INTERNATIONAL JOURNAL OF HUMAN–COMPUTER INTERACTION, *13*(4), 445–479

Evaluation of Procedures for Adjusting Problem-Discovery Rates Estimated From Small Samples

James R. Lewis
IBM Corporation

There are 2 excellent reasons to compute usability problem-discovery rates. First, an estimate of the problem-discovery rate is a key component for projecting the required sample size for a usability study. Second, practitioners can use this estimate to calculate the proportion of discovered problems for a given sample size. Unfortunately, small-sample estimates of the problem-discovery rate suffer from a serious overestimation bias. This bias can lead to serious underestimation of required sample sizes and serious overestimation of the proportion of discovered problems. This article contains descriptions and evaluations of a number of methods for adjusting small-sample estimates of the problem-discovery rate to compensate for this bias. A series of Monte Carlo simulations provided evidence that the average of a normalization procedure and Good–Turing (Jelinek, 1997; Manning & Schutze, 1999) discounting produces highly accurate estimates of usability problem-discovery rates from small sample sizes.

1. INTRODUCTION

1.1. Problem-Discovery Rate and Its Applications

Many usability studies (such as scenario-based usability evaluations and heuristic evaluations) have as their primary goal the discovery of usability problems (Lewis, 1994). Practitioners use the discovered problems to develop recommendations for the improvement of the system or product under study (Norman, 1983). The problem-discovery rate (p) for a usability study is, across a sample of participants, the average of the proportion of problems observed for each participant (or the average of the proportion of participants experiencing each observed problem).

It is important to keep in mind that no single usability method can detect all possible usability problems. In this article, reference to the proportion of problems detected means the number of problems detected over the number of detectable prob-

Requests for reprints should be sent to James R. Lewis, IBM Corporation, 8051 Congress Avenue, Suite 2227, Boca Raton, FL 33487. E-mail: jimlewis@us.ibm.com

lems. Furthermore, the focus is solely on the frequency of problem occurrence rather than severity or impact. Despite some evidence that the discovery rate for very severe problems is greater than that for less severe problems (Nielsen, 1992; Virzi, 1992), it is possible that this effect depends on the method used to rate severity (Lewis, 1994) or the expertise of the evaluators (Connell & Hammond, 1999).

Discovering usability problems is fundamentally different from discovering diamonds in a diamond mine. The diamonds in a mine may be well hidden, but a group of skilled miners would have excellent agreement about the methods to use to discover them and would not need to spend time discussing whether a particular object in the mine was or was not a diamond. Usability problems, on the other hand, do not exist as a "set of objectively defined, nicely separated problems just waiting for discovery" (M. Hertzum, personal communication, October 26, 2000). Different evaluators can disagree about the usability problems contained in an interface (Hertzum & Jacobsen, this issue). Reference to the number of detectable problems in this article means nothing more than the number of detectable problems given the limitations of a specific usability evaluation setting.

The number of detectable problems can vary as a function of many factors, including but not limited to the number of observers, the expertise of observers, the expertise of participants, and the specific set of scenarios-of-use in problem-discovery observational studies. In heuristic evaluations, the expertise of evaluators and specific set of heuristics can affect the number of detectable problems. For either observational or heuristic methods, the stage of product development and degree of implementation (e.g., paper prototype vs. full implementation) can also affect the number of detectable problems. By whatever means employed, however, once investigators have a set of usability problems in which they can code the presence and absence of problems across participants or observers as a series of 0s and 1s, those sets have some interesting properties.

A hypothetical example. Suppose a usability study of 10 participants performing a set of tasks with a particular software application (or 10 independent evaluators in a heuristic evaluation) had the outcome illustrated in Table 1.

Because there are 10 problems and 10 participants, the table contains 100 cells. An "x" in a cell indicates that the specified participant experienced the specified problem. With 50 cells filled, the estimated problem-discovery rate (also known as the average likelihood of problem detection) is .50 (50/100). (Note that the averages of the elements in the Proportion column and the Proportion row in the table are both .50.)

Projection of sample size requirements for usability studies. A number of large-sample usability and heuristic studies (Lewis, 1994; Nielsen & Landauer, 1993; Virzi, 1992) have shown that usability practitioners can use Equation 1 to project problem discovery as a function of the sample size n (number of participants or evaluators) and the problem-discovery rate p.

$$1-(1-p)^n \tag{1}$$

Table 1: Hypothetical Results for a Problem-Discovery Usability Study

Participant	*Prob 1*	*Prob 2*	*Prob 3*	*Prob 4*	*Prob 5*	*Prob 6*	*Prob 7*	*Prob 8*	*Prob 9*	*Prob 10*	*Count*	*Proportion*
1	x	x		x		x		x		x	6	0.6
2	x	x		x		x		x			5	0.5
3	x	x		x	x	x					5	0.5
4	x	x		x			x				4	0.4
5	x	x	x	x		x			x		6	0.6
6	x	x	x					x			4	0.4
7	x	x	x		x						4	0.4
8	x	x	x		x		x				5	0.5
9	x		x		x		x		x		5	0.5
10	x		x		x		x		x	x	6	0.6
Count	10	8	6	5	5	4	4	3	3	2	50	
Proportion	1.0	0.8	0.6	0.5	0.5	0.4	0.4	0.3	0.3	0.2		0.50

Note. Prob = problem; x = specified participant experienced specified problem.

It is possible to derive Equation 1 from either the binomial probability formula (Lewis, 1982, 1994) or the constant probability path independent Poisson process model (T. K. Landauer, personal communication, December 20, 1999; Nielsen & Landauer, 1993). Regardless of derivational perspective, an accurate estimate of p is essential for the accurate projection of cumulative problem discovery as a function of n, which is in turn essential for the accurate estimation of the sample size required to achieve a given problem-discovery goal (e.g., 90% problem discovery). Figure 1 illustrates projected problem-discovery curves for a variety of values of p.

Estimation of the proportion of discovered problems as a function of sample size. There are times when a usability practitioner does not have full control over the sample size due to various time or other resource constraints. In those cases, the practitioner might want to estimate the proportion of problems detected from the full set of problems available for detection by the method employed by the practitioner. Suppose the practitioner has observed the first 3 participants and has obtained the results presented in the first three rows of Table 1. Because $p = .5$ and $n = 3$, the estimated proportion of problems detected is .875, or $1 - (1 - .5)^3$.

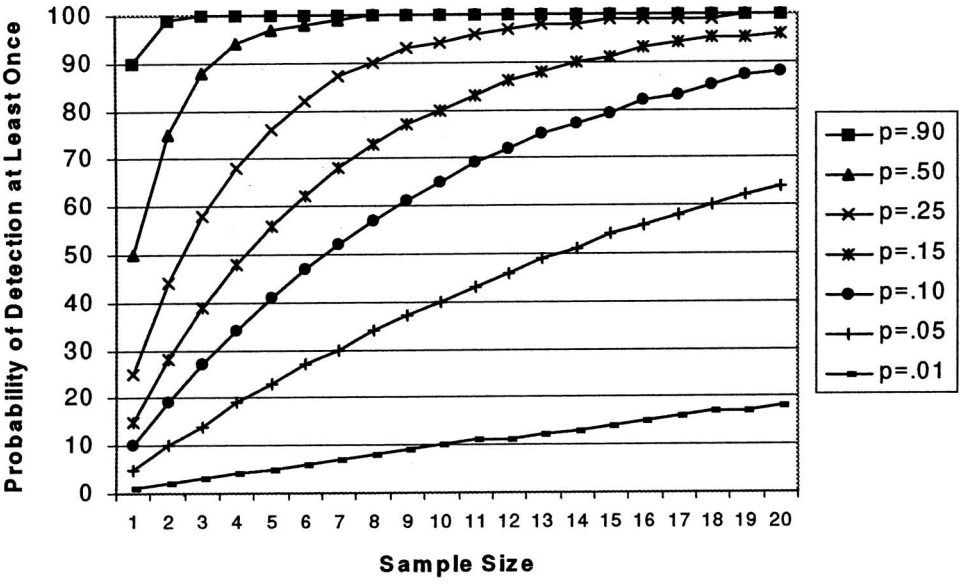

FIGURE 1 Projected problem-discovery curves as a function of n for various values of p.

1.2. Small-Sample Overestimation of the Problem-Discovery Rate

In the preceding example, however, how does the practitioner know that the problem-discovery rate is .5? Observation of the first 3 participants only uncovers Problems 1, 2, 4, 5, 6, 8, and 10, with the participant-to-problem distribution shown in Table 2. Because there are 21 cells (7 observed problems × 3 participants), with 16 of the cells containing an "x," the estimated value of p is .76, about a 50% overestimation relative to the full set of data presented in Table 1. Given this estimate of p, a practitioner would conclude that the study had uncovered 98.6% of the detectable problems.

Hertzum and Jacobsen (this issue) were the first to identify the problem of overestimating p from small-sample usability studies. They pointed out that the smallest possible value of p estimable from a given study is $1/n$, where n is the number of participants in the study. If a study has only 1 participant, the estimate of p will necessarily be 1.0. If a study has 2 participants, the smallest possible estimate of p, regardless of its true value, is .5. Furthermore, the study will produce this minimum value only when every problem is idiosyncratic (no duplication of any problem across multiple participants). Any duplication necessarily increases the estimate of p.

Goals of this research. In the remainder of this article, I describe research undertaken to accomplish the following goals:

1. Develop an understanding of the extent to which the overestimation of p is a potential problem for usability practitioners.
2. Investigate different procedures for the adjustment of the initial estimate of p to an estimate closer to the true value of p.
3. For the best adjustment procedures, determine how their overestimation or underestimation of p affects projected sample sizes and deviations from problem-discovery goals.

2. METHOD

In the first part of this section, I provide some background on the different adjustment procedures investigated in this series of experiments. In the remaining

Table 2: Hypothetical Results for a Problem-Discovery Usability Study: First 3 Participants

Participant	*Prob 1*	*Prob 2*	*Prob 4*	*Prob 5*	*Prob 6*	*Prob 8*	*Prob 10*	*Count*	*Proportion*
1	x	x	x		x	x	x	6	0.86
2	x	x	x		x	x		5	0.71
3	x	x	x	x	x			5	0.71
Count	3	3	3	1	3	2	1	16	
Proportion	1.00	1.00	1.00	0.33	1.00	0.67	0.33		0.76

Note. Prob = problem; x = specified participant experienced specified problem.

sections, I describe the dependent measurements used to evaluate the adjustment procedures and the problem-discovery databases used as source material for the investigations.

2.1. Adjustment Methods

I investigated three approaches for adjusting the initial estimate of *p:* discounting, normalization, and regression.

Discounting. One way to reduce the magnitude of overestimation of p 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 & Schutze, 1999).

The oldest discounting procedure is LaPlace's law of succession (Jelinek, 1997), sometimes referred to as the "add One" 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 & Schutze, 1999).

A widely used procedure that is more accurate than LaPlace's law is Good–Turing estimation (GT; Jelinek, 1997; Manning & Schutze, 1999). There are a number of paths that lead to the derivation of the GT 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 2, $E(N_1)$ 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 2, this would result in estimates of .696, or 16/(21 + 2); .571, or 16/(21 + 7); and .432, or 16/(21 + 16), respectively.

Suppose one discount method consistently fails to reduce the estimate of p sufficiently, and a different one consistently reduces it to too great an extent. It is then possible to use simple linear interpolation to arrive at a better estimate of p. In the examples used previously, averaging the Add Occs estimation with the Add Probs

estimation results in an estimate of .502, or (.571 + .432)/2—the closest estimate in this set of examples to the true p of .500.

Normalization. In their discussion of the problem of overestimation of p, Hertzum and Jacobsen (this issue) pointed out that the smallest possible value of p from a small-sample problem-discovery study is $1/n$. For estimates based on a small sample size, this limit can produce considerable overestimation of p. With larger sample sizes, the effect of this limit on the lowest possible value of p becomes less important. For example, if a study includes 20 participants, then the lower limit for p is 1/20, or .05.

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 2, 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 is a considerable underestimation of true p. It is not clear, though, how serious the underestimation would typically be, so submitting this procedure to more systematic evaluation would be reasonable.

Regression. Another approach for the estimation of true p from a small sample would be to develop one or more 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.

2.2. Measurements

I wrote a BASIC program that used a Monte Carlo procedure to sample data from published problem-discovery databases with known values of true p (Lewis, 2000e) to produce the following measurements:

1. Mean value of p.
2. Root mean square error (*RMSE*) for estimated p against true p.
3. The central 50% (interquartile) range of the distribution of p.
4. The central 90% range of the distribution of p.

The program could produce this set of statistics for the unadjusted estimate of p and up to five adjusted estimates (based on 1,000 Monte Carlo iterations for each problem-discovery database at each level of sample size). A preliminary experiment (Lewis, 2000e) confirmed that the programmed Monte Carlo procedure sampled randomly and produced results that were virtually identical to complete fac-

torial arrangement of participants for sample sizes of 2, 3, and 4. The impact of combinatorial expansion (the large number of possible combinations of participants) prevented the use of factorial arrangement for the investigation of larger sample sizes.

2.3. Problem-Discovery Databases

The published problem-discovery databases evaluated in this article were:

1. 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 6 hr, 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.
2. 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.
3. MANTEL (Nielsen & Molich, 1990): These usability problems came from 76 submissions to a contest presented in the Danish edition of *Computerworld* (excluding data from one submission that did not list any problems). 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.
4. SAVINGS (Nielsen & Molich, 1990): For this study, 34 computer science students taking a course in user interface design performed heuristic evaluations of an interactive voice response system (working and deployed) designed to give banking customers information such as their account balances and currency exchange rates. The participants uncovered 48 different usability problems with a problem-discovery rate (p) of .26.

(For figures depicting the MANTEL and SAVINGS databases, see Nielsen & Molich, 1990. For the VIRZI90 database, see Virzi, 1990. The MACERR database is available in the Appendix of this article.)

These were the only large-scale problem-discovery databases available to me for analysis. Fortunately, they had considerable variation in total number of participants, total number of usability problems uncovered, basic problem-discovery rate, and method of execution (observational with and without talk-aloud vs. heuristic, different error classification procedures). Given this variation among the databases, any consistent results obtained by evaluating them stands a good chance of generalizing to other problem-discovery databases (Chapanis, 1988).

3. MONTE CARLO SIMULATIONS OF PROBLEM DISCOVERY

This section of the article contains descriptions of Monte Carlo simulations of problem discovery conducted to

1. Evaluate the degree of overestimation produced by small-sample estimates of p (Section 3.1).
2. Investigate a set of discounting procedures to determine which method best adjusted the initial estimates of p (Section 3.2).
3. Develop a set of regression equations for estimating true p from an initial estimate (Section 3.3).
4. Investigate the normalization procedure and the regression equations to determine which procedure in the set best adjusted the initial estimates of p (Section 3.4).
5. Investigate the effectiveness of adjustment using a combination of GT discounting and normalization (Section 3.5).
6. Replicate and extend the previous findings for the best procedures from each class of procedure (discounting, normalization, and regression) using a greater range of sample sizes (Section 3.6).

3.1. How Serious Is the Small-Sample Overestimation of Problem-Discovery Rates?

The primary purpose of the first Monte Carlo experiment was to investigate the extent to which calculating p from small-sample usability studies results in overestimation.

Table 3: Mean Estimates of Discovery Rates as a Function of Sample Size for Published Databases

Source	*True p*	*N = 2*	*N = 3*	*N = 4*	*N = 5*	*N = 6*
MACERR[a]	.16	0.568	0.421	0.346	0.301	0.269
VIRZI90[b]	.36	0.661	0.544	0.484	0.448	0.425
MANTEL[c]	.38	0.724	0.622	0.572	0.536	0.511
SAVINGS[c]	.26	0.629	0.505	0.442	0.406	0.380

[a]Lewis (1994) and Lewis, Henry, and Mack (1990). [b]Virzi (1990, 1992). [c]Nielsen and Molich (1990).

Table 4: Mean Estimates of Root Mean Square Error as a Function of Sample Size for Published Databases

Source	*True p*	*N = 2*	*N = 3*	*N = 4*	*N = 5*	*N = 6*
MACERR[a]	.16	0.406	0.259	0.185	0.140	0.107
VIRZI90[b]	.36	0.306	0.189	0.130	0.094	0.071
MANTEL[c]	.38	0.354	0.254	0.204	0.167	0.143
SAVINGS[c]	.26	0.377	0.253	0.191	0.155	0.129

[a]Lewis (1994) and Lewis, Henry, and Mack (1990). [b]Virzi (1990, 1992). [c]Nielsen and Molich (1990).

Small-sample estimates of p. Table 3 shows the mean Monte Carlo estimates of *p* from the published databases for sample sizes from 2 to 6.

RMSE for unadjusted estimates. Table 4 shows the average *RMSE* from the Monte Carlo simulation. The *RMSE* is similar to a standard deviation, but rather than computing the mean squared difference between each data point and the mean of a sample (the standard deviation), the computation is the mean squared difference between each data point and true *p* (based on all participants in a given database). The *RMSE* has the desirable characteristic (for a measure of accuracy) of being sensitive to both the central tendency and variance of a measure. In other words, if two measurement methods are equally accurate with regard to the deviation of their mean from a known true value, the measurement with lower variance will have the lower *RMSE.* A perfect estimate would have an *RMSE* of 0.

Discussion. The initial estimates of *p* for all four databases clearly overestimated the true value of *p,* with the extent of the overestimation declining as the sample size increased, but still overestimating when $n = 6$. The *RMSE* showed a similar pattern, with the amount of error declining as the sample size increased, but with nonzero error remaining when $n = 6$. These data indicate that the overestimation problem is serious and provide baselines for the evaluation of adjustment procedures. (For a more comprehensive analysis of the data, see Lewis, 2000b).

3.2. Evaluation of Discounting Procedures

The purpose of the second Monte Carlo experiment was to evaluate five different adjustments based on the discounting procedures discussed previously: GT, Add Ones (ONES), Add Probs (PROBS), Add Occs (OCCS), and the mean of PROBS and OCCS (PR–OCC).

Estimates of p. Table 5 shows the mean Monte Carlo estimates of *p* (unadjusted and adjusted using the discounting procedures) from the published databases for sample sizes from 2 to 6.

Table 5: Adjusted Discovery Rates as a Function of Sample Size and Discounting Method

Source	*True p*	*Adjustment*	*N = 2*	*N = 3*	*N = 4*	*N = 5*	*N = 6*
MACERR[a]	.16	NONE	0.568	0.421	0.346	0.301	0.269
		ONES	0.397	0.334	0.294	0.266	0.243
		PROBS	0.378	0.315	0.277	0.251	0.230
		OCCS	0.362	0.296	0.257	0.231	0.212
		PR–OCC	0.372	0.304	0.264	0.237	0.217
		GT	0.305	0.237	0.202	0.181	0.164
VIRZI90[b]	.36	NONE	0.661	0.544	0.484	0.448	0.425
		ONES	0.495	0.463	0.437	0.418	0.405
		PROBS	0.441	0.408	0.387	0.374	0.364
		OCCS	0.397	0.352	0.326	0.309	0.298
		PR–OCC	0.427	0.383	0.356	0.338	0.327
		GT	0.397	0.358	0.340	0.331	0.327
MANTEL[c]	.38	NONE	0.724	0.622	0.572	0.536	0.511
		ONES	0.571	0.549	0.528	0.506	0.489
		PROBS	0.483	0.467	0.458	0.447	0.438
		OCCS	0.419	0.383	0.363	0.348	0.338
		PR–OCC	0.466	0.436	0.418	0.402	0.390
		GT	0.473	0.446	0.431	0.415	0.403
SAVINGS[c]	.26	NONE	0.627	0.502	0.442	0.403	0.381
		ONES	0.458	0.421	0.394	0.371	0.359
		PROBS	0.418	0.377	0.354	0.335	0.326
		OCCS	0.385	0.334	0.306	0.287	0.275
		PR–OCC	0.401	0.355	0.330	0.311	0.301
		GT	0.361	0.318	0.298	0.284	0.280

Note. NONE = no adjustment; ONES = discounted adjustment with the Add Ones method; PROBS = discounted adjustment with the Add Problems method; OCCS = discounted adjustment with the Add Occurences method; PR–OCC = mean of PROBS and OCCS estimates; GT = Good– Turing estimation.
[a]Lewis (1994) and Lewis, Henry, and Mack (1990). [b]Virzi (1990, 1992). [c]Nielsen and Molich (1990).

Evaluation of accuracy. I conducted an analysis of variance using *RMSE* as the dependent variable and treating databases as subjects in a within-subjects design. The independent variables were sample size (from 2 to 6) and discounting method (NONE, ONES, PROBS, OCCS, PR–OCC, 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, $F(20, 60) = 29.0$, $p = .00035$. Figure 2 illustrates the interaction.

Figure 2 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 *RMSE*s for OCCS, PR–OCC, and GT were almost identical, especially for sample sizes of 4, 5, and 6. The lines suggest potential convergence at some larger sample size.

A set of planned *t* tests showed that all discounting methods improved estimation accuracy of *p* relative to no discounting at every level of sample size (25

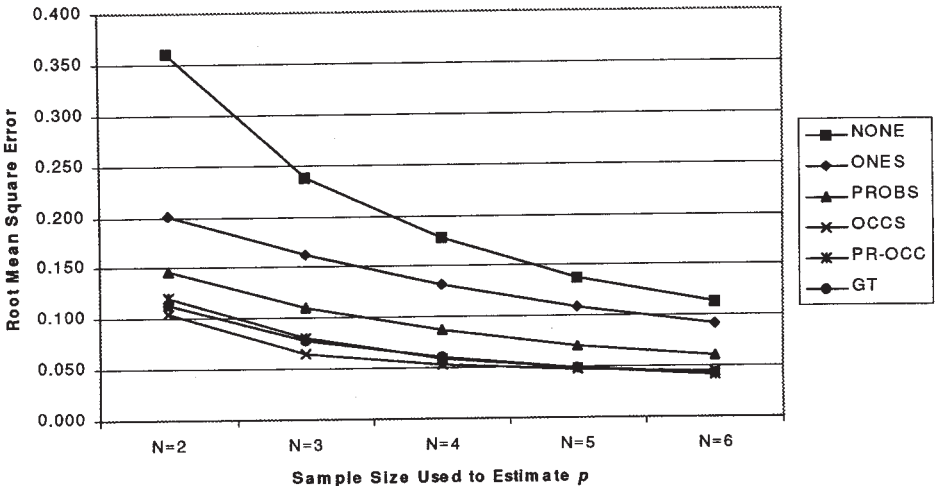

FIGURE 2 Discounting Method × Sample Size interaction for root mean square error.

tests with 3 *df* each, all *p*s < .05). A similar set of planned *t* tests showed that, at all levels of sample size, GT estimation was more accurate than the Add Ones method (5 tests with 3 *df* each, all $p < .05$), but was not significantly more accurate than the other discounting methods (15 tests with 3 *df* each, all $p > .16$). Because no other evaluated discounting procedure was significantly more accurate than GT, all additional analyses involving discounting methods focus on that well-known estimator.

Estimates of required sample size. GT discounting clearly improves the accuracy of small-sample estimation of *p*, but to what extent does it improve the accuracy of sample size estimation? To investigate this, I used true *p* from each problem-discovery database in Equation 1 to project the sample size required to achieve both 90% and 95% problem discovery for each database (true *n*, shown in Table 6). Next, I projected the required sample sizes using both the unadjusted and the GT-adjusted estimate of *p* at each level of sample size (2–6). The final step was to cal-

Table 6: Projected Sample Size Requirements for Each Problem-Discovery Database

		90% Problem Discovery		*95% Problem Discovery*	
Database	*True p*[a]	*n*	*Proportion*[b]	*n*	*Proportion*[b]
MACERR[c]	.16	14	.913	18	.957
VIRZI90[d]	.36	6	.931	7	.956
MANTEL[e]	.38	5	.905	7	.963
SAVINGS[e]	.26	8	.906	11	.961

[a]True *p* for the specified database. [b]Proportion of problems discovered at that sample size. [c]Lewis (1994) and Lewis, Henry, and Mack (1990). [d]Virzi (1990, 1992). [e]Nielsen and Molich (1990).

Table 7: Deviation From Required Sample Size for No Adjustment (NONE) and Good–Turing (GT) Discounting

Sample Size (N)	*NONE 90*	*GT 90*	*NONE 95*	*GT 95*
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. 90 = 90% problem discovery; 95 = 95% problem discovery.

culate the difference between the values of true n and the unadjusted and adjusted estimates of n. These differences appear in Table 7, with positive values indicating underestimation of true n (the expected consequence of overestimating p).

An analysis of variance (ANOVA) on the underestimation data revealed significant main effects of sample size, $F(4, 12) = 21.4$, $p = .00002$ and discounting, $F(1, 3) = 15.3$, $p = .03$, and a significant Sample Size × Goal interaction (in which the goals are 90% and 95% problem discovery), $F(4, 12) = 3.6$, $p = .04$. As the size of the sample used to estimate p increased, the magnitude of underestimation in the projected sample size decreased. GT estimation reduced the magnitude of underestimation relative to no adjustment. 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.

Discussion. Adjustment of p using discount methods provided a much more accurate estimate of the true value of p than unadjusted estimation. The best known of the methods, GT discounting, was as effective as any of the other evaluated discounting methods. GT estimation, though, still generally left 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 decreased as the size of the initial sample used to estimate p increased. For initial sample sizes of 4 to 6 participants, the magnitude of underestimation ranged from about 1.5 to 0.5 participants. Thus, final sample sizes projected from GT estimates based on initial sample sizes of 6 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 1 participant when estimating p from a 6-participant sample. (For a more comprehensive analysis of this data, see Lewis, 2000d.)

3.3. Development of Regression Equations for Predicting True p

The purpose of the third Monte Carlo experiment was to develop linear regression equations that estimate the true value of p using data available from a problem-discovery usability study.

Source data. To obtain data for the generation of the regression equations, I divided the errors in the MACERR database into four groups with mean *p* of .10, .25, .50, and .73. The use of these four groups ensured the presence of training data for the regression equations with a range of values for true *p.* A Monte Carlo procedure generated 1,000 cases for each group for each level of sample size from 2 to 6 (20,000 cases). Each case included the following measurements:

- Unadjusted estimate of *p.*
- Normalized estimate of *p.*
- Sample size.
- True value of *p.*

Regression equations. I used SYSTAT (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). Table 8 contains the resulting regression equations, the percentage of variance explained by the regression (R^2) and the observed significance level of the regression (osl).

In Table 8, *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 used to estimate *p.* All regressions (REG) except for REG3 (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 REG2, REG4, REG5, and REG6. Because the previous Monte Carlo studies indicated that the sample size plays an important role when estimating *p*, REG4, REG5, and REG6 received further evaluation in the following experiment.

Table 8: Regression Equations for Predicting True *p*

Key	*Equation*	R^2	*osl*
REG1	*truep* = –.109 + 1.017**estp*	.699	0.000
REG2	*truep* = .16 + .823**normp*	.785	0.000
REG3	*truep* = .396 + 0**n*	.000	1.000
REG4	*truep* = –.387 + 1.145**estp* + .054**n*	.786	0.000
REG5	*truep* = .210 + .829**normp* – .013**n*	.791	0.000
REG6	*truep* = –.064 + .520**estp* + .463**normp* + .017**n*	.799	0.000

Note. R^2 = percentage of variance explained by the regression; osl = observed significance level of the regression; REG = regression; *truep* = true value of *p* as predicted by the equation; *estp* = unadjusted estimate of *p* from the sample; *normp* = normalized estimate of *p* from the sample; *n* = sample size used to estimate *p*.

3.4. Evaluation of Normalization and Regression

The purpose of the fourth Monte Carlo experiment was to evaluate and compare unadjusted accuracy and the accuracy of adjustments using the normalization procedure (described previously in Section 2.1), regression equations (REG4, REG5, and REG6), and the GT procedure.

Evaluation of accuracy. An ANOVA on the *RMSE* data 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, $F(20, 60) = 20.6$, $p = .000008$. Figure 3 illustrates the interaction.

Figure 3 shows that as the sample size increased, accuracy generally increased for all estimation procedures (the main effect of sample size). A set of planned t tests evaluated, at each level of sample size, the significance of difference between no adjustment and each adjustment method and the significance of difference between GT estimation (the discounting method selected for further evaluation) and the other procedures. For sample sizes of 2, 3, and 4, all adjustment methods improved estimation accuracy of p relative to no adjustment (15 tests with 3 df each, all ps < .02). At a sample size of 5, the accuracy of REG6 was no longer significantly more accurate than no adjustment ($p > .10$). At a sample size of 6, both REG4 and REG6 (which included the unadjusted estimate of p as an element in the equation) failed to be more accurate than the unadjusted estimate

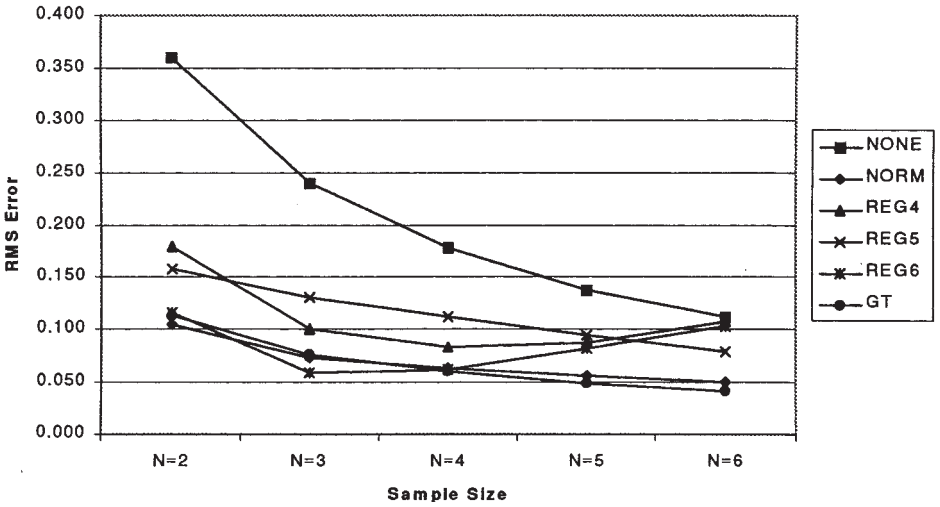

FIGURE 3 Adjustment Method × Sample Size interaction for root mean square error (RMSE).

Table 9: Deviation From Required Sample Size as a Function of Adjustment Method, Sample Size, and Discovery Goal

Sample Size (N)	*NONE 90*	*GT 90*	*NORM 90*	*NONE 95*	*GT 95*	*NORM 95*
2	5.5	2.8	–0.8	7.3	4.0	–0.8
3	4.5	1.8	–0.8	6.0	2.8	–0.8
4	4.0	1.0	–0.8	5.0	1.5	–1.0
5	3.5	0.8	–1.3	4.5	1.0	–1.5
6	2.8	0.5	–1.3	3.8	0.5	–1.5

Note. NONE = no adjustment; 90 = 90% problem discovery; GT = Good–Turing estimate; NORM = normalization estimate; 95 = 95% problem discovery.

(both *p*s > .60). Overall, GT and normalization produced the most accurate adjustments in this evaluation.

Estimates of required sample size. Using the same method as that used to evaluate sample-size projections based on GT adjustment, Table 9 shows the difference between the values of true *n* (see Table 6) and the unadjusted and adjusted (both GT and normalization) estimates of *n.* Positive values indicate underestimation of true *n;* negative values indicate overestimation.

An ANOVA revealed significant main effects of sample size, $F(4, 12) = 6.8$, $p = .004$ and adjustment, $F(2, 6) = 4.7$, $p = .05$, and significant Sample Size × Problem-Discovery Goal, $F(4, 12) = 3.2$, $p = .05$ and Sample Size × Adjustment Method, $F(8, 24) = 3.6$, $p = .007$ interactions. As the size of the sample used to estimate *p* increased, the magnitude of deviation from true *n* in the projected sample size decreased. Both GT and normalized estimation reduced the magnitude of deviation relative to no adjustment; but this magnitude decreased as the sample size increased. Although the deviation for 95% discovery consistently exceeded that for 90% discovery, as the sample size increased, the difference between the magnitude of deviation for 90% and 95% discovery decreased.

Discussion. The attempts to develop multiple regression equations for the prediction of true *p* did not fare as well as the nonregression approaches of GT and normalization. Even if the regression-based approaches had been as accurate (as measured by *RMSE*) as the nonregression approaches, the nonregression approaches would be preferable because they do not rely on statistically estimated parameters, making them solutions that have potentially greater generalizability. (For a more comprehensive analysis of this data, see Lewis, 2000c.)

As in the previous evaluation, adjustment with GT discounting consistently resulted in a slight underestimation of the required sample size. Adjustment via normalization consistently resulted in a slight overestimation of the required sample size. This result suggests the intriguing possibility that adjustment based on a com-

bination of GT and normalization procedures might yield highly accurate estimates of true p and n.

It is possible that nonlinear regression might have yielded superior results to linear regression. Exploring such alternatives, however, increases the potential for capitalizing on chance patterns in the data. If the combination of GT and normalization were to yield the expected estimation accuracy, then there would be no need to investigate solutions based on more complex regression models.

3.5. Improved Estimation Through the Combination of GT and Normalization

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 GT methods. This is important because even though the results of the previous experiment indicated that combining GT and normalization should work well, that judgment came from the averaging of average estimates—something that would be impossible for a practitioner conducting a single usability study. To control for this in this experiment, the program computed the combination estimate for each case generated via Monte Carlo simulation. Doing the computation at this level made it possible to evaluate the properties of the distribution of the combined estimate (which was not possible given the data in the previous experiment).

The adjustment methods explored in this experiment were no adjustment (NONE), GT discounting, normalization (NORM), and the combination (through simple averaging) of GT and NORM (COMB). The formula for this combination was:

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

where *truep* is the adjusted estimate of p calculated from the estimate of p derived from the participant by problem matrix (*estp*), n is the sample size used to compute the initial estimate of p, and *GTadj* is the GT adjustment to probability space, which is the proportion of the number of problems that occurred once divided by the number of different problems (see Section 2.1).

Estimates of RMSE. An ANOVA on the *RMSE* data 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, $F(12, 36) = 36.5$, $p = .0000002$. Figure 4 illustrates the interaction.

Figure 4 shows that as the sample size increased, accuracy generally increased for all estimation procedures (the main effect of sample size). The lines for NORM, GT, and COMB almost overlaid one another, with COMB having slightly less

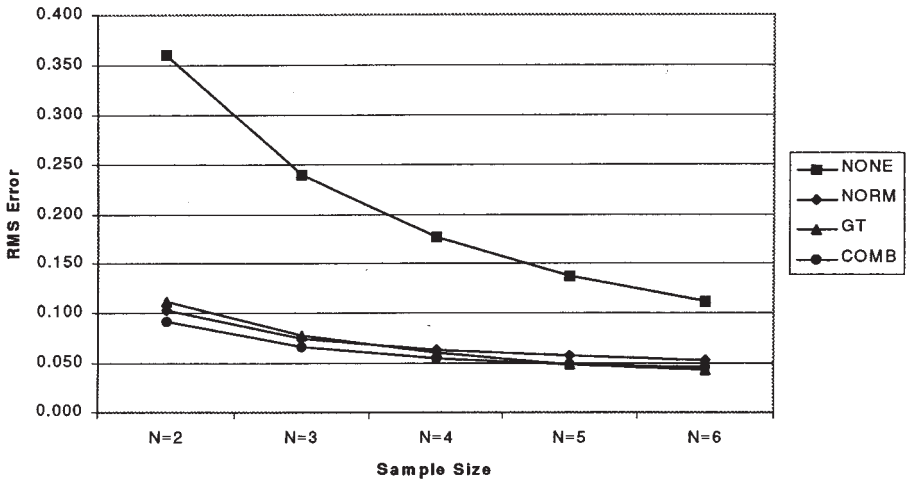

FIGURE 4 Root mean square error (RMSE) as a function of sample size and adjustment method.

RMSE than either GT or NORM. A set of planned *t* tests showed that estimates based on this combination resulted in significantly lower *RMSE* than unadjusted estimates for all sample sizes (all *p*s < .02). A similar set of *t* tests showed that none of the *RMSE* differences among GT, normalization, or their combination were significant (all *p*s > .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*.

Estimates of required sample size. I conducted an ANOVA on the deviations from required sample size for unadjusted *p*, GT discounting, normalization, and the averaging of GT and normalization, treating databases as subjects in a within-subjects design with independent variables of sample size, type of 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 × Adjustment Type interaction, $F(3, 9) = 3.9$, $p = .05$ and the Sample Size × Adjustment Type interaction, $F(12, 36) = 2.9$, $p = .006$. In the Discovery Goal × Adjustment Type interaction, the underestimation of the required sample size for 95% was generally greater than for 90%, except for the normalization adjustment type, which had equal deviation for both levels of discovery goal (see Figure 5). The Sample Size × 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*, GT estimates, and the combination estimate, but not for estimates based on normalized *p* (see Figure 6).

As expected, across all sample sizes the GT estimate tended to underestimate the required sample size and the normalized estimate tended to overestimate the re-

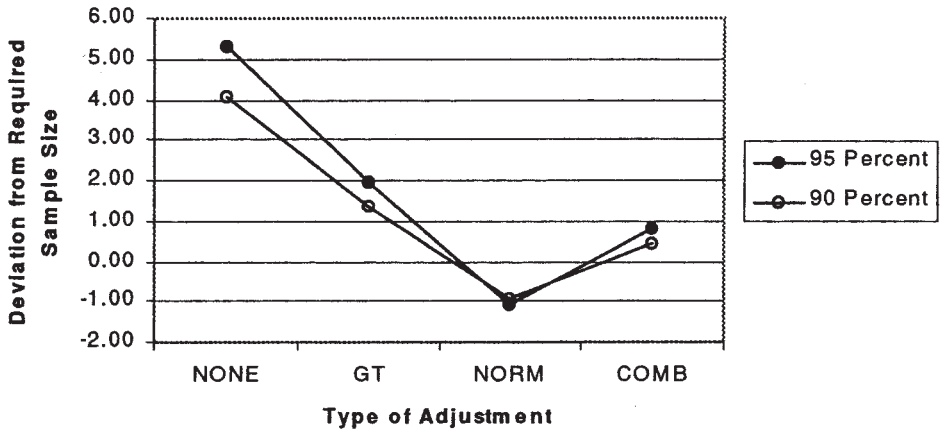

FIGURE 5 Discovery Goal × Adjustment Method interaction.

quired 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 true n. For estimates of p computed from initial sample sizes of 2 and 3 participants, the mean underestimations of true n projected from combination-adjusted estimates of p were 2 participants and 1 participant, respectively.

Variability of combination estimate. The preceding analyses focused on the means of various distributions. The analyses in this section address the distribution of p after adjustment with the normalization and GT combination procedure (spe-

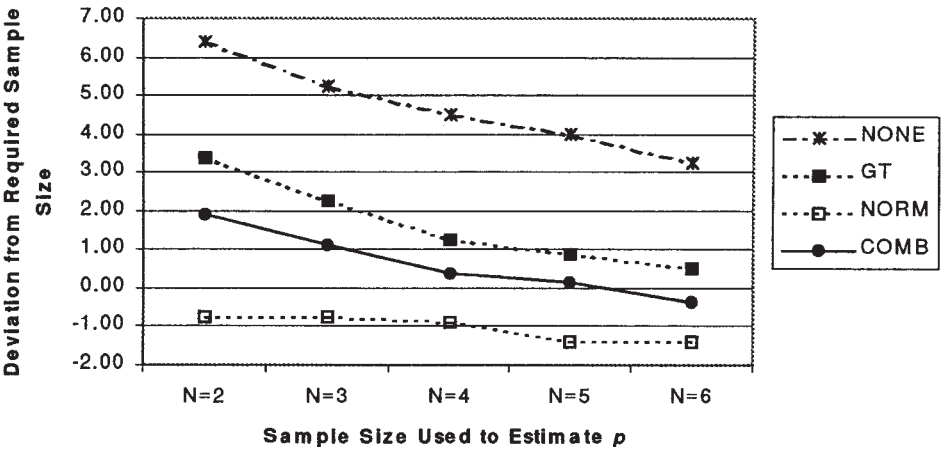

FIGURE 6 Sample Size × Adjustment Method interaction.

Table 10: Distribution of Deviations From True *p* as a Function of Sample Size for Combination Adjustment

Percentile	*N* = 2	*N* = 3	*N* = 4	*N* = 5	*N* = 6
1st	–0.10	–0.09	–0.09	–0.09	–0.08
5th	–0.06	–0.06	–0.06	–0.06	–0.06
10th	–0.04	–0.04	–0.05	–0.05	–0.05
25th	0.00	–0.01	–0.02	–0.03	–0.03
50th	0.04	0.02	0.01	0.00	–0.01
75th	0.10	0.06	0.04	0.03	0.02
90th	0.14	0.10	0.07	0.05	0.04
95th	0.17	0.12	0.09	0.07	0.06
99th	0.24	0.16	0.12	0.10	0.09

Note. The 25th and 75th percentiles define the interquartile range. The 5th and 95th percentiles define the 90% range

cifically, the deviation of this adjustment from true *p*). These analyses will help practitioners understand the variability of this distribution so they can take this variability into account when planning problem-discovery usability studies.

Table 10 provides the average deviations from true *p* (collapsed across databases) as a function of sample size for adjustments using Equation 2 (the combination adjustment). Table 11 shows the corresponding magnitude of interquartile and 90% ranges for these deviations. Relatively smaller deviations from true *p* and relatively smaller ranges are indicators of better accuracy. The results shown in the tables indicate that increasing the sample size used to estimate *p* decreased both the magnitude and variability of deviation from true *p.*

Discussion. GT estimation generally left the estimate of *p* slightly inflated, leading to some underestimation of required total sample sizes when projecting from the initial sample. Estimating *p* with the normalization procedure had the same accuracy as GT estimation but tended to underestimate true *p,* leading to some overestimation of required sample sizes when projecting from an initial sample. Averaging the GT and normalization estimates (combination adjustment) provided a highly accurate estimate of true *p* from very small samples,

Table 11: Widths of Interquartile and 90% Ranges as a Function of Sample Size for Combination Adjustment

Sample Size (N)	*Interquartile Range*	*90% Range*
2	.10	.23
3	.07	.18
4	.06	.15
5	.05	.13
6	.05	.13

which in turn led to highly accurate estimates of required sample sizes for specified problem-discovery goals. These estimates appear to be accurate enough that a practitioner should be able to make an initial projection from a combination-adjusted estimate of p using a sample with as few as 2 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 2 or 3 participants (which should generally overestimate the required sample size slightly). 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 reestimate p and project the revised sample size requirement. (For a more comprehensive analysis of this data, see Lewis, 2000c.)

The apparent trends in Figure 6 indicated that it might not be wise to use the combination or normalization approaches when the sample size exceeds 6 participants. At 6 participants, normalization continued to underestimate p, and the combination approach began to slightly underestimate p. The GT approach appeared 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 was not clear from the data at what sample size a practitioner should abandon GT and move to the unadjusted estimate of p.

3.6. Evaluation of Best Procedures for Sample Sizes From 2 to 10

One goal of the final Monte Carlo experiment was to replicate the previous investigation of a variety of approaches (NORM, REG2, REG5, GT, COMB) for adjusting observed estimates of p to bring them closer to true p using sample sizes from 2 to 10 participants for the initial estimate of p. In particular, would the combination approach continue to provide accurate estimates of true p for sample sizes from 7 to 10 participants?

Another goal was to investigate the extent to which inaccuracy in estimating problem-discovery sample size using these methods affects the true proportion of discovered problems. The previous investigations assessed the deviation from the sample size required to achieve 90% and 95% problem-discovery goals but did not assess the magnitude of deviation from the problem-discovery goals caused by overestimating or underestimating the required sample size.

Estimates of p. An ANOVA (within-subjects ANOVA treating problem-discovery databases as subjects), conducted on the problem-discovery rates (p) for each of the six estimation methods at each of the nine levels of sample size, revealed a significant main effect for type of adjustment, $F(8, 24) = 92.6$, $p = .00000015$; a significant main effect of sample size, $F(5, 15) = 138.5$, $p = .00000015$; and a significant Ad-

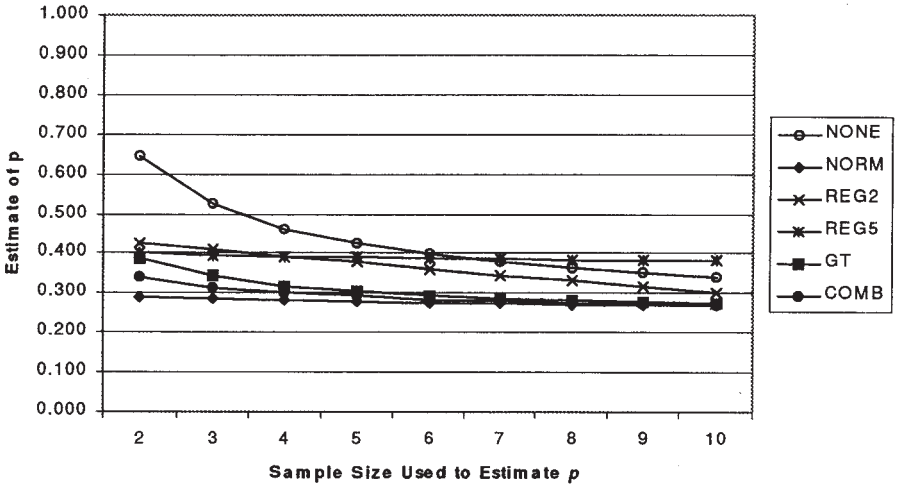

FIGURE 7 Adjustment Method × Sample Size interaction for problem-discovery rate.

justment Type × Sample Size interaction, $F(40, 120) = 72.8$, $p = .0001$. Figure 7 illustrates the interaction.

Estimates of RMSE. An ANOVA conducted on *RMSE* revealed a significant main effect for type of adjustment, $F(8, 24) = 120.6$, $p = .00000004$; a significant main effect of sample size, $F(5, 15) = 27.2$, $p = .0000006$; and a significant Adjustment Type × Sample Size interaction, $F(40, 120) = 32.9$, $p = .0000001$ (see Figure 8).

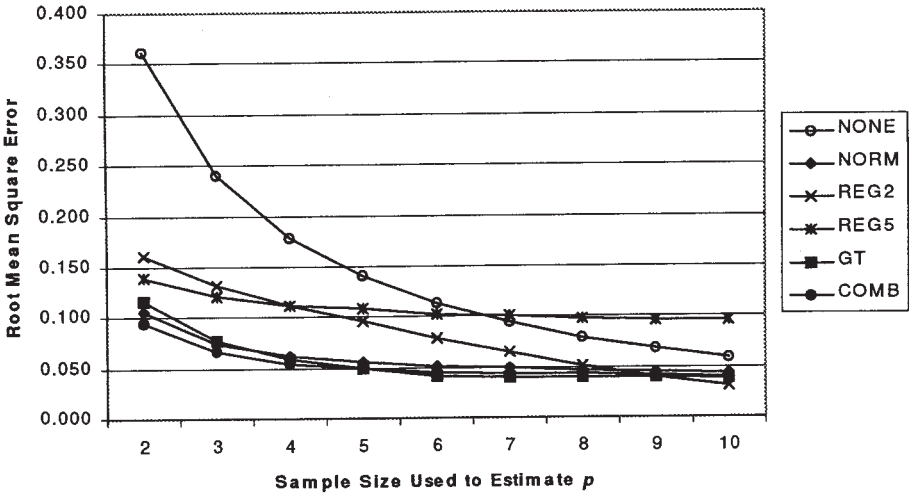

FIGURE 8 Adjustment Method × Sample Size interaction for root mean square error.

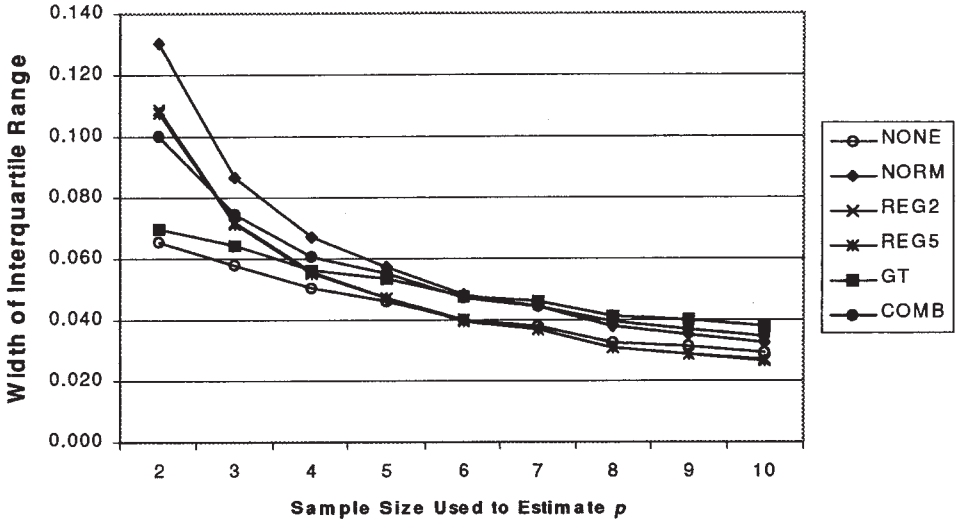

FIGURE 9 Adjustment Method × Sample Size interaction for interquartile range.

Estimation variability. The *interquartile range* is the size of the interval that contains the central 50% of a distribution (the range from the 25th to the 75th percentile). The smaller this range, the less variable is the distribution. An ANOVA conducted on interquartile ranges revealed a significant main effect for type of adjustment, $F(8, 24) = 109.1$, $p = .0000003$; a significant main effect of sample size, $F(5, 15) = 10.3$, $p = .0002$; and a significant Adjustment Type × Sample Size interaction, $F(40, 120) = 69.8$, $p = .00000001$. Figure 9 illustrates this interaction.

Estimates of required sample size. I conducted a within-subjects ANOVA on the deviations from required sample sizes, treating problem-discovery databases as subjects. The independent variables were adjustment method (NONE, NORM, REG2, REG5, GT, COMB), sample size used to estimate p (2–10), and problem-discovery goal (90%, 95%). The analysis indicated the following significant effects:

- Main effect of adjustment method, $F(5, 15) = 4.1$, $p = .015$.
- Main effect of sample size, $F(8, 24) = 3.8$, $p = .005$.
- Adjustment Method × Discovery Goal interaction, $F(5, 15) = 3.9$, $p = .019$.
- Adjustment Method × Sample Size interaction, $F(40, 120) = 3.2$, $p = .0000006$.
- Adjustment Method × Sample Size × Problem-Discovery Goal interaction, $F(40, 120) = 2.0$, $p = .002$.

Collapsed over discovery goal and sample size, the deviations from true n for the various methods of adjustment (main effect of adjustment method) were underestimations of 3.7 for NONE, 3.1 for REG5, 2.5 for REG2, 0.7 for GT, and overestimations of 1.3 for NORM and 0.1 for COMBO. The pattern for the main ef-

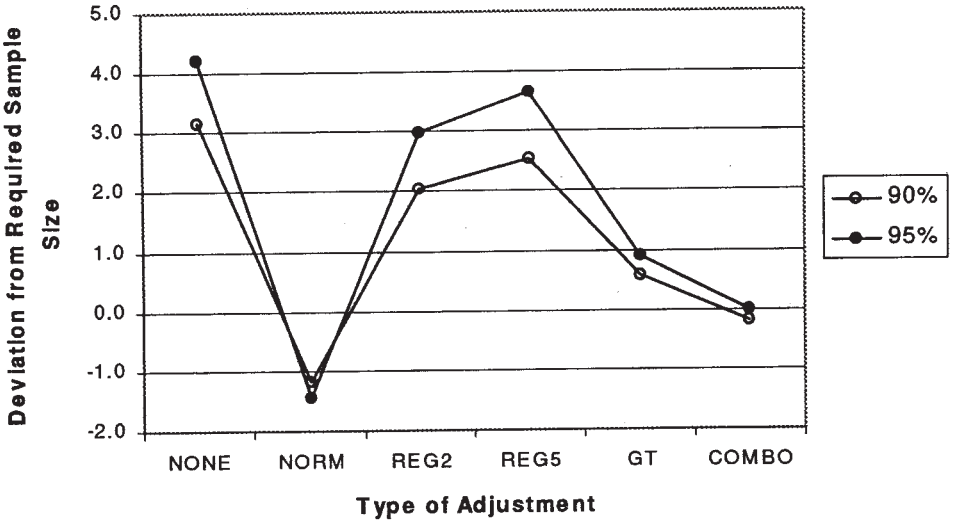

FIGURE 10 Deviation from required sample size as a function of adjustment method and discovery goal.

fect of sample size was a steady decline in overestimation from 3.1 at a sample size of 2 to 0.3 at a sample size of 10.

Figures 10 and 11 show the Adjustment Method × Discovery Goal interaction and the Adjustment Method × Sample Size interaction. The pattern for the Adjustment Method × Sample Size × Discovery Goal interaction was very similar to that

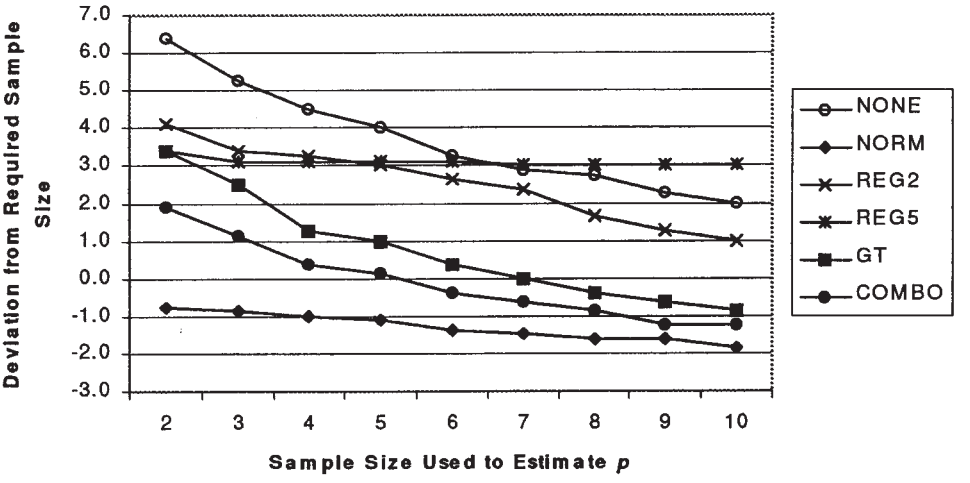

FIGURE 11 Deviation from required sample size as a function of adjustment method and sample size.

for the Adjustment Method × Sample Size interaction, with the difference between deviations for 90% and 95% discovery declining as a function of sample size. (Positive values indicate underestimation of the required sample size; negative values indicate overestimation.)

Deviation from problem-discovery goal. Because the fundamental problem in underestimating a required sample size is the failure to achieve a specified problem-discovery goal, I also conducted a within-subjects ANOVA on the deviations from specified problem-discovery goals of 90% and 95% discovery, treating problem-discovery databases as subjects. The independent variables were adjustment method (NONE, NORM, REG2, REG5, GT, COMB), sample size used to estimate p (2–10), and problem-discovery goal (90%, 95%). The dependent variable was the difference between the specified problem-discovery goal and the magnitude of problem discovery for the projected sample sizes calculated in the previous section. Because sample sizes are discrete rather than continuous variables, there are cases in which a sample size that is 1 participant smaller than the sample size required to achieve (meet or exceed) a specified problem-discovery goal is actually much closer to the goal (although just under it) than the required sample size. In other words, underestimating the sample size requirement for a specified problem-discovery goal might, in many cases, still allow a practitioner to come very close to achieving the specified goal.

The analysis indicated the following significant effects:

- Main effect of adjustment method, $F(5, 15) = 13.6$, $p = .00004$.
- Main effect of sample size, $F(8, 24) = 15.5$, $p = .0000001$.
- Adjustment Method × Sample Size interaction, $F(40, 120) = 12.5$, $p = .005$.
- Sample Size × Problem-Discovery Goal interaction, $F(8, 24) = 3.8$, $p = .006$.

The pattern for the main effect of adjustment method was an underestimation of .113 for no adjustment, underestimation of .068 for REG5, underestimation of .053 for REG2, overestimation of .020 for NORM, underestimation of .008 for GT, and overestimation of .008 for COMBO. The pattern for the main effect of sample size was a steady decline in the magnitude of underestimation from .100 for a sample size of 2 to .008 for a sample size of 10. The basis for the Sample Size × Discovery Goal interaction was that deviation tended to be smaller for 90% discovery for sample sizes smaller than 6, with the opposite trend for sample sizes greater than 6. Figure 12 illustrates the interaction between adjustment method and sample size.

Iterative sample size estimation strategy using the combination adjustment. One practical application for the use of the combined estimator is to allow usability practitioners to estimate their final sample size requirement from their first few participants. To do this, practitioners must keep a careful record of which

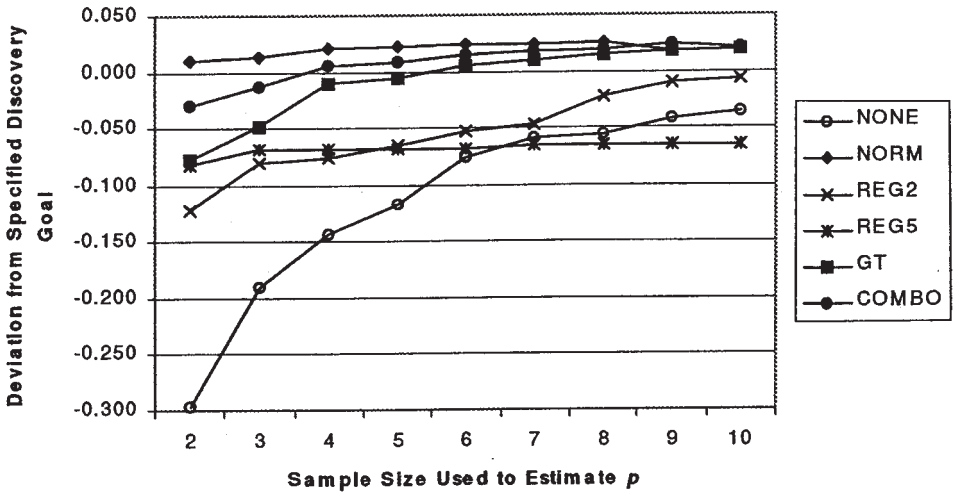

FIGURE 12 Deviation from specified discovery goal as a function of adjustment method and sample size.

participants experience which usability problems to enable calculation of an unadjusted initial estimate of p that they then adjust using the combined estimator.

Suppose, for example, that a practitioner is conducting a study on a product with problem-discovery characteristics similar to MACERR (true p of .16), and the practitioner has set a goal of discovering 90% of the problems. Referring to Table 5, the initial unadjusted estimate of p calculated from the first 2 participants would be, on average, .568. Adjusting this initial estimate with the combined procedure results in a value for p of .218. Using Equation 1 to project the sample size until the estimated proportion of discovery exceeds .900 yields a preliminary sample size requirement of 10 participants. After the practitioner runs 2 more participants toward the goal of 10 participants, he or she can recalculate the adjusted value of p to be .165 and can project the final required sample size to be 13 participants. As shown in Table 6, the sample size actually required to exceed a problem-discovery proportion of .900 is 14 participants. With 13 participants, the true proportion of problem discovery is .896, which misses the specified problem-discovery target goal by only .004—less than 0.5%.

Table 12 shows the outcome of repeating this exercise for each database and for each investigated problem-discovery goal. These outcomes show that following this procedure leads to very accurate estimates of sample sizes and very little deviation from problem-discovery goals—a remarkable outcome given the differences in the usability studies that produced these problem-discovery databases. The mean deviation from problem-discovery goal across cases was .006—overachievement by just over 0.5%.

A similar evaluation conducted for lower, less-aggressive problem-discovery goals (70%, 75%, and 80%) revealed a similar pattern of results. The overall mean

Table 12: Problem-Discovery Outcomes Achieved by Using Combination Adjustment to Project a Final Sample Size from Initial Sample Sizes of 2 and 4 Participants

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

Note. True *p* = value of *p* estimated from the entire database; Est *p* | *N* = 2 = unadjusted estimate of *p* given a sample size of 2 participants; Comb *p* | *N* = 2 = combination-adjusted estimate of *p* given a sample size of 2; *n* | *N* = 2 = projected sample size requirement given the combination-adjusted estimate of *p* estimated from a sample size of 2; Est *p* | *N* = 4 = unadjusted estimate of *p* given a sample size of 4 participants; Comb *p* | *N* = 4 = combination-adjusted estimate of *p* given a sample size of 4; *n* | *N* = 4 = projected sample size requirement given the combination-adjusted estimate of *p* estimated from a sample size of 4; True *n* = sample size requirement projected from True *p*.

[a]Lewis (1994) and Lewis, Henry, and Mack (1990). [b]Virzi (1990, 1992). [c]Nielsen and Molich (1990).

deviation from the discovery goal was overachievement of .021 (2.1%). The greatest mean deviation across the problem-discovery goals occurred for the VIRZI90 database (overachievement of .081), with smaller deviations for the other three databases (.000, .006, and –.004 for MACERR, MANTEL, and SAVINGS, respectively). Averaging over databases indicated overachievement of .047 and .028 for 70% and 75% discovery, and underachievement of .012 for 80% discovery. In all cases, the projected sample size given, $n = 4$, was within one participant of true n, with an overall average deviation of estimated sample size from true sample size requirement of –0.17 participants.

3.7. Discussion

The decision about which adjustment procedure or procedures to use should take into account both the central tendency and variability of distributions created by applying the adjustment procedure or procedures. A measure that produces mean estimates close in value to true p will generally be more accurate in the long run. A measure with low variability is less likely to produce an extreme outlier in any single study. Usability practitioners do not typically conduct large-scale studies, however, so it is important to balance benefits associated with the statistical long run with the benefits associated with reducing the risk of encountering an extreme outlier.

The regression equations (REG2 and REG5) tended to be less variable than other adjustment procedures, but their accuracy was very poor relative to all other adjustment procedures. The accuracy of REG2 improved as a function of the sample size used to estimate p, but the accuracy of REG5 did not. Their relatively poor performance removes these regression equations from consideration as a recommended method for adjusting p.

The remaining procedures (NORM, GT, and COMB) showed similar patterns to one another for deviations from true p. When estimating p with smaller sample sizes (2–4 participants), the curves for these three measures showed some separation, with the differences diminishing and the curves converging as the size of the sample used to estimate p increased. For these measures, especially at small sample sizes, the normalization procedure appeared to produce the best results and the combined estimator produced the second-best results.

The measures of interquartile range, however, indicated that the estimates produced by the normalization procedure were much more variable than those produced by the GT or the combined estimator. The results for the *RMSE* (which take both central-tendency accuracy and variability into account) showed that all three measures had essentially equal accuracy at all levels of sample size from 2 to 10. As expected, the variability of all measures decreased as a function of the sample size used to estimate p.

Considering all the information, the combined estimator seemed to provide the best balance between central-tendency accuracy and lower variability, making it the preferred adjustment procedure. What really matters, though, is the extent to

which the adjustment procedure leads to accurate sample size estimation and achievement of specified problem-discovery goals.

The analyses of underestimation of required sample sizes and deviation from problem-discovery goals also support the use of the combined estimator. After averaging across all problem-discovery databases (MACERR, VIRZI90, MANTEL, and SAVINGS), both problem-discovery goals (90%, 95%), and sample sizes from 2 to 10, the accuracy of sample size estimation with the combined estimator was almost perfect, overestimating the required sample size by only 0.1 participant on average. The results were similar for mean deviation from problem-discovery goal. The magnitude of deviation was about the same for the combined estimator and the GT estimator. On average, however, the combined estimator tended to slightly overachieve the discovery goal, whereas the GT estimator tended to slightly underachieve the discovery goal.

With one exception, the patterns of results for the significant interactions supported the unqualified use of the combined estimator. The exception was the interaction between adjustment method and the sample size used to estimate p. As shown in Figure 11, at sample sizes of 2 and 3 the combined estimator tended to underestimate the required sample size, but the normalization procedure tended to overestimate it. The same interaction for the deviation from the discovery goal, however, indicates that the consequence of this underestimation of the required sample size was slight, even when the sample size used to estimate p was only 2 participants (in which case the underachievement was, on average, 3%). This does suggest some need on the part of practitioners to balance the cost of additional participants against their need to achieve a specific problem-discovery goal. If the former is more important, then the practitioner should use the combined estimator. If the latter is more important and the practitioner is estimating p once (not iteratively) from a very small sample size, then it would be reasonable to use the more conservative normalization procedure. (For a more comprehensive analysis of this data, see Lewis, 2000a.)

3.8. Summary of Analyses and Results

A series of Monte Carlo simulations provided evidence that the average of a normalization procedure and GT discounting produces highly accurate estimates of usability problem-discovery rates from small sample sizes. The motivation for conducting the research was the observation that unadjusted estimates of p derived from small-sample usability studies have a bias in a direction that would lead usability practitioners to believe that their studies have been more effective than the data really warrants (Hertzum & Jacobsen, this issue).

The simulations allowed investigation of three broad classes of methods for reducing initial estimates of p from small-sample usability studies: discounting procedures, regression, and normalization. Accuracy assessments of the various procedures (using *RMSE* as the measure of accuracy) indicated that (a) GT discounting (the best known of the discounting methods investigated) was as accurate as or

more accurate than the other discounting procedures, (b) normalization was also a very accurate adjustment procedure, and (c) regression did not produce very satisfactory adjustments.

An unexpected (but fortuitous) result was that two of the most accurate adjustment procedures—GT discounting and normalization—had residual biases in opposite directions and of about equal magnitude. Investigation of an adjustment procedure based on the combination of these methods indicated that this approach provided the most satisfactory adjustments (high accuracy and low variability). Using this combined procedure to adjust initial estimates of p (derived from several published large-sample usability studies) resulted in highly accurate estimates of required sample sizes.

4. CONCLUSIONS AND RECOMMENDATIONS FOR PRACTITIONERS

- The overestimation of p from small-sample usability studies is a real problem with potentially troubling consequences for usability practitioners.
- It is possible to compensate for the overestimation bias of p calculated from small-sample usability studies.
- The combined normalization and GT estimator (Equation 2) is the best procedure for adjusting initial estimates of p calculated from small samples (2 to 10 participants).
- If (a) the cost of additional participants is low, (b) the sample size used to estimate p is very small (2 or 3 participants), and (c) it is very important to achieve or exceed specified problem-discovery goals, then practitioners should use the normalization procedure to adjust the initial estimate of p.
- Practitioners can obtain accurate sample size estimates for problem-discovery goals ranging from 70% to 95% by making an initial estimate of the required sample size after running 2 participants, then adjusting the estimate after obtaining data from another 2 (total of 4) participants.

REFERENCES

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

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

Connell, I. W., & Hammond, N. V. (1999). Comparing usability evaluation principles with heuristics: Problem instances vs. problem types. In M. A. Sasse & C. Johnson (Eds.), *Proceedings of INTERACT '99—Human–Computer Interaction* (Vol. 1, pp. 621–629). Edinburgh, Scotland: International Federation for Information Processing.

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

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

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

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

Lewis, J. R. (2000a). *Evaluation of problem discovery rate adjustment procedures for sample sizes from two to ten* (Tech. Rep. No. 29.3362). Raleigh, NC: IBM. Available from the author.
Lewis, J. R. (2000b). *Overestimation of p in problem discovery usability studies: How serious is the problem?* (Tech. Rep. No. 29.3358). Raleigh, NC: IBM. Available from the author.
Lewis, J. R. (2000c). *Reducing the overestimation of p in problem discovery usability studies: Normalization, regression, and a combination normalization/Good–Turing approach* (Tech. Rep. No. 29.3361). Raleigh, NC: IBM. Available from the author.
Lewis, J. R. (2000d). *Using discounting methods to reduce overestimation of p in problem discovery usability studies* (Tech. Rep. No. 29.3359). Raleigh, NC: IBM. Available from the author.
Lewis, J. R. (2000e). *Validation of Monte Carlo estimation of problem discovery likelihood* (Tech. Rep. No. 29.3357). Raleigh, NC: IBM. Available from the author.
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, International Federation for Information Processing.
Manning, C. D., & Schutze, H. (1999). *Foundations of statistical natural language processing.* Cambridge, MA: MIT Press.
Nielsen, J. (1992). Finding usability problems through heuristic evaluation. In *Conference Proceedings on Human Factors in Computing Systems—CHI '92* (pp. 373–380). Monterey, CA: ACM.
Nielsen, J., & Landauer, T. K. (1993). A mathematical model of the finding of usability problems. In *Conference Proceedings on Human Factors in Computing Systems—CHI '93* (pp. 206–213). New York: ACM.
Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Conference Proceedings on Human Factors in Computing Systems—CHI '90* (pp. 249–256). New York: ACM.
Norman, D. A. (1983). Design rules based on analyses of human error. *Communications of the ACM, 4,* 254–258.
Pedhazur, E. J. (1982). *Multiple regression in behavioral research: Explanation and prediction.* Fort 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.

APPENDIX

MACERR Problem-Discovery Database

The MACERR database contains discovery information for 145 usability problems uncovered by observation of 15 participants, with p estimated to be .16 (Lewis, Henry, & Mack, 1990). The first column in the database table is the problem identification number. The next 15 columns represent the observation of the experience of each participant with that problem, with a 0 indicating that the participant did not experience the problem and a 1 indicating that the participant did experience that problem. The last column is the modal impact rating for the problem across participants experiencing the problem, using the behavioral rating scheme described in Lewis (1994). The criteria for the impact ratings were

Prob	*P1*	*P2*	*P3*	*P4*	*P5*	*P6*	*P7*	*P8*	*P9*	*P10*	*P11*	*P12*	*P13*	*P14*	*P15*	*Impact*
1	0	1	1	1	0	1	1	1	1	1	1	0	0	1	1	2
2	1	0	1	0	0	0	1	1	1	1	1	0	0	1	0	2
3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
4	1	1	1	1	0	0	1	0	1	1	1	0	1	1	0	4
5	1	0	0	1	1	0	1	0	0	0	1	0	1	0	1	3
6	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	4
7	1	0	0	1	1	1	0	0	1	0	1	0	0	0	1	3
8	1	0	0	0	1	0	0	0	1	1	0	0	0	0	0	3
9	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	3
10	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	2
11	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
12	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3
13	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	2
14	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	2
15	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	4
16	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	3
17	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	3
18	1	1	1	1	0	1	1	0	1	1	1	0	1	1	0	2
19	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	1
20	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	2
21	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	3
22	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	3
23	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
24	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	3
25	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3
26	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	3
27	1	0	1	1	0	0	1	1	1	1	0	1	1	0	1	2
28	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	2
29	1	0	0	0	0	1	1	1	1	1	0	1	0	0	1	3
30	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
31	0	0	1	1	0	1	1	0	1	0	0	1	0	0	0	1
32	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	2
33	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	3
34	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	3
35	0	0	0	0	1	1	0	0	1	0	0	0	1	0	0	1
36	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	4
37	1	0	1	1	0	0	0	1	0	1	0	0	1	0	0	4
38	0	1	0	0	0	0	1	0	0	1	0	0	1	0	0	2
39	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2
40	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2
41	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	2
42	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	1
43	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
44	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2
45	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2
46	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0	1
47	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
48	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2
49	1	0	1	0	0	0	0	0	0	0	0	0	0	0	1	2
50	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2

(*continued*)

Prob	*P1*	*P2*	*P3*	*P4*	*P5*	*P6*	*P7*	*P8*	*P9*	*P10*	*P11*	*P12*	*P13*	*P14*	*P15*	*Impact*
51	1	1	1	1	0	0	1	0	1	0	0	0	0	0	1	1
52	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
53	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	3
54	1	0	1	1	0	1	0	0	0	0	0	0	0	0	1	2
55	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3
56	0	1	0	1	0	1	0	1	1	1	0	0	0	0	1	2
57	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3
58	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	3
59	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2
60	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	2
61	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2
62	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2
63	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
64	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	3
65	0	1	1	0	0	0	1	0	0	0	0	0	0	0	1	1
66	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	2
67	1	0	0	1	1	0	0	0	0	0	0	0	0	0	1	2
68	1	0	0	0	0	0	0	1	1	0	0	0	0	0	1	1
69	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	4
70	1	0	0	0	0	0	0	1	0	0	0	0	0	0	1	2
71	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2
72	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	1
73	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	2
74	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	3
75	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	4
76	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
77	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	2
78	0	1	1	1	1	0	1	0	0	0	0	0	0	0	0	1
79	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
80	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0
81	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
82	0	1	1	1	0	0	0	0	0	1	0	0	0	0	0	1
83	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1
84	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0
85	0	0	1	0	0	1	1	0	1	0	0	0	0	0	0	1
86	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
87	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1
88	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	2
89	0	0	0	1	0	1	0	1	1	0	0	0	0	0	0	4
90	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	2
91	0	1	0	1	1	1	1	0	0	0	0	0	0	0	0	1
92	1	1	0	1	1	1	0	1	0	1	0	0	0	0	0	3
93	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	2
94	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	2
95	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	3
96	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2
97	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	2
98	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	4
99	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2
100	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1

(*continued*)

Prob	*P1*	*P2*	*P3*	*P4*	*P5*	*P6*	*P7*	*P8*	*P9*	*P10*	*P11*	*P12*	*P13*	*P14*	*P15*	*Impact*
101	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	2
102	0	1	1	0	1	0	0	1	0	0	0	0	0	0	0	3
103	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2
104	0	1	1	0	1	1	0	1	1	1	0	0	0	0	0	1
105	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	1
106	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	3
107	1	1	1	1	1	1	0	0	1	0	0	0	0	0	0	3
108	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	3
109	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
110	0	1	0	0	1	0	0	0	1	0	0	0	0	0	0	1
111	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
112	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	2
113	0	0	0	1	1	1	0	0	1	0	0	0	0	0	0	2
114	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	3
115	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2
116	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	2
117	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2
118	0	0	0	0	0	1	0	0	0	1	0	0	0	0	0	2
119	1	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1
120	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	1
121	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	1
122	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	2
123	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
124	1	0	0	0	1	1	0	0	0	0	0	0	0	0	0	2
125	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	3
126	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2
127	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2
128	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	2
129	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
130	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1
131	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
132	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	1
133	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
134	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
135	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
136	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
137	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1
138	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	2
139	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1
140	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	2
141	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	1
142	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
143	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
144	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	2
145	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1

Note. Prob = problem identification number; P = participant; Impact = modal impact rating for the problem across participants; 0 = participant did not experience the problem; 1 = participant did experience the problem.

1. *Scenario failure.* Participants failed to successfully complete a scenario if they either requested help to complete it or produced an incorrect output (excluding minor typographical errors).
2. *Considerable recovery effort.* The participant either worked on error recovery for more than a minute or repeated the error within a scenario.
3. *Minor recovery effort.* The participant experienced the problem only once within a scenario and required less than a minute to recover.
4. *Inefficiency.* The participant worked toward the scenario's goal but deviated from the most efficient path.

In the analyses in this article I did not use the impact ratings, but I have provided them for any researcher who needs them. The database is available in electronic form in Lewis (2000e).