Legitimate Use of Small Samples in Usability Studies: Three Examples

TR 54.594
June 1, 1991

James R. Lewis

IBM Human Factors

Boca Raton, Florida

Abstract

Efficiency is an important consideration in the design of industrial usability studies. One way to reduce the cost of a usability study is to reduce its sample size. However, sample size reduction will improve the efficiency of the study only if it does not lead to a significant loss of information. Three legitimate uses of small samples are described in this report: (1) problem discovery studies, (2) criterion-based usability studies in which the criterion is clearly met, and (3) binomial confidence intervals to describe usability defect rates. In these situations, relatively small samples are often adequate to meet the goals of the usability engineer. The risks of using small samples in these situations are also discussed.

ITIRC Keywords

Usability study
Sample size
Confidence interval
Criterion-based
Problem discovery
Binomial confidence interval
Usability defect rate

Contents

Introduction

To be as cost-effective as possible, industrial usability studies must be efficient (Lewis, 1990; Virzi, 1990). A study conducted with a small sample is less costly than one with a large sample. However, sample size reduction will improve the efficiency of the study only if it does not lead to a significant loss of information. Three legitimate uses of small samples are: (1) problem discovery studies, (2) criterion-based usability studies in which the criterion is clearly met, and (3) binomial confidence intervals to describe usability defect rates. In these situations, relatively small samples are often adequate to meet the goals of the usability engineer. The risks of using small samples in these situations are also discussed.

Problem Discovery Studies

The primary goal of a problem discovery study is to find system usability problems by watching participants try to use the system. Problems are usually categorized along two or more dimensions, such as frequency and impact. This information is used to guide system development and to prioritize the effort expended to correct usability defects. It is the experience of many usability engineers that the additional information gained by running more than five or six participants is small relative to that obtained from the initial participants. A recent empirical study (Virzi, 1990) and a mathematical model based on the binomial probability distribution (Lewis, 1990) both support this experience. Lewis (1990) published tables based on the binomial probability distribution (Bradley, 1976) to help human factors engineers and other usability researchers to determine appropriate sample sizes for their usability studies. In most cases, the diminishing returns after the first few participants suggest that the sample sizes for problem discovery studies should not exceed ten participants. In many cases, particularly if the testing is iterative, sample sizes as small as three participants will be adequate (Lewis, 1982; Lewis, 1990).

Table 1 illustrates the diminishing returns predicted by the binomial probability formula. The data in Table 1 were generated assuming a .50 average likelihood of problem discovery. This likelihood is consistent with the rate observed in many usability studies. (For example, the likelihood in the study reported by Virzi, 1990 was .36.) The table shows the following:

1. If you don't run any participants, you can't find any problems. (With a sample size of 0, the percentage of problems discovered is 0.0).
2. Because, in this example, the base rate of problem occurrence is .5, on the average 50 percent of the problems that can be discovered will be discovered with the first participant. This means that the improvement in problem discovery from no participants to one participant is 50 percent.
3. The improvement from one participant to two participants is 25 percent. The improvement from two participants to three participants is 12.5 percent, and the improvement from three participants to four participants is 6.3 percent. The additional gain expected from each additional participant is clearly diminishing.
4. The improvement from four participants to an infinite number of participants is 6.2 percent -- a little less than the gain from three to four participants.

TABLE 1. Binomial Probability Distribution: Diminishing Returns

Sample Size	Likelihood of Problem Discovery	Percentage of Problems Discovered
0	0.000	0.0
1	0.500	50.0
2	0.750	75.0
3	0.875	87.5
4	0.938	93.8
…		
Infinite	1.000	100.0

Thus, Table 1 illustrates the legitimacy of using small samples in problem discovery usability studies, at least when the likelihood of problem occurrence is not too small. However, when the likelihood of problem occurrence is small, then the product must be well designed and therefore demands less application of usability resources than other products.

The risk associated with using a small sample for a problem discovery study has been described by Lewis (1990, p. 6):

> It is important to understand the risks as well as the gains if a small sample is used in an observational usability study. . . . The only characteristic that can cause a problem to turn up early is its frequency of occurrence in the user population. . . . Therefore, there is always a risk that a low-frequency but important problem will go undetected with a small sample. However, in many circumstances (especially if the problems are not expected to have a great cost and testing is iterative), studying a small sample is reasonable and efficient.

The importance of iterative testing to the development of usable products is well known (Gould, 1988). For efficient iterative problem discovery studies, I recommend the following iterative testing cycle:

1. Start with an expert evaluation or one-participant pilot study to uncover the majority of high-frequency problems (J. H. McTyre, personal communication, February 15, 1991). As many of these problems as possible should be corrected before starting the iterative cycles with Step 2. All unresolved problems should be listed and be carried to Step 2.
2. Watch a sample of three participants use the system. Record all observed usability problems.
3. Redesign based on the problems discovered. Fix all problems that were experienced by two or more participants during the studies. Fix as many of the remaining problems as possible. Any outstanding problems should be recorded and should remain open for all following iterations.
4. If there are any outstanding problems and there is time for another iteration, go back to Step 2. If there are no outstanding problems or there is insufficient time for another iteration, then stop. Maintain a running average of usability measures (such as time-on-

task and user satisfaction scores) based on the last three participants. If the running averages at the end of testing do not meet the established criteria, then inform management that additional design and tests are required.

5. Any outstanding problems at the end of testing should be recorded and

This strategy blends the benefits of large and small sample studies. During each iteration, only three participants are observed before the system is redesigned. Therefore, the most frequent problems are quickly identified and corrected. With five iterations, for example, the total sample size would be 15 participants. With several iterations many less frequent problems will also be identified and corrected because uncorrected problems are recorded and tracked through all iterations. (From the tables in Lewis, 1990, a problem discovery study with 15 participants should uncover almost all of the problems that would affect 25 percent or more of the user population. The same study would be expected to uncover about 80 percent of the problems that would affect as few as 10 percent of the user population.)

Criterion-Based Usability Studies: Criterion Clearly Met

Usability criteria can be established in a number of ways. For example, they may represent a development group's best guess at acceptable user performance, they may be derived from recommendations by customers, or they may be based on studies of competitive systems (Lewis, 1982). Regardless of the source of the criteria, there are some cases in which a criterion is clearly met, based on sample sizes as small as two.

In a recent user-oriented systems test, participants were asked to install an option in a PS/2 Model 90. The criterion time was 60 minutes. The two participants completed the task in 35 and 37 minutes, respectively. The obtained mean time of 36 minutes was statistically significantly faster than the criterion ($t(1)=24$, $p=.013$). In the same study, two participants installed an option card in six and ten minutes respectively. The observed mean of 8 minutes was significantly faster than the criterion of 30 minutes ($t(1)=11.0$, $p=.03$).

When the difference between the observed mean and the criterion is large, the difference can be reliably detected with as few as two participants. The minimum number of participants is two because the *t* statistic cannot be calculated without an estimate of variability. The formula for the sample variance (Walpole, 1976) from a random sample $x_1, x_2, \ldots, x_n$ is given in Equation 1. If the sample size is 1, then the denominator of Equation 1 will be 0, and the variance is undefined. If the sample size is 2, then the denominator will be equal to 1.

$$\underline{s}^2 = \{\Sigma[\underline{x}-(\underline{x})]^2\}/(\underline{n}-1) \qquad (1)$$

The formula for a *t*-test to compare a sample mean to a criterion (Walpole, 1976) is given in Equation 2.

$$t = (x-u_o)/(s/n^{1/2}) \qquad (2)$$

The numerator of Equation 2 is the difference between the sample mean (x) and the criterion (u_o). This difference is sometimes abbreviated *d*. The denominator is the sample standard deviation (s, the square root of the sample variance, s^2) divided by the square root of the sample size (n).

The critical value of *t* is the value that is considered statistically significant. When the observed *t*-value is obtained from a sample, it is compared to the critical *t*-value. If the magnitude of the observed *t* is greater than the critical *t*, then the *t*-test indicates a statistically significant difference.

The critical value of *t* changes depending upon the degrees of freedom (v) and the α level used to determine the acceptable Type I error rate. The Type I error rate is the likelihood of mistakenly rejecting the hypothesis that the observed mean and the criterion are essentially equal (usually called the null hypothesis). The degrees of freedom are directly related to the sample size (n). For the type of test illustrated in Equation 2, the degrees of freedom are equal to $n-1$ (Walpole, 1976). For a test with few degrees of freedom, the critical value of *t* will be fairly large. As the degrees of freedom approach infinity, the *t*-distribution gets smaller and approaches the *z*-distribution. For example, the critical value of the *t*-distribution with one

degree of freedom and α equal to .05 (a common criterion for statistical significance) is 6.314 (Walpole, 1976). For an infinite number of degrees of freedom, the critical *t* is 1.645.

The obtained value of *t* is affected by three characteristics of the sample. The larger the numerator of Equation 2 (*d*, the difference between the obtained mean and the criterion), the larger the obtained *t* will be. The smaller the standard deviation (*s*), then the smaller the denominator of Equation 2, and the larger will be the obtained *t*. Finally, the larger the sample size (*n*), the smaller the denominator of Equation 2, and the larger will be the obtained value of *t*.

In both of the examples cited at the beginning of this section, the observed times were close in value (35 and 37 minutes for the first example, 6 and 10 minutes for the second example), resulting in small variances and standard deviations (*s*). Both means were considerably smaller than their respective criteria (a mean of 36 minutes against a criterion of 60 minutes for the first example, and a mean of 8 minutes against a criterion of 30 minutes for the second example, both large *d*s). These two conditions (large *d*, small *s*) result in a significant *t*-test with a sample size of two. The calculation of the observed *t*-values for these examples is shown in Table 2.

TABLE 2. Calculation of *t* for Two Examples

Calculation Step	Example 1	Example 2
Sample size	2	2
Degrees of freedom	1	1
Alpha (α)	0.05	0.05
Critical *t* value	6.314	6.314
Observed values	35, 37 minutes	6, 10 minutes
Observed mean	36 minutes	8 minutes
Criterion	60 minutes	30 minutes
Difference (*d*)	24 minutes	22 minutes
Sample variance	2	8
Standard deviation	1.414	2.828
Obtained *t* value	24	11
Significant?	Yes	Yes

These examples, taken from a recent usability test, show that it is possible to determine from a small sample that user performance is significantly better than a given criterion. The risk involved in planning to use a small sample is that it is impossible to predict that the conditions required to achieve statistical significance with a small sample (large d, small s) will occur in a study. If these conditions do not occur and the t-test is not significant, there are two possible explanations:

1. There really isn't any difference between the observed mean and the criterion. The null hypothesis is true.
2. There is a difference, and if the study is continued and the sample size is increased (making the denominator of the obtained t-value smaller and, thus, the obtained t larger) then the difference will be proven to be statistically significant. The null hypothesis is really false, but more information must be obtained in order to establish sufficient proof.

I recommend the following strategy for industrial usability studies in which an observed mean will be compared to a criterion.

If previous studies are available from which the sample mean and standard deviation can be estimated, and the criterion is known or can be estimated, then use Equation 3 (Diamond, 1981) to estimate the sample size required.

$$n = (t_a+t_b)^2/(s^2/d^2) \qquad (3)$$

1. Use the standard deviation from the previous study to estimate s.
2. Determine the value of *d*, the difference between the obtained mean and the criterion, which it is important to be able to detect. Increasing the value of *d* will lead to a smaller *n*.
3. t_a is the *t*-value associated with the Type I error, the likelihood that the .10 as a default setting for α. Use the degrees of freedom from the previous study to select the specific t_a from the *t*-table (found at the back of most introductory statistics textbooks).
4. t_b is the *t*-value associated with the Type II error, the likelihood that the null hypothesis will not be rejected when it is actually false. For industrial usability studies, I recommend .20 as a default setting for β. Again, use the degrees of freedom from the previous study to select the specific t_b from the *t*-table.
5. Although I have recommended default settings for α and β, these defaults should be carefully examined for applicability to any specific situation. To properly set the *t*-values, the expected costs of Type I and Type II errors should be estimated and compared.

If no previous studies are available, plan to run a relatively small sample. Six participants would be reasonable, but the time required to observe a participant should be considered. If the effort is small, plan to observe a few more participants, but if the effort is great, plan to observe fewer participants. If significant results are obtained with this initial sample, stop the study. If significant results are not obtained, use the initial sample to derive estimates of *s* and *d*, then use Equation 3 to estimate the sample size that would be required to achieve statistical significance. If the additional sample required is small, continue the study. If the additional sample required is large, consult with management to determine if the study should be continued or abandoned.

A Numerical Example

Suppose a previous study of six participants installing an option showed that the option required an average of 20 minutes to install. The standard deviation was five minutes. To drive for constant improvement, the new criterion value is set for 15 minutes. An analysis of the new installation procedures has indicated that participants may install the option in as little as ten minutes. Because the sample size from the previous study was six, the degrees of freedom were five. Therefore, t_a=.10=1.48 and t_b=.20=0.92. *d*, the difference between the expected mean and the criterion is five minutes, and *s*, the estimate of the standard deviation is also five minutes. Putting these values into Equation 3, the estimate for *n* is (1.48+0.92)2/(52/52)=5.76

participants. For sample size formulas, all fractional values of *n* should be rounded up to the next whole number (Walpole, 1976). In other words, given the assumptions derived from the previous study, a study of six participants should be adequate for the current study to determine statistical significance as long as the true mean is five minutes (or more than five minutes) better than the criterion.

In conclusion, if the difference between the true mean and a criterion is large, then that difference can be detected by a study with relatively few participants. If the difference is small, then a larger number of participants would be required to prove that the difference is statistically significant. However, if the difference is small, it may not be of any practical significance. The strategy described above should help in the design of efficient and reasonable usability tests in which observed means are to be compared to criteria.

Binomial Confidence Intervals for Usability Defect Rates

A problem observed during a usability study may be an indication of a defect in the design of the system (Norman, 1983). In usability studies, a usability defect rate for a specific problem is the number of participants who experience the problem divided by the total number of participants. Usability defect rates can be measured as proportions or percentages. A percentage is a proportion multiplied by 100. (I usually use percentages in reports to decision makers because I believe most non-scientific business professionals understand percentages more easily than proportions.)

The statistical term for a study to estimate a defect rate is a binomial experiment, because a given problem either will or will not occur during the study. For example, a participant either will or will not install an option correctly. The point estimate of the defect rate is the observed percentage of failures. However, the likelihood is very small that the point estimate from a study is exactly the same as the true percentage of failures, especially if the sample size is small (Walpole, 1976). To compensate for this, interval estimates that have a known likelihood of containing the true percentage can be calculated. These binomial confidence intervals can be used to describe the percentage of usability defects effectively, often with a small sample. (See Appendix A for a BASIC program for binomial confidence intervals.) The report of a binomial confidence interval usually takes the form of:

> The observed percentage of (a particular usability defect) was PP percent. The lower limit of the CC-percent binomial confidence interval was XX percent and the upper limit was YY percent,
>
> where:
>
> The percentage PP is the observed percentage of failures.
>
> The value of CC is the likelihood (confidence) that the interval will contain the true defect rate.
>
> XX is the lower limit of the interval.
>
> YY is the upper limit of the interval.

"Ideally, we prefer a short interval with a high degree of confidence." (Walpole, 1976, p. 123) The factors that affect the width of a confidence interval are very similar to those that affect the sample size estimates discussed in the previous section. All other factors being equal, a confidence interval computed from a large sample will be shorter than one computed from a small sample. A confidence interval with a high degree of confidence will be wider than one with a lower degree of confidence. The effects of these factors are illustrated in Table 3.

TABLE 2. Effect of Varying Sample Size and Degree of Confidence on Interval Width

Sample Size	Confidence Level	Lower Limit	Observed Value	Upper Limit	Interval Width
8	90	40	75	96	56
80	90	66	75	83	17
8	95	35	75	97	62
80	95	64	75	84	20

For a 95-percent binomial confidence interval, it is 95 percent likely that the interval contains the true percentage of defects. If the true percentage is high, then the lower limit of a 95-percent binomial confidence interval will be high, even with a small sample. If the lower limit of the confidence interval is unacceptable, then it is legitimate to conclude that the defect rate is unacceptable, regardless of the sample size. The following examples illustrate the use of this technique.

Errata Sheet Effectiveness: A Risk Assessment

King, Lee and Lewis (1990) studied the effectiveness of an errata sheet placed on top of the documentation ship group of a product. The primary variable of interest was whether a user unpacking the system would use the errata sheet. The errata sheet was designed to attract the user's attention, and large (24-point) print at the top of the page stated "DO THIS FIRST!" Six out of eight participants ignored the errata sheet. Ignoring the errata sheet was defined as a usability defect, so the observed defect rate was 75 percent. A 95-percent binomial confidence interval for this defect rate ranged from 35-percent to 97-percent failures. Even with this small sample, it was possible to predict that a 35-percent defect rate was the lowest rate of failure that could be obtained with this type of errata sheet. We concluded that the use of an errata sheet will generally be an unacceptable strategy, and cannot be recommended.

Evaluation of Graphic Symbols for PHONE and LINE

Lewis and Pallo (1991) studied the effectiveness of graphic symbols for phone and phone line connection for attaching telephony equipment to a computer. Computer-naive participants were asked to attach a telephone to a computer with only the proposed graphic symbols to guide the installation. Nine of eleven installations (82 percent) were incorrect. The 95-percent confidence interval for this percentage ranged from 48 percent to 98 percent, and was considered to be unacceptably high. We were 95-percent confident that unless additional information was provided to users, the failure rate for installation would be at least 48 percent.

If a study is planned in which the percentage of usability defects will be measured, I recommend a strategy similar to that described for studies in which a mean is compared to a criterion. Study a small sample of participants and record the number of usability defects (such as incorrect installations or failures to complete assigned tasks). Use the procedure described in Walpole (1976) or the program listed in Appendix A to calculate the binomial confidence interval.

Report the observed defect percentage and the lower limit of the binomial confidence interval to the product developer. Ask if the lower limit is an acceptable defect rate because the 95-percent confidence interval indicates that it is 95-percent likely that the true defect rate will be at least as high as the lower limit of the interval. When the defect rate is high, this can be a very convincing argument for redesigning the product or system that has been studied. When the defect rate is low, this procedure may not work without a larger sample size. The decision to continue collecting samples or to stop the study should be determined by a reasonable business case that balances the cost of continued data collection against the potential cost of allowing the defect to go uncorrected.

Discussion

It would be misleading to suggest that all types of usability studies can be effectively conducted with small samples. However, these three types of studies (problem discovery studies, studies to compare a mean to a criterion, and binomial experiments to estimate usability defect rates) comprise a large proportion of the usability studies conducted in industrial settings. As a general rule, efficiency is better accomplished by using the smallest sample possible while still gathering enough data to allow decision makers to make informed judgments regarding the usability of products and systems.

References

Bradley, J. V. (1976). *Probability; decision; statistics*. Englewood Cliffs, NJ: Prentice-Hall.

Gould, J. D. (1988). How to design usable systems. In M. Helander (Ed.), *Handbook of Human-Computer Interaction* (pp. 757-789). New York, NY: North-Holland.

King, L., Lee, R., and Lewis, J. R. (1990). *Errata sheet effectiveness: A risk assessment* (IBM Tech. Report 54.567). Boca Raton, FL: International Business Machines, Inc.

Lewis, J. R. (1982). Testing small system customer setup. In *Proceedings of the Human Factors Society 26th Annual Meeting* (pp. 718-720). Seattle, WA: Human Factors Society.

Lewis, J. R. (1990). *Sample sizes for observational usability studies: Tables based on the binomial probability formula* (Tech. Report 54.571). Boca Raton, FL: International Business Machines, Inc.

Lewis, J. R. and Pallo, S. (1991). *Evaluation of graphic symbols for phone and line* (IBM Tech. Report 54.572). Boca Raton, FL: International Business Machines, Inc.

Norman, D. (1983). Design rules based on analyses of human error. *Communications of the ACM*, 4, 254-258.

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

Appendix A. A BASIC Program for Binomial Confidence Intervals

```
5 ' Approximate binomial confidence limits:  Level 2.0, 8/29/90
10 '
11 ' Z and P9 (line 15) are set for calculating 2-sided limits
15 z(1)=1.645:Z(2)=1.96:Z(3)=2.575
16 cls
17 Print "Approximate 2-sided binomial confidence intervals (90%, 95%, 99%)"
19 print: Print "Enter the observed number of occurrences (x) and the number of"
20 print "opportunities for occurrence (n, the sample size) separated "
21 print "by a comma.  The results displayed are proportions.  "
22 print "Move the decimal place over two positions (i.e., multiply by 100)
23 print "to convert to percentages.  ":print:print
30 INPUT "Enter x,n: ",X1,N:print:print
40 for cnt=1 to 3
50 let z=z(cnt)
70 X=X1:IF X=N THEN p2=1:if p2 = 1 then goto 95
75 GOSUB 5815
80 P2=P8
95 ' Get lower limit P1 by replacing x1 by n-x1 and
96 ' carry out calculation as before ; then P1 = 1 - P8.
100 IF X1=0 THEN p1=0:if p1=0 then goto 130
103 x=n-x1
105 GOSUB 5815
110 P1=1-P8
130 print
132 if cnt=1 then print "90% confidence interval: ";
133 if cnt=2 then print "95% confidence interval: ";
134 if cnt=3 then print "99% confidence interval: ";
135 print using "#.###";p1;:print " - ";:print using "#.###";x1/n;
136 print " - ";:print using "#.###";p2
137 next cnt
138 print:print:print "(Use Print Screen if hardcopy is required.)"
139 PRINT:print:INPUT "Do you want to do more calculations? (Y/N)",A$
140 IF left$(A$,1)="y" THEN LET A$="Y"
145 IF left$(A$,1)="Y" THEN 16
150 system
155 '
```

```
5800 '***** Approximate binomial confidence limits *****
5801 ' Use the Paulson-Takeuchi approximation described in:
5802 ' Yoritake Fujino: Approximate binomial confidence limits,
5803 ' Biometrika (1980) 67, 677-681.
5804 ' Line 5830 calculates the upper binomial probability limit,
5805 ' where F2 is an approximation to the requisite upper
5806 ' percentage point of an F-statistic with degrees of freedom of
5807 ' 2*(x+1) and 2*(n-x)
5808 ' The approximation to the inverse of the cumulative F is
5809 ' derived from Paulson's approximation to the cumulative F.
5810 ' If desired the user may input the exact F percentage point
5811 ' prior to line 5830.
5812 '
5815 A=1/(9*(X+1)):A1=1-A
5820 B=1/(9*(N-X)):B1=1-B
5825 F2=((A1*B1+Z*SQR(A1^2*B+A*B1^2-A*B*Z^2))/(B1^2-B*Z^2))^3
5830 P8=(X+1)*F2/((N-X)+(X+1)*F2)
5835 RETURN
5840 ***** END BINOMIAL CONFIDENCE LIMITS *****
```