** | **90%** | **95%** |
|---|---|---|
| *2* | 4.1 | 5.6 |
| *3* | 3.1 | 4.3 |
| *4* | 2.6 | 3.3 |
| *5* | 2.1 | 2.9 |
| *6* | 1.6 | 2.1 |

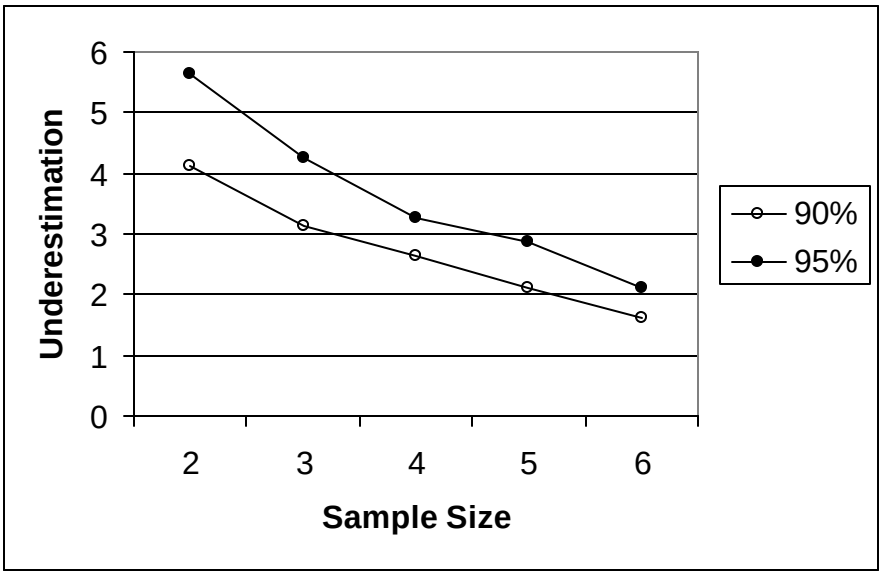


*Figure 6. Projected sample size underestimation: Sample size by discovery goal interaction*

## Discussion

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 (although other discounting procedures based on problem discovery measures also appear to work well). 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. The magnitude of this underestimation of required sample size decreases as the size of the initial sample used to estimate *p* increases. For initial sample sizes of four to six participants, the magnitude of underestimation ranged from about 1.5 to 0.5 participants. Thus, final sample sizes projected from Good-Turing estimates based on initial sample sizes of six participants should generally be quite accurate. For each of the investigated problem discovery databases and both problem discovery goals, the mean extent of underestimation of the required sample size never exceeded one participant when estimating *p* from a six-participant sample. This suggests that usability practitioners can use this technique to adjust a small sample estimate of *p* when planning a usability study. As the study continues, they can re-estimate *p* and project the revised sample size requirement.

## References


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

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

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

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

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

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

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

Lewis, J. R. (2000a). *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., Henry, S. C., & Mack, R. L. (1990). Integrated office scenario benchmarks: A case study. In *Human Computer Interaction -- INTERACT '90* (pp. 337-343). Cambridge, England: Elsevier Science Publishers, IFIP.

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

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

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

## Appendix A. The SAVE Problem Discovery Database

In this database, the first two values are (1) the number of problems and (2) the number of participants. After those values, the leftmost column is a list of problem identification numbers and the rightmost column is a placeholder for problem impact (severity) ratings. Because impact ratings were not available for this database, the rightmost column contains all zeros. The intervening columns represent the participants in the study. In the cells of the participant columns, a '1' indicates that the participant experienced the identified problem during the usability evaluation and a '0' indicates that the participant did not experience that problem. The MACERR, VIRZI90 and MANTEL databases are available in Lewis (2000a).

| 48 | 34 | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 2 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 |
| 3 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 |
| 4 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |
| 5 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| 6 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 |
| 7 | 1 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 |
| 8 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| 9 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 0 |
| 10 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |
| 11 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 |
| 12 | 1 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 |
| 13 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 1 | 1 | 0 |
| 14 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 |
| 15 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 1 | 1 | 1 | 1 | 1 | 0 | 1 | 0 |
| 16 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 |
| 17 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 1 | 0 |
| 18 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 1 | 0 |
| 19 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| 20 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 1 | 0 |
| 21 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 22 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 23 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| 24 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 25 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 1 | 0 | 1 | 0 |
| 26 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 |
| 27 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 |
| 28 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 1 | 0 |
| 29 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 |
| 30 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 | 1 | 0 |
| 31 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 |
| 32 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 |
| 33 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 34 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 |
| 35 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 36 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 1 | 0 | 1 | 0 | 0 |
| 37 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 38 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |
| 39 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 40 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 41 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 |
| 42 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 |
| 43 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 44 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 45 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 46 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 47 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| 48 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |