Sample Size Estimation and Use of Substitute Audiences

TR 29.3385

January 3, 2001

James R. Lewis
Speech Product Design and Usability
West Palm Beach, FL

Abstract

This paper discusses two important usability evaluation topics: sample size estimation and the use of substitute audiences. The first section of the paper contains explanations and examples for estimating required sample sizes for parameter estimation studies, comparative studies, and problem-discovery studies (both user-based and heuristic). The second section discusses the use of substitute audiences and the generalizability of results obtained with substitute audiences.

ITIRC Keywords

Usability evaluation
Sample size estimation
Substitute audiences
Generalizability

Contents

Executive Summary

This paper discusses two important usability evaluation topics: sample size estimation and the use of substitute audiences. The first section of the paper contains explanations and examples for estimating required sample sizes for parameter estimation studies, comparative studies, and problem-discovery studies (both user-based and heuristic). The second section discusses the use of substitute audiences and the generalizability of results obtained with substitute audiences. For each of these topics, this summary provides an overview, indications and contraindications of use, a brief description, and a discussion of required resources and available tools.

Sample Size Estimation

Overview

When preparing to run a usability evaluation study, it is important to have some idea about the required sample size. If you run too few participants, then you might fail to achieve the goals of your study. If you run too many participants, then you will expend more resources than necessary, making the study inefficient.

Indications

Plan to perform sample size estimation for any study in which you will take measurements or count events (which is a form of measurement). Some examples of applicable types of studies are:

- Standard scenario-based problem-discovery usability studies (estimating the number of participants)
- Heuristic evaluations (estimating the number of evaluators)
- Competitive assessment of speech dictation throughput (estimating the number of participants)
- Characterization of typing speed for a new keyboard (estimating the number of typists)
- Competitive assessment of the frequency of passive verb usage in software documentation (estimating the number of text samples to analyze)

Contraindications

If you have a set deadline or limited participant availability that constrains your sample size to a fairly small number (say, three or fewer), then there is no point in computing sample size estimates. Even with such a small sample, though, it is usually advantageous and informative to compute standard statistics related to sample size estimation (mean, standard deviation, standard error of the mean, confidence intervals, *t*-tests as appropriate), if for no other reason than to quantify the limits of the accuracy of your study's measurements due to the small sample size.

Description

For parameter estimation and comparative studies, it is important to have an estimate of the variance of key measurements. With this estimate and decisions about the required precision of measurement and confidence level, it is reasonably easy to estimate required sample sizes. For problem-discovery studies, it is important to have an estimate of p, the mean probability of problem discovery. This estimate can come from previous studies of similar systems or from the literature describing problem discovery as a function of sample size for both user-based and heuristic evaluations. The body of this paper contains the details and numerous examples of how to use estimates of variance and p to estimate required sample sizes, as well as a discussion of other factors that can affect sample size decisions.

Resources Required

Any contemporary spreadsheet program provides the functions needed to perform the computations described in this paper. In addition, you will need a table of the t-statistic, which is available in almost any textbook on introductory or advanced statistics. With these resources and an estimate of variance (or p), it should only take a few minutes to perform the computations.

Available Tools

Because the resources required to perform the computations are relatively light, there are no specialized tools available for this purpose (although some statistics packages might include sample size estimation features for parameter estimation and comparative studies).

Use of Substitute Audiences

Overview

A key issue in the generalizability of results from a sample to a population is the extent to which the sample is representative of the population of interest. Theoretically, the only way to ensure representativeness (in the statistical sense) is to create the sample by drawing sample units (for example, test participants) randomly from the population of interest. Practically, this is almost never possible. The purpose of this section of the paper is to discuss issues related to the use of substitute audiences and the associated risks.

Indications

When you don’t have access to a random sample of the population of interest, you will need to either use a substitute audience or abandon the proposed study. In most cases, you should be able to find participants that are reasonably similar to most of our target audiences, either within IBM or through a temporary employment agency.

Contraindications

If the target audience is extremely specialized (chairmen of Fortune 500 companies, Japanese Java programmers with at least 15 years of programming experience, etc.), then you might not be able to reasonably argue for the use of a substitute audience and, if you can’t get access to the target audience, might have to abandon the proposed study.

Description

The body of this paper addresses topics such as the types of human characteristics (experience, individual differences) that affect human performance with computer systems. For example, it is very important to make sure that members of a substitute audience are as similar as possible to the target audience with regard to critical types of expertise. There is some indication that individual differences in field dependence/independence can affect user performance for certain types of computer-related tasks. The paper also provides tips on enhancing generalizability and some guidance on when to study multiple groups.

Resources Required

You need to have access to participants, possibly through IBM internal volunteers, external volunteers, or through a temporary employment agency (although there are other possible sources – see the paper on UCD Logistics for more information). It usually should not take much time to determine if a substitute audience has reasonable similarity to a target audience.

Available Tools

It would be useful for us to begin systematic use of the Windows[1] 95 version of the WCEQ (Windows Computer Experience Questionnaire – first described in Miller, Stanney, and Wooten, 1997, and available from jimlewis@us.ibm.com) and collection of this information into a UCD database. This would make it easier to quantify the similarity of different audiences on the important user characteristic of computer experience with Windows.

[1] Windows is a trademark or registered trademark of Microsoft Corp.

Introduction

Sample Size Estimation

One key purpose of this paper is to discuss the principles of sample size estimation for three types of evaluation: comparative (also referred to as experimental), population parameter estimation, and problem-discovery (also referred to as diagnostic, observational, or formative)[2]. The paper also covers certain special issues of interest to the IBM UCD community, such as sample sizes for international tests, for various types of design evaluation activities, and for less traditional areas of UCD evaluation (such as the assessment of advertising materials) in an effort to accomplish the goal of evaluating the total user experience (TUE).

Substitute Audiences

The second key purpose of this paper is to discuss the use of substitute audiences. All researchers would agree that the ideal sample is one randomly drawn from the population of interest. For us as IBMers, though, getting to that population is sometimes logistically impossible. When is it advisable to use substitute audiences? What are the risks in extrapolating to the target audiences? In sharp contrast to the mathematical sophistication applied to the issue of sample size estimation (although sample size estimation also requires careful human judgement), it will be necessary to address this topic primarily with rational argument because of the paucity of empirical quantitative work in the area.

[2] This paper assumes knowledge of introductory applied statistics. If you're not comfortable with terms such as mean, variance, standard deviation, *p*, *t*-score, and *z*-score, then refer to an introductory statistics text such as Walpole (1976) for definitions of these and other fundamental terms.

Sample Size Estimation

Sample size estimation requires a blend of mathematics and judgement. The mathematics are straightforward, and it is possible to make reasoned judgements (for example, judgements about expected costs and precision requirements) for those values that the mathematics cannot determine.

Sample size estimation for comparative and population parameter estimation studies
Traditional sample size estimation for comparative and estimation studies depends on having an estimate of the variance of the dependent measure(s) of interest and an idea of how precise (the magnitude of the critical difference and the statistical confidence level) the measurement must be (Walpole, 1976). Once you have that, the rest is mathematical mechanics (typically using the formula for the *t* statistic[3], but sometimes using binomial probabilities).

You can (1) get an estimate of variance from previous studies using the same method (same or similar tasks and measures), (2) you can run a quick pilot study to get the estimate (for example, piloting with four participants should suffice to provide an initial estimate of variability), or (3) you can set the critical difference your are trying to detect to some fraction of the standard deviation (Diamond, 1981). (See the following examples for more details about these different methods).

Certainly, people prefer precise measurement to imprecise measurement, but all other things being equal, the more precise a measurement is, the more it will cost, and running more participants than necessary is wasteful of resources (Kraemer and Thiemann, 1987). The process of carrying out sample size estimation can also lead UCD workers and their management to a realistic determination of how much precision they really need to make their required decisions.

Alreck and Settle (1985) recommend using a 'what if' approach to help decision makers determine their required precision. Start by asking the decision maker what would happen if the average value from the study was off the true value by one percent. Usually, the response would be that a difference that small wouldn't matter. Then ask what would happen if the measurement were off by five percent. Continue until you determine the magnitude of the critical difference. Then start the process again, this time pinning down the required level of statistical confidence. Note that statistically unsophisticated decision makers are likely to start out by demanding 100% confidence (which is only possible by sampling every unit in the population). Presenting them with the sample sizes required to achieve different levels of confidence can help them settle in on a more realistic confidence level.

[3] Because, in my experience (supported by published studies of robustness, including Lewis, 1993), *t*-tests are more likely to work effectively than to provide misleading results when applied to the types of data with which usability researchers typically work, I do not bring up the statistical assumptions that underlie the *t* statistic. These assumptions are available in any introductory statistics text.

Example 1. Population Parameter Estimation – Estimate of Variability Available and Realistic Measurement Criteria

For speech recognition applications, the recognition accuracy is an important value to track due to the adverse effects misrecognitions have on product usability. Part of the process of evaluating a speech recognition product, then, is estimating its accuracy.

For this example (based on work reported in Lewis, 1997), suppose:

- Recognition variability from a previous similar evaluation = 6.35
- Critical difference (d) = 2.5%
- Desired level of confidence: 90%

The appropriate procedure for estimating a population parameter is to construct a confidence interval (Bradley, 1976). You determine the upper and lower limits of a confidence interval by adding to and subtracting from the observed mean the value:

[1] sem * crit-t

where sem is the standard error of the mean (the standard deviation, S, divided by the square root of the sample size, n) and crit-t is the t-value associated with the desired level of confidence (found in a t-table, available in most statistics texts). Setting the critical difference to 2.5 is the same as saying that the value of sem * crit-t should be equal to 2.5. In other words, you don't want the upper or lower bound of the confidence interval to be more than 2.5 percentage points away from the observed mean percent correct recognition, for a confidence interval width equal to 5.0.

Calculating the sem depends on knowing the sample size, and the value of crit-t also depends on the sample size, but you don't know the sample size yet. Iterate using the following method.

1. Start with the z-score for the desired level of confidence in place of t-crit[4]. For 90% confidence, this is 1.645.

2. Algebraic manipulations based on the formula sem * $z = d$ results in $n = (z^2 * S^2)/d^2$ which, for this example, is $n = (1.645^2 * 6.35)/2.5^2$, which equals 2.7. Always round sample size estimates up to the next whole number, so this initial estimate is 3.

3. Now you need to adjust the estimate by replacing the z-score with the t-score for a sample size of 3. For this estimate, the degrees of freedom (df) to use when looking up the value in

[4] By the way, if you actually **know** the true variability for the measurement rather than just having an estimate, you're done at this point because it's appropriate to use the z-score rather than a t-score. Unfortunately, we almost never know the true variability, but must work with estimates.

a *t* table is *n*-1, or 2. This is important because the value of *z* will always be smaller than the appropriate value of *t*, making the initial estimate smaller than it should be. For this example, crit-*t* is 2.92.

4. Recalculating for *n* using 2.92 in place of 1.645 produces 8.66, which rounds up to 9.

5. Because the appropriate value of *t*-crit is now a little smaller than 2.92 (because the estimated sample size is now larger, with 9-1 or 8 degrees of freedom), recalculate *n* again, using *t*-crit equal to 1.860. The new value for *n* is 3.515, which rounds up to 4.

6. Stop iterating when you get the same value for *n* on two iterations or you begin cycling between two values for *n*, in which case you should choose the larger value[5]. Table 1 shows the full set of iterations for this example, which ends by estimating the appropriate sample size as 5.

Table 1. Full set of iterations for Example 1

				Iteration		
	Initial	**1**	**2**	**3**	**4**	**5**
crit-t	1.645	2.92	1.86	2.353	2.015	2.132
crit-t^2	2.71	8.53	3.46	5.54	4.06	4.55
S^2	6.35	6.35	6.35	6.35	6.35	6.35
d	2.5	2.5	2.5	2.5	2.5	2.5
Estimated n	2.7493	8.663	3.515	5.6252	4.1252	4.618
Rounded up	3	9	4	6	5	5
df	2	8	3	5	4	4

The measurement criteria in Example 1 are reasonable – 90% confidence that the observed interval (limited to a total length of 5%) contains the true mean[6].

[5] Diamond (1981) points out that you can usually get by with an initial estimate and one iteration because most researchers don't mind having a sample size that's a little larger than necessary. If the cost of each sample is high, though, it probably makes sense to iterate until reaching one of the stopping criteria. Note that the initial estimate establishes the lower bound for the sample size (3 in this example), and the first iteration establishes the upper bound (9 in this example).

[6] Note that there is nothing in this equation that makes reference to the size of the population! Unless the size of the sample is a significant percentage of the total population under study (which is rare), the size of the population is irrelevant. Alreck and Settle (1985) explain this with a soup-tasting analogy. Suppose you're cooking soup in a one-quart saucepan, and want to test if it's hot enough. You would stir it thoroughly, then taste one teaspoon. If it were a two-quart saucepan, you'd follow the same procedure – stir thoroughly, then taste one teaspoon.

Example 2. Population Parameter Estimation – Estimate of Variability Available and Less Realistic Measurement Criteria
Example 2 shows what would have happened if the measurement criteria were less realistic, illustrating the potential cost associated with high confidence and high measurement precision.

Suppose the measurement criteria for the situation in Example 1 were less realistic, with:

- Recognition variability from a previous similar evaluation = 6.35
- Critical difference (*d*) = .5%
- Desired level of confidence: 99%

In that case, the initial *z*-score would be 2.576, and the initial estimate of *n* would be:

[2] $n = (2.576^2*4.9)/.5^2 = 168.549$, which rounds up to 169.

Recalculating *n* with crit-*t* equal to 2.605 (*t* with 168 degrees of freedom) results in *n* equal to 172.37, which rounds up to 173. (Rather then continuing to iterate, note that the final value for the sample size must lie between 169 and 173.)

There might be some industrial environments in which usability investigators would consider 169 to 173 participants a reasonable and practical sample size, but they are probably rare. On the other hand, collecting data from this number of participants in a mailed survey is common.

Example 3. Population Parameter Estimation – No Estimate of Variability Available
For both Examples 1 and 2, it doesn't matter if the estimate of variability came from a previous study or a quick pilot study. Suppose, however, that you don't have any idea what the measurement variability is, and it's too expensive to run a pilot study to get the initial estimate.

Example 3 illustrates a technique (from Diamond, 1981) for getting around this problem. To do this, though, you need to give up a definition of the critical difference (*d*) in terms of the variable of interest and replace it with a definition in terms of a fraction of the standard deviation.

In this example, the measurement variance is unknown. To get started, the researchers have decided that, with 90% confidence, they do not want *d* to exceed half the value of the standard deviation[7]. The measurement criteria are:

- Recognition variability from a previous similar evaluation = N/A
- Critical difference (*d*) = .5*S*
- Desired level of confidence: 90%

The initial sample size estimate is:

[7] Using the data from Example 1, the value of *d* in this example would be .5**S* where *S* = 2.5, or 1.25.

$$[3]\ n = (1.645^2 * S^2)/(.5S)^2 = 10.824,\text{ which rounds up to 11.}$$

The result of the first iteration, replacing 1.645 with crit-*t* for 10 degrees of freedom (1.812), results in a sample size estimation of 13.13, which rounds up to 14. The appropriate sample size is therefore somewhere between 11 and 14, with the final estimate determined by completing the full set of iterations.

Example 4. Population Parameter Estimation for Binomial Variables

A problem[8] observed during a usability study may be an indication of a defect in the design of the system (Lewis and Norman, 1986; 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.

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 for each trial (participant) in the experiment. For example, a participant either will or will not install an option correctly. The point estimate of the defect rate is the observed proportion of failures (p). 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, you can calculate interval estimates that have a known likelihood of containing the true proportion. You can use these binomial confidence intervals to describe the proportion of usability defects effectively, often with a small sample (Lewis, 1991, 1994b, 1996). (See Appendix A for a BASIC program that calculates approximate binomial confidence intervals based on the Paulson-Takeuchi formula (Fujino, 1980; Lewis, 1996.))

The variance for a binomial experiment is $p(1-p)$. Thus, the maximum variance possible occurs when p=.5, making Var(p)=.25. As p becomes larger or smaller, the variance becomes smaller, reaching 0 when p equals either 1.0 or 0.0. Substituting $p(1-p)$ for the variance in the previous formula for n results in the following initial sample size estimate for a binomial experiment:

$$[4]\ n = (z^2 * p * (1-p))/d^2$$

If you have some idea about what p is (on the basis of previous studies or pilot studies), then you should use it. If you have no idea, then assume p is equal to .5.

[8] One of the current issues in usability evaluation is determining exactly what constitutes a problem. Tackling this topic in detail is outside the scope of this paper. For scenario-based problem-discovery studies, I have usually tried to define a usability problem as any event that degrades a user's ability to complete a task as efficiently as possible, with different impact levels assigned to an observed problem on the basis of whether the problem prevented the user from completing the task, took more than one minute to resolve, or took less than one minute to resolve. See Lewis (1994a) for more details.

For this example, assume you are studying the effectiveness of a wordless graphic instruction that has the purpose of illustrating to users how to connect a telephone to a computer. (This example is from Lewis and Pallo, 1991.) You have no idea what to expect for a failure rate, so you establish the following measurement criteria:

- Binomial variance: .25 (maximum possible)
- Critical difference (*d*): .15
- Desired level of confidence: 90%

Table 2 shows the cycle of iterations to arrive at the estimated sample size for this experiment:

Table 2. Sample size iteration for Example 4

	Initial	**1**
Crit-t	1.645	1.697
Crit-t^2	2.706	2.880
p	0.5	0.5
1-p	0.5	0.5
p(1-p)*	0.25	0.25
d	0.15	0.15
d^2	0.0225	0.0225
Estimated n	30.067	31.998
Rounded up	31	32
df	30	31

If $p = .5$ and $n = 32$, then the 90% confidence interval should range from about .35 to .65 (.50 +/- .15). Table 3 shows how the size of the confidence interval gets smaller as the observed *p* moves away from .5 (using the program given in Appendix A to compute the confidence intervals[9]).

[9] The program in Appendix A produces approximate binomial confidence intervals that always contain the true binomial confidence interval, but are sometimes slightly conservative. Note that this is true for the computed interval at *p*=.5 which has a size of 0.32, slightly larger than the projected size of 0.30. Normally, this small difference will not cause any serious measurement problems. In cases where it is necessary to be extremely precise, compute the exact binomial confidence interval using procedures given in Steele and Torrie (1960). You can't just add and subtract *d* from the observed *p* because, except at *p*=.5, the appropriate binomial confidence interval is not symmetrical around *p*.

Table 3. 90% binomial confidence intervals for various values of p with n=32

Defects	**16**	**13**	**9**	**8**	**6**	**3**	**1**
p	0.50	0.41	0.28	0.25	0.19	0.09	0.03
Upper limit	0.66	0.57	0.44	0.41	0.34	0.23	0.14
Lower limit	0.34	0.26	0.16	0.13	0.09	0.03	0.001
Interval size	0.32	0.31	0.28	0.28	0.25	0.20	0.14

Note that if you have reason to suspect a spectacular failure rate, you do not need a very large sample to acquire convincing evidence. In the first evaluation of the wordless graphic instruction mentioned above (Lewis and Pallo, 1991), 9 of 11 installations (82%) were incorrect. The 90% binomial confidence interval for this outcome ranged from .53 to .97. This interval allowed us to be 90% confident that unless Development provided additional information to users, the failure rate for installation would be at least 53%.

This suggests that a reasonable strategy for binomial experiments is to start with a small sample size and record the number of failures. From these results, compute a confidence interval. If the lower limit of the confidence interval indicates an unacceptably high failure rate, stop testing. Otherwise, continue testing and evaluating in increments until you reach the maximum sample size allowed for the study.

Example 5. Using Population Parameter Estimation to Compare an Observed Measure to a Criterion

One reason in usability engineering to compute a population parameter is to compare it to a usability criterion. One can establish usability criteria in a number of ways (Lewis, 1982). For example, usability criteria could:

- represent a development group’s best guess (consensus) at acceptable user performance
- be the result of customer feedback
- come from usability studies of previous or competitive systems
- appear in industry standards.

For example, suppose that you have a product requirement that installation should take no more than 30 minutes. In a preliminary evaluation, participants needed an average of 45 minutes to complete installation. Development has fixed a number of usability problems found in that preliminary study, so you’re ready to measure installation time again, using the following measurement criteria:

- Performance variability from the previous evaluation = 10.0
- Critical difference (*d*) = 3 minutes
- Desired level of confidence: 90%

The interpretation of these measurement criteria is that we want to be 90% confident that we can detect a difference as small as 3 minutes between the mean of the data we gather in the study and the criterion we're trying to beat. In other words, the installation will pass if the observed mean time is 27 minutes or less, because the sample size should guarantee an upper limit to the confidence interval that is no more than 3 minutes above the mean[10]. The procedure for determining the sample size in this situation is the same as that of Experiment 1, shown in Table 4. The outcome of these iterations is an estimated sample size of 6.

Table 4. Full set of iterations for Example 5

	Initial	**1**	**2**	**3**	**4**
t	1.645	2.353	1.943	2.132	2.015
t^2	2.706	5.537	3.775	4.545	4.060
s^2	10	10	10	10	10
d	3	3	3	3	3
d^2	9	9	9	9	9
Estimated n	3.006694	6.151788	4.194721	5.050471	4.511361
Rounded up	4	7	5	6	5
df	3	6	4	5	4

Example 6. Sample Size for a Paired t-test

When you obtain two measurements from participants, you are in a position to compare the results using a paired *t*-test[11]. Another name for the paired *t*-test is the difference score *t*-test, because the measurement of concern is the difference between the participants' two scores.

For example, suppose you plan to obtain recognition accuracy scores from participants who have dictated test texts into your product under development and a competitor's product (following all the appropriate experimental design procedures – Lewis, 1997), using the following criteria:

- Difference score variability from a previous evaluation = 5.0
- Critical difference (*d*) = 2%
- Desired level of confidence: 90%

This situation is similar to that of the previous example, because the typical goal of a difference scores *t*-test is to determine if the average difference between the scores is statistically

[10] Because the initial estimate of variance is certainly not accurate, it is important to calculate the confidence interval from the collected data. Just because the mean is 27 minutes doesn't mean that the system passed the test unless the variance is equal to or less than the initial estimate of variance.

[11] It's beyond the scope of this paper to get into details about appropriate experimental design for these cases. For example, it's important to counterbalance the order in which the participants experience the experimental conditions, if at all possible. For information on appropriate experimental design, consult texts such as Bradley (1976) or Myers (1979).

significantly different from 0. Thus, the usability criterion in this case is 0, and we want to be 90% confident that if the true difference between the systems' accuracies is 2% or more, then we will be able to detect it because the confidence interval for the difference scores will not contain 0.

Table 5 shows the iterations for this situation, leading to a sample size estimate of 6.

Table 5. Full set of iterations for Example 6

	Initial	**1**	**2**	**3**	**4**
t	1.645	2.353	1.943	2.132	2.015
t^2	2.706	5.537	3.775	4.545	4.060
s^2	5	5	5	5	5
d	2	2	2	2	2
d^2	4	4	4	4	4
Estimated n	3.382531	6.920761	4.719061	5.68178	5.075281
Rounded up	4	7	5	6	6
df	3	6	4	5	5

Example 7. Sample Size for a Two-Groups t-test

Up to this point, the examples have all involved one group of scores, and have been amenable to similar treatment. If you have a situation in which you plan to compare scores from two independent groups, then things get a little more complicated. For one thing, you now have two sample sizes to consider – one for each group. It's beyond the scope of this paper to get into this in too much detail, and in many cases the examples already given cover a lot of comparative usability evaluation situations.

To simplify things in this example, assume that the groups are essentially equal (especially with regard to performance variability). For example, this should be true if the groups contain participants from a single population who have received random assignment to treatment conditions. In this case, it is reasonable to believe that the sample size for both groups will be equal[12], which simplifies things. For this situation, the formula for the initial estimate of the sample size for each group is:

[6] $n = (2*z^2*S^2)/d^2$

[12] It gets more complicated if you have reason to believe that the groups are different, especially with regard to variability of performance. In that case, you would want to have a larger sample size for the group with greater performance variability in an attempt to obtain more equal precision of measurement for each group. Advanced market research texts (such as Brown, 1980) provide sample size formulas for these situations.

Note that this is the similar to the formula presented in Example 1, with the numerator multiplied by 2. After getting the initial estimate, begin iterating using the appropriate value for crit-t in place of z.

For example, suppose we needed to conduct the experiment described in Example 6 with independent groups of participants, keeping the measurement criteria the same:

- Estimate of variability from a previous evaluation = 5.0
- Critical difference (d) = 2%
- Desired level of confidence: 90%

In that case, iterations would converge on a sample size of 9 participants per group, for a total sample size of 18, as shown in Table 6.

Table 6. Full set of iterations for Example 7

	Initial	**1**	**2**	**3**
t	1.645	1.943	1.833	1.86
t^2	2.706	3.775	3.360	3.460
s^2	5	5	5	5
d	2	2	2	2
d^2	4	4	4	4
Estimated n	6.765	9.438	8.400	8.649
Rounded up	7	10	9	9
df	6	9	8	8

This illustrates the well-known measurement efficiency of experiments that produce difference scores (within-subjects designs) relative to experiments involving independent groups (between-subjects designs). For the same measurement precision, the estimated sample size for Example 6 was six participants, 1/3 the sample size requirement estimated for Example 7.

Example 8. Making Power Explicit in the Sample Size Formula

The power of a procedure is not an issue when you estimate the value of a parameter, but it is an issue when you test a hypothesis (as in Example 7). In traditional hypothesis testing, you have a null (H_o) and an alternative (H_a) hypothesis. The typical null hypothesis is that there is no difference between groups, and the typical alternative hypothesis is that the difference is greater than zero[13]. When you test a hypothesis (for example, that the difference in recognition

[13] When the alternative hypothesis is that the difference is nonzero, the test is two-tailed because you can reject the null hypothesis with either a sufficiently positive or a sufficiently negative outcome. If you have reason to believe that you can predict the direction of the outcome, or if an outcome in only one direction is meaningful, you can construct an alternative hypothesis that considers only a sufficiently positive or a sufficiently negative outcome. This is a one-tailed test. For more information, see an introductory statistics text (such as Walpole, 1976).

accuracy between two competitive dictation products is nonzero), there are two ways to make a correct decision and two ways to be wrong, as shown in Table 7.

Table 7. Possible outcomes of a hypothesis test

	Reality	
Decision	*H_o is true*	*H_o is false*
Insufficient evidence to reject H_o	Fail to reject H_o	Type II error
Sufficient evidence to reject H_o	Type I error	Reject H_o

Strictly speaking, you never accept the null hypothesis, because the failure to acquire sufficient evidence to reject the null hypothesis could be due to (1) no significant difference between groups or (2) a sample size too small to detect an existing difference. Rather than accepting the null hypothesis, you fail to reject it[14].

Returning to Table 7, the two ways to be right are (1) to fail to reject the null hypothesis (H_0) when it is true or (2) to reject the null hypothesis when it is false. The two ways to be wrong are (1) to reject the null hypothesis when it is true (Type I error) or (2) to fail to reject the null hypothesis when it is false (Type II error)[15].

Table 8 shows the relationship between these concepts and their corresponding statistical testing terms:

Table 8. Statistical testing terms

Statistical Concept	**Testing Term**
Acceptable probability of a Type I error	Alpha
Acceptable probability of a Type II error	Beta
Confidence	1-alpha
Power	1-beta

The formula presented in Example 7 for an initial sample size estimate presented was:

$$[7]\ n = (2*z^2*S^2)/d^2$$

In Example 7, the *z*-score was set for 90% confidence (which means alpha = .10). To take power into account in this formula, you need to add another *z*-score to the formula – the *z*-score associated with the desired power of the test (as defined in Table 8). Thus, the formula becomes:

$$[8]\ n = (2*(z_a+ z_b)^2*S^2)/d^2$$

[14] This is a subtle but important difference.

[15] Note the similarity of these outcomes with the outcomes of signal detection theory: correct rejection, hit, miss and false alarm.

So, what was the value for power in Example 7? When beta equals .5 (in other words, when the power is 50%), the value of z_b is 0, so z_b disappears from the formula. Thus, in Example 7, the implicit power was 50%. Suppose you want to increase the power of the test to 80%, reducing beta to .2.

- Estimate of variability from a previous evaluation = 5.0
- Critical difference (*d*) = 2%
- Desired level of confidence: 90% (z_a=1.645)
- Desired power: 80% (z_b=1.282)

With this change, the iterations converge on a sample size of 24 participants per group, for a total sample size of 48, as shown in Table 9. To achieve the stated goal for power results in a considerably larger sample size.

Table 9. Full set of iterations for Example 8

	Initial	**1**	**2**	**3**
t(alpha)	1.645	1.721	1.714	1.717
t(beta)	1.282	1.323	1.319	1.321
t(total)	2.927	3.044	3.033	3.038
t(total)^2	8.567	9.266	9.199	9.229
s^2	5	5	5	5
d	2	2	2	2
d^2	4	4	4	4
Estimated n	21.418	23.165	22.998	23.074
Rounded up	22	24	23	24
df	21	23	22	23

Note that the stated power of a test is relative to the critical difference – the smallest effect worth finding. Either increasing the value of the critical difference or reducing the power of a test will result in a smaller required sample size.

A discussion of appropriate statistical criteria for industrial testing

In scientific publishing, the usual criterion for statistical significance is to set the permissible Type I error (alpha) equal to 0.05. This is equivalent to seeking to have 95% confidence that the effect is real rather than random, and is focused on controlling the Type I error (the likelihood that you decide that an effect is real when it's random). There is no corresponding scientific recommendation for the Type II error (beta, the likelihood that you will conclude an effect is random when it's real), although some suggest setting it to .20 (Diamond, 1981). The rationale behind the emphasis on controlling the Type I error is that it is better to delay the introduction of good information into the scientific database (a Type II error) than to let erroneous information in (a Type I error).

In industrial evaluation, the appropriate values for Type I and Type II errors depend on the demands of the situation – whether the cost of a Type I or Type II error would be more damaging to the organization. Because we are often resource-constrained, especially with regard to making timely decisions to compete in dynamic marketplaces, this paper has used measurement criteria (such as 90% confidence rather than 95% confidence[16] and fairly large values for *d*) that seek a greater balance between Type I and Type II errors than is typical in work designed to result in scientific publications. For an excellent discussion of this topic for usability researchers, see Wickens (1998). For other technical issues and perspectives, see Landauer (1997).

Another way to look at the issue is to ask the question, "Am I typically interested in small high-variability effects or large low-variability effects?" The correct answer depends on the situation, but in usability testing, the emphasis is on the detection of large low-variability effects (either large performance effects or frequently-occurring problems). This also makes sense for other aspects of the total user experience (TUE), such as measuring the effects of different advertising methods or different product support strategies. If the differences between advertising methods or support strategies are subtle and highly variable, then it doesn't make sense to make a large investment in their investigation unless you can demonstrate that decisions based on their investigation will increase profits to the point of recovering the investment in the study. You shouldn't need a large sample to verify the existence of large low-variability effects[17].

Coming from a different tradition than usability research, many market research texts provide rules of thumb recommending large sample sizes. For example, Aaker and Day (1986) recommend a minimum of 100 per group, with 20-50 for subgroups. For national surveys with many subgroup analyses, the typical total sample size is 2500 (Sudman, 1976). These rules of thumb do not make any formal contact with statistical theory, and may in fact be excessive, depending on the goals of the study. Other market researchers (for example, Banks, 1965) do promote a careful evaluation of the goals of a study.

> It is urged that instead of a policy of setting uniform requirements for type I and II errors, regardless of the economic consequences of the various decisions to be made from experimental data, a much more flexible approach be adopted. After all, if a researcher sets himself a policy of always choosing the apparently most effective of a group of alternative treatments on the basis of data from

[16] Nielsen (1997) suggests that 80% confidence is appropriate for practical development purposes.

[17] Some writers equate sample size with population coverage. This isn't true. A small sample size drawn from the right population provides better coverage than a large sample size drawn from the wrong population. The statistics involved in computing confidence intervals from small samples compensate for the potentially smaller variance in the small sample by forcing the confidence interval to be wider than that for a larger sample.

> unbiased surveys or experiments and pursues this policy consistently, he will find that in the long run he will be better off than if he chose any other policy. This fact would hold even if none of the differences involved were statistically significant according to our usual standards or even at probability levels of 20 or 30 percent. (Banks, 1965, p. 252)

Finally, Alreck and Settle (1985) provide an excellent summary of the factors indicating appropriate use of large and small samples.

Use a large sample size when:

1. Decisions based on the data will have very serious or costly consequences
2. The sponsors (decision-makers) demand a high level of confidence
3. The important measures have high variance
4. Analyses will require the dividing of the total sample into small subsamples
5. Increasing the sample size has a negligible effect on the cost and timing of the study
6. Time and resources are available to cover the cost of data collection

Use a small sample size when:

1. The data will determine few major commitments or decisions
2. The sponsors (decision-makers) require only rough estimates
3. The important measures have low variance
4. Analyses will use the entire sample, or just a few relatively large subsamples
5. Costs increase dramatically with sample size
6. Budget constraints or time limitations limit the amount of data you can collect

Some tips on reducing variance

Because measurement variance is such an important factor in sample size estimation for these types of studies, it generally makes sense to attempt to manage variance (although in some situations, such management is sometimes out of a practitioner’s control). Here are some ways to reduce variance:

- Make sure participants understand what they are supposed to do in the study. Unless potential participant confusion is part of the evaluation (and it sometimes is), it can only add to measurement variance.
- One way to accomplish this is through practice trials that allow participants to get used to the experimental situation without unduly revealing study-relevant information.
- If appropriate, use expert rather than novice participants. Almost by definition, expertise implies reduced performance variability (increased automaticity) (Mayer, 1997). With regard to reducing variance, the farther up the learning curve, the better.

- A corollary of this is that if you need to include both expert and novice users, you should be able to get equal measurement precision for both groups with unequal sample sizes (fewer experts required than novices[18]).
- Study simple rather than complex tasks.
- Use data transformations for measurements that typically exhibit correlations between means and variances or standard deviations. For example, frequency counts often have proportional means and variances (treated with the square root transformation), and time scores often have proportional means and standard deviations (treated with the logarithmic transformation (Myers, 1979).
- For comparative studies, use within-subjects designs rather than between-subjects designs whenever possible.
- Keep user groups as homogeneous as possible[19].

Keep in mind that it is reasonable to use these tips only when their use does not adversely affect the validity and generalizability of the study. Having a valid and generalizable study is far more important than reducing variability. (For information on enhancing generalizability, see the section in this paper on *Substitute Audiences*.)

Some tips for estimating unknown variance

Parasuraman (1986) described a method for estimating variability if you have an idea about the largest and smallest values for a population of measurements, but don't have the information you need to actually calculate the variability. Estimate the standard deviation (the square root of the variability) by dividing the difference between the largest and smallest values by 6. This technique assumes that the population distribution is normal, and then takes advantage of the fact that 99% of a normal distribution will lie in the range of plus or minus three standard deviations of the mean.

Churchill (1991) provided a list of typical variances for data obtained from rating scales. Because the number of items in the scale affects the possible variance (with more items leading to more variance), the table takes the number of items into account. For five-point scales, the typical variance is 1.2-2.0; for seven-point scales[20] it is 2.4-4.0; and for ten-point scales it is 3.0-7.0. Because data obtained using rating scales tends to have a more uniform than normal

[18] Which is good, because experts are typically harder than novices to acquire as participants.

[19] But keep in mind that while this reduces variability, it can simultaneously pose a threat to a study's external validity (Campbell and Stanley, 1963) if the test group is more homogenous than the population under study. See the section, presented later in this paper, on the use of substitute audiences and the factors that increase the generalizability of a study.

[20] Psychometric theory can quantify the relationship between the number of scale steps and item reliability (Nunnally, 1976), with seven scale steps appearing to be a reasonable number to use. Research conducted at IBM (Lewis, 1993; Lewis, 1995) is consistent with the suggestion to use seven rather than five scale steps.

distribution, he advises using a number nearer the high end of the listed range when estimating sample sizes[21].

Sample size estimation for problem-discovery (formative) studies

Estimating sample sizes for studies that have the primary purpose of discovering the problems in an interface depends on having an estimate of *p*, the average likelihood of problem occurrence. As with comparative studies, this estimate can come from previous studies using the same method and similar system under evaluation, or can come from a pilot study. For standard scenario-based usability studies, the literature contains large-sample examples with *p* ranging from .16 to .42 (Lewis, 1994a). For heuristic evaluations, the reported value of *p* from large-sample studies ranges from .22 to .60 (Nielsen and Molich, 1990).

When estimating *p* from a small sample, it is important to adjust its estimated value because small-sample (for example, fewer than 20 participants) estimates of *p* have a bias that results in potentially substantial overestimation (Hertzum and Jacobsen, in press). A series of recent Monte Carlo experiments (Lewis, 2000a-e) have demonstrated that a formula combining Good-Turing discounting with a normalization procedure provides a very accurate adjustment of initial estimates of *p*, even when the sample size for that initial estimate has as few as two participants. This formula for the adjustment of *p* is:

$$[9]\ truep = ½[(estp - 1/n)(1 - 1/n)] + ½[estp/(1+GTadj)]$$

where *GTadj* is the Good-Turing adjustment to probability space (which is the proportion of the number of problems that occurred once divided by the total number of different problems). The *estp*/(1+*GTadj*) component in the equation produces the Good-Turing adjusted estimate of *p* by dividing the observed, unadjusted estimate of *p* (*estp*) by the Good-Turing adjustment to probability space. The (*estp* – 1/*n*)(1 – 1/*n*) component in the equation produces the normalized estimate of *p* from the observed, unadjusted estimate of *p* and *n* (the sample size used to estimate *p*). The reason for averaging these two different estimates is that the Good-Turing estimator tends to overestimate the true value of p, and the normalization tends to underestimate it (Lewis, 2000b, 2000d). For more details and experimental data supporting the use of this formula for estimates of *p* based on sample sizes from two to ten participants, see Lewis (2000a).

[21] Measurement theorists who agree with S. S. Steven's (1951) principle of invariance might yell 'foul' at this point because they believe it is not permissible to calculate averages or variances from rating scale data. There is considerable controversy on this point (for example, see Lord, 1953, Nunnally, 1976, or Harris, 1985). My own experiences with rating scale data have convinced me that taking averages and conducting *t*-tests on the data provides far more interpretable and consistent results than the alternative of taking medians and conducting Mann-Whitney *U*-tests (Lewis, 1993). You do have to be careful not to act as if rating scale data is interval data rather than ordinal data when you make claims about the meaning of the outcomes of your statistical tests. An average satisfaction rating of 4 might be better than an average rating of 2, but you can't claim that it is twice as good.

Once you have an appropriate estimate for *p*, you use the formula 1-(1-*p*)^n (derived both from the binomial probability formula, Lewis, 1982, 1990, 1994a, and from the Poisson probability formula, Nielsen and Landauer, 1993) for various values of *n* from, say, 1 to 20, to generate the curve of diminishing returns expected as a function of sample size. It is possible to get even more sophisticated, taking into account the fixed and variable costs of the evaluation (especially the variable costs associated with the study of additional participants) to determine when running an additional participant will result in costs that exceed the value of the additional problems discovered (Lewis, 1994a).

The recent Monte Carlo experiment reported in Lewis (2000a) demonstrated that an effective strategy for planning the sample size for a usability study is first to establish a problem discovery goal (for example, 90% or 95%). Run the first two participants and, based on those results, calculate the adjusted value of *p* using the equation in [9]. This provides an early indication of the likely required sample size, which might estimate the final sample size exactly or, more likely, underestimate by one or two participants. Collect data from two more participants (for a total of four). Recalculate the adjusted estimate of *p* using the equation in [9] and project the required sample size using 1-(1-*p*)^n. The estimated sample size requirement based on data from four participants will generally be highly accurate, allowing accurate planning for the remainder of the study. Practitioners should do this even if they have calculated a preliminary estimate of the required sample size from an adjusted value for *p* obtained from a previous study.

Figure 1 shows the predicted discovery rates for problems of differing likelihoods of observation during a usability study. Several independent studies have verified that these types of predictions fit observed data very closely for both usability and heuristic evaluations (Lewis, 1994a; Nielsen and Landauer, 1993; Nielsen and Molich, 1990; Virzi, 1990, 1992; Wright and Monk, 1991). Furthermore, the predictions work both for predicting the discovery of individual problems with a given probability of detection and for modeling the discovery of members of sets of problems with a given mean probability of detection (Lewis, 1994a). For usability studies, the sample size is the number of participants. For heuristic evaluations, the sample size is the number of evaluators.

Figure 1. Predicted Discovery as a Function of Problem Likelihood

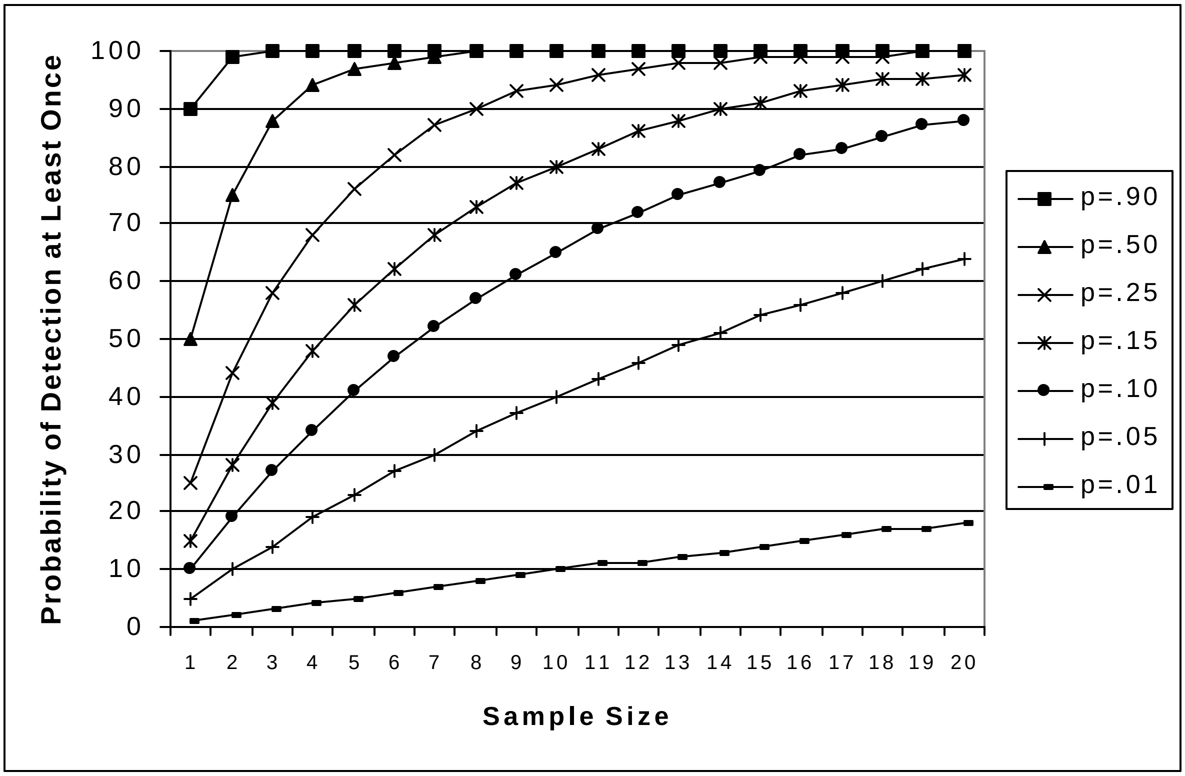

Table 10 (from Lewis, 1994a) shows problem detection sample size requirements as a function of problem detection probability and the cumulative likelihood of detecting the problem at least once during the study. The required sample size for detecting the problem twice during a study appears in parentheses.

Table 10. Sample Size Requirements for Problem Discovery (Formative) Studies

Problem Occurrence Probability	**Cumulative Likelihood of Detecting the Problem at Least Once (Twice)**					
	0.50	**0.75**	**0.85**	**0.90**	**0.95**	**0.99**
0.01	68 (166)	136 (266)	186 (332)	225 (382)	289 (462)	418 (615)
0.05	14 (33)	27 (53)	37 (66)	44 (76)	57 (91)	82 (121)
0.10	7 (17)	13 (26)	18 (33)	22 (37)	28 (45)	40 (60)
0.15	5 (11)	9 (17)	12 (22)	14 (25)	18 (29)	26 (39)
0.25	3 (7)	5 (10)	7 (13)	8 (14)	11 (17)	15 (22)
0.50	1 (3)	2 (5)	3 (6)	4 (7)	5 (8)	7 (10)
0.90	1 (2)	1 (2)	1 (3)	1 (3)	2 (3)	2 (4)

To use this information to establish a usability sample size, you need to determine three things.

First, what is the average likelihood of problem detection probability (*p*)? This plays a role similar to the role of variance in the previous examples. If you don't know this value (from previous studies or a pilot study), then you need to decide on the lowest problem detection probability that you want to (or have the resources to) tackle. The smaller this number, the larger the required sample size.

Second, you need to determine what proportion of the problems that exist at that level you need (or have the resources) to discover during the study (in other words, the cumulative likelihood of problem detection). The larger this number, the larger the required sample size.

Finally, you need to decide whether you are willing to take single occurrences of problems seriously or if problems must appear at least twice before receiving consideration. Requiring two occurrences results in a larger sample size.

Lewis (1994a) created a return-on-investment (ROI) model to investigate appropriate cumulative problem detection targets. It turned out that the appropriate target depended on the average problem detection probability in the evaluation – the same value that has a key role in determining the sample size. The model indicated that if the expected value of *p* was small (say, around 0.10), practitioners should plan to discover about 86% of the problems. If the expected value of *p* was larger (say, around .25 or .50), practitioners should plan to discover about 98% of the problems. For expected values of *p* between 0.10 and 0.25, practitioners should interpolate between 86 and 97% to determine an appropriate goal for the percentage of problems to discover.

Contrary to expectation, the cost of an undiscovered problem had a minor effect on sample size at maximum ROI, but it had a strong effect on the magnitude of the maximum ROI. Usability practitioners should be aware of these costs in their settings and their effect on ROI (Boehm, 1981), but these costs have relatively little effect on the appropriate sample size for a usability study.

In summary, the law of diminishing returns, based on the cumulative binomial probability formula, applies to problem discovery studies. To use this formula to determine an appropriate sample size, practitioners must form an idea about the expected value of *p* (the average likelihood of problem detection) for the study and the percentage of problems that the study should uncover. Practitioners can use the ROI model from Lewis (1994a) or their own ROI formulas to estimate an appropriate goal for the percentage of problems to discover and can examine data from their own or published usability studies to estimate *p* (which published studies to date indicate can range from 0.16 to 0.60). With these two estimates, practitioners can use Table 10 to estimate appropriate sample sizes for their usability studies.

It is interesting to note that a new product that has not yet undergone any usability evaluation is likely to have a higher *p* than an established product that has gone through several development iterations (including usability testing). This suggests that it is easier (takes fewer participants) to improve a completely new product than to improve an existing product (as long as that existing product has benefited from previous usability evaluation). This is related to the idea that usability evaluation is a hill-climbing procedure, in which the results of a usability evaluation are applied to a product to push its usability up the hill. The higher up the hill you go, the more difficult it becomes to go higher, because you have already weeded out the problems that were easy to find and fix.

Practitioners who wait to see a problem at least twice before giving it serious consideration can see from Table 10 the sample size implications of this strategy. Certainly, all other things being equal, it is more important to correct a problem that occurs frequently than one that occurs infrequently. However, it is unrealistic to assume that the frequency of detection of a problem is the only criterion to consider in the analysis of usability problems. The best strategy is to consider problem frequency and impact simultaneously[22] to determine which problems are most important to correct rather than establishing a cutoff rule such as “fix every problem that appears two or more times.”

Note that in contrast to the results reported by Virzi (1992), the results reported by Lewis (1994a) did not indicate any consistent relationship between problem frequency and impact (severity). The safest strategy is for practitioners to assume independence of frequency and impact until further research resolves the discrepancy between the outcomes of these studies.

It is important for practitioners to consider the risks as well as the gains when using small samples for usability studies. Although the diminishing returns for inclusion of additional participants strongly suggest that the most efficient approach is to run a small sample (especially if *p* is high, if the study will be iterative, and if undiscovered problems will not have dangerous or expensive outcomes), human factors engineers and other usability practitioners must not become complacent regarding the risk of failing to detect low-frequency but important problems[23].

[22] When prioritizing usability problems, it is important to consider not only frequency, but also impact (which could include both information about how damaging the problem is to users and how frequently in the field users will perform the task that elicited the problem). The recommended sample size strategy for problem-discovery studies only addresses the sample size required to have some confidence that you will uncover a reasonable proportion of the problems that are available for discovery. As you discover problems, you can also get an idea about their impact on users (typically a minor annoyance vs. typically causes the user to fail to complete the scenario). When constructing scenarios, you might well have an idea (or even data) regarding the relative frequency of use of the functions under investigation in the study (for example, in speech dictation systems, the tasks of dictating and correcting are far more frequent than the task of creating a dictation macro). Clearly, the problems that should receive the highest priority for repair from a usability point of view are those that have high frequency and high impact. Note that the literature has shown inconsistent results regarding the relative rates of discovery of problems as a function of impact (severity). However, it appears that high-impact problems are either discovered faster or, more likely, at the same rate as low-impact problems. Because the strongest evidence is in favor of equal discovery rates as a function of impact (Lewis, 1994a), the sample size estimation procedures given here do not take impact (severity) into account.

[23] Note that one could argue that the true number of possible usability problems in any interface is essentially infinite, with an essentially infinite number of problems with non-zero probabilities that are extremely close to zero. For the purposes of determining sample size, the *p* we are really dealing with is the *p*

Example 9. Sample Sizes for Usability Studies

Here are some examples using Table 10 as an aid in selecting an appropriate sample size for a usability study.

A. Given the following set of discovery criteria:

- Detect problems with average probability of: 0.25
- Minimum number of detections required: 1
- Planned proportion to discover: 0.90

The appropriate sample size is 8 participants.

B. Given the same discovery criteria, except the practitioner requires problems to be detected twice before receiving serious attention:

- Detect problems with average probability of: 0.25
- Minimum number of detections required: 2
- Planned proportion to discover: 0.90

The appropriate sample size would be 14 participants.

C. Returning to requiring a single detection, but increasing the planned proportion to discover to .99:

- Detect problems with average probability of: 0.25
- Minimum number of detections required: 1
- Planned proportion to discover: 0.99

The appropriate sample size would be 15 participants.

D. Given extremely stringent discovery criteria:

- Detect problems with average probability of: 0.01
- Minimum number of detections required: 1
- Planned proportion to discover: 0.99

The required sample size would be 418 participants (an unrealistic requirement in most settings, implying unrealistic study goals).

that represents the number of discovered problems divided by the number of discoverable problems, where the definition of a discoverable problem is vague, but almost certainly constrained by details of the experimental setting, such as the studied scenarios and tasks and the skill of the observer(s). Despite this, these techniques seem to work reasonably well in practice.

Note that there is no requirement that you run the entire planned sample through the usability study before reporting clear problems to development and getting those problems fixed before continuing. These required sample sizes are total sample sizes, not sample sizes per iteration. The importance of iterative testing to the development of usable products is well known (Gould, 1988). The following testing strategy promotes efficient iterative problem discovery studies.

1. Start with an expert (heuristic) evaluation or one-participant pilot study to uncover the majority of high-frequency problems. Correct as many of these problems as possible before starting the iterative cycles with Step 2. List all unresolved problems and carry them to Step 2.

2. Watch a small sample (for example, three or four participants) use the system. Record all observed usability problems. Calculate an adjusted estimate of p based on these results and re-estimate the required sample size.

3. Redesign based on the problems discovered. Focus on fixing high frequency and high impact problems. Fix as many of the remaining problems as possible. Record any outstanding problems so they can remain open for all following iterations.

4. Continue iterating until you have reached your sample size goal (or must stop for any other reason, such as you ran out of time).

5. Record any outstanding problems remaining at the end of testing and carry them over to the next product for which they are applicable.

This strategy blends the benefits of large and small sample studies. During each iteration, you observe only three participants before redesigning the system. Therefore, you can quickly identify and correct the most frequent problems (which means you waste less time watching the next set of participants encounter problems that you already know about). With five iterations, for example, the total sample size would be 15 participants. With several iterations you will identify and correct many less frequent problems because you record and track the uncorrected problems through all iterations[24].

Example 10. Evaluating sample size effectiveness with fixed n

Suppose you know you only have time to run a limited number of participants, are willing to treat a single occurrence of a problem seriously, and want to determine what you can expect to get out of a problem-discovery study with that number of participants. If that number were six, for example, examination of Table 10 indicates:

[24] Note that using this sort of iterative procedure lowers p as you go along. The value of p in the system you end with will be lower than the p you started with (as long as the process of fixing problems doesn't create as many other problems).

- You are almost certain to detect problems that have a .90 likelihood of occurrence (it only takes two participants to have a 99% cumulative likelihood of seeing the problem at least once).
- You are almost certain (between 95 and 99% likely) to detect problems that have a .50 likelihood of occurrence (for this likelihood of occurrence, the required sample size at 95% is 5, and at 99% is 7).
- You've got a reasonable chance (about 80% likely) of detecting problems that have a .25 likelihood of occurrence (for this likelihood of occurrence, the required sample size at 75% is 5,and at 85% is 7).
- You have a little better than even odds of detecting problems that have a .15 likelihood of occurrence (the required sample size at 50% is 5).
- You have a little less than even odds of detecting problems that have a .10 likelihood of occurrence (the required sample size at 50% is 7).
- You are not likely to detect many of the problems that have a likelihood of occurrence of .05 or .01 (for these likelihoods of occurrence, the required sample size at 50% is 14 and 68 respectively).

This analysis illustrates that although a problem-discovery study with a sample size of six participants will typically not discover problems with very low likelihoods of occurrence, the study is almost certainly worth conducting (a finding consistent with the six-participant rule-of-thumb for usability studies that has been around since the late 1970s).

Applying this procedure to a number of different sample sizes produces Table 11. The cells in Table 11 are the probability of having a problem with a specified occurrence probability happen at least once during a usability study with the given sample size.

Table 11. Likelihood of discovering problems of probability p at least once in a study with sample size n

Problem Occurrence Probability (p)	**Sample Size (*n*)**						
	3	**6**	**9**	**12**	**15**	**18**	**21**
.01	0.03	0.06	0.09	0.11	0.14	0.17	0.19
.05	0.14	0.26	0.37	0.46	0.54	0.60	0.66
.10	0.27	0.47	0.61	0.72	0.79	0.85	0.89
.15	0.39	0.62	0.77	0.86	0.91	0.95	0.97
.25	0.58	0.82	0.92	0.97	0.99	0.99	1.00
.50	0.88	0.98	1.00	1.00	1.00	1.00	1.00
.90	1.00	1.00	1.00	1.00	1.00	1.00	1.00

Example 11. Sample Sizes for Heuristic Evaluations

The range of *p* reported in the literature appears to be similar for usability and heuristic evaluations, and the cumulative binomial probability formula ($1-(1-p)^n$) appears to accurately model problem discovery for both methods (Lewis, 1994a; Nielsen and Molich, 1990; Nielsen and Landauer[25], 1993). Nielsen and Landauer (1993) reported the value of *p* for five usability studies and six heuristic evaluations, with an average of 0.28 for the usability studies and 0.31 for the heuristic evaluations. Therefore, it seems that both types of evaluations would have similar sample size requirements.

With heuristic evaluations, some of the factors that can affect *p* are:

- The stage of the usability lifecycle – as mentioned above for usability tests, it is generally easier to find problems in new interfaces than in older, more polished interfaces.
- The degree of implementation – it is harder to find usability problems in paper mockups than in running prototypes.
- The skill of the heuristic evaluators – evaluators with usability expertise have a higher problem discovery rate than evaluators without such expertise, and evaluators with expertise in both usability and the application domain have the highest discovery rates[26].

A potential downside of heuristic evaluations is the difficulty of acquiring skilled evaluators with usability (and, if possible, domain) expertise. This provides further support for Step 1 of the iterative testing strategy, with the following potential modification (which is still a blend of heuristic and usability evaluation, with just a little more heuristic evaluation):

1. Start with a single heuristic evaluation (preferably by an evaluator with both usability and domain knowledge) to uncover the majority of high-frequency problems. Correct as many of these problems as possible before starting Step 2. List all unresolved problems and carry them to Step 2.

2. Conduct an additional set of heuristic evaluations with one to three usability or usability/domain experts. Redesign based on the problems discovered, focusing on problems with high consensus. Record any outstanding problems to carry to the next step.

3. Begin usability evaluation[27]. Watch a small sample (for example, three participants) use the system. Record all observed usability problems. Redesign based on the problems

[25] As a technical note, Nielsen and Landauer (1993) derived the same formula – $1-(1-p)^n$ – from a Poisson process model, constant probability path independent (T.K. Landauer, personal communication, December 20, 1999).

[26] Note that the skill of the observer in a usability evaluation is analogous to evaluator skill in heuristic evaluation. Because "both detection of usability problems and selection of the most severe problems are subject to considerable individual variability" (Jacobsen, Hertzum, and John, 1998, p. 1336) in usability testing, it is important to use highly skilled observers.

[27] Why do user-based usability evaluation at all rather than just continuing to perform iterative reviews? In the field of usability evaluation, problem-discovery studies with users is widely considered to be the gold

discovered. Focus on fixing high frequency and high impact problems. Fix as many of the remaining problems as possible. Record any outstanding problems so they can remain open for all following iterations.

4. Continue iterating the usability study (Step 3) until you have reached your sample size goal (or must stop for any other reason, such as you ran out of time).

5. Record any outstanding problems remaining at the end of testing and carry them over to the next product for which they are applicable.

Some tips on managing p

Because *p* (the average likelihood of problem discovery) is such an important factor in sample size estimation for usability and heuristic evaluations, it generally makes sense to attempt to manage it (although in some situations, such management is sometimes out of a practitioner's control). Here are some ways to increase *p*:

- Use highly-skilled observers for usability studies[28].
- Focus evaluation on new products with newly-designed interfaces rather than older, more refined interfaces.
- Study less-skilled participants in usability studies (as long as they are appropriate participants).
- Make the user sample as heterogeneous as possible, within the bounds of the population to which you plan to generalize the results.
- Make the task sample as heterogeneous as possible.
- Emphasize complex rather than simple tasks.
- For heuristic evaluations, use examiners with usability and application domain expertise.
- For heuristic evaluations, if you must make a tradeoff between having a single evaluator spend a lot of time examining an interface versus having more examiners spend less time each examining an interface, choose the latter option (Dumas, Sorce, and Virzi, 1995; Virzi, 1997).

standard for detecting problems. This is because the problems discovered via review might or might not turn out to be problems when users actually use the system (in other words, some problems discovered via review might turn out to be false alarms). In general, most usability evaluators believe that problems discovered via user testing are more likely to be real problems (with probable impact to users in the field) rather than false alarms.

[28] The task of the observer in a usability study involves classical signal-detection issues (Massaro, 1975). The observer monitors participant behavior, and at any given moment must decide whether that observed behavior is indicative of a usability problem. Thus, there are two ways for an observer to make correct decisions (rejecting non-problem behaviors correctly, identifying problem behaviors correctly) and two ways to make incorrect decisions (identifying non-problem behaviors as indicative of a usability problem, failing to identify problem behaviors as indicative of a usability problem). In signal detection terms, the names for these right and wrong decisions are Correct Rejection, Hit, False Alarm, and Miss. The rates for these types of decisions depend independently on both the skill and the bias of the observer (in signal detection terms, d' and β). Applying signal detection theory to the assessment of the skill and bias of usability test observers is a potentially rich, but to date untapped, area of research, with potential application for both selection and training of observers.

Note that some of the tips for increasing *p* are the opposite of those that reduce variability. This is reasonable because increasing *p* (the likelihood that users experience problems with the system) will necessarily increase the mean time users need to complete tasks, which will typically increase the variance for the task times.

International tests

There is no fundamental difference between international and non-international tests with regard to sample size estimation. If a researcher has reason to believe that an international group varies more or less on a key comparative variable, then he or she should adjust the sample size for that group accordingly to maintain equal precision of measurement with the domestic group, if possible. It is reasonable to apply a similar strategy with regard to expected values of *p* for problem-discovery studies.

Different types of design evaluation activities

There is a fundamental difference in sample size estimation for comparative/parameter estimating and problem-discovery studies, so you need to be able to tell the difference. Fortunately, the difference is easy to determine. Any activity that resembles a traditional experiment with independent variables is comparative. If you are trying to estimate a value for a population of users (for example, acceptability and interest ratings for advertising materials planned for use with a product), you are planning a population parameter estimation study. Any activity that involves the assessment of a single system for the purpose of uncovering problems in the use of that system is a problem-discovery study. The two most common types of problem-discovery studies are standard usability studies and heuristic evaluations. Once having made this determination and providing estimates for the key variables (variance, confidence and precision for comparative/estimating studies; *p* and proportion of problems to discover for problem-discovery studies), sample size estimation is typically a trivial exercise.

There is no evidence that there is any systematic difference in variability or *p* as a function of the type of product under evaluation (for example, hardware vs. software or user interaction with a graphical user interface vs. user interpretation of documentation vs. user assessments of marketing material). This lack of evidence could be a simple reflection of our current state of ignorance on this topic. It seems likely, however, that measurement variability is far more affected by user and task (for example, longer, more complex tasks performed by novice users will have greater variability than shorter, less complex tasks performed by experts) than by whether you are evaluating a hardware or software product.

Non-traditional areas of evaluation

Non-traditional areas of evaluation include activities such as the evaluation of visual design and marketing materials. As with traditional areas of evaluation, the first step is to determine if the evaluation is comparative/parameter estimation or problem-discovery.

Part of the problem with non-traditional areas is that we have less information regarding the values of the variables needed to estimate sample sizes. Perhaps one of the things we should start doing is recording those values from different IBM activities to develop a database from which one could extract appropriate values for the various activities we do that require or would benefit from sample size estimation.

Another issue for non-traditional areas of evaluation is whether these areas are inherently focused on detecting more subtle effects than is the norm in usability testing, which has a focus on large low-variability effects (and correspondingly small sample size requirements). Determining this requires the involvement of someone with domain expertise in these non-traditional areas. It seems, however, that even these non-traditional areas would benefit from focusing on the discovery of large low-variability effects. Only if there was a business case that the investment in a study to detect small, highly-variable effects would ultimately pay for itself should you conduct such a study.

For example, in *The Survey Research Handbook*, Alreck and Settle (1985) point out that the reason that survey samples rarely contain fewer than several hundred respondents is due to the cost structure of surveys. The fixed costs of the survey include activities such as determining information requirements, identifying survey topics, selecting a data collection method, writing questions, choosing scales, composing the questionnaire, etc. For this type of research, the additional or ‘marginal’ cost of including hundreds of additional respondents can be very small relative to the fixed costs. Contrast this with the cost (or feasibility) of adding participants to a usability study in which there might be as little as a week or two between the availability of testable software and the deadline for affecting the product, with resources limiting the observation of participants to one at a time and the test scenarios requiring two days to complete.

“Since the numbers don’t know where they came from, they always behave just the same way, regardless.” (Lord, 1953, p. 751) What differs for non-traditional areas of evaluation isn’t the behavior of numbers or statistical procedures, but the researchers’ goals and economic realities.

Need for a database of variance and p

Nielsen (1997) surveyed 36 published usability studies, and found that the mean standard deviation for measures of expert performance was 33% of the mean value of the usability measure. For novice-user learning the mean standard deviation was 46%, and for measures of error rates the value was 59%. With regard to *p*, Nielsen and Landauer (1993) reported its value for five usability studies and six heuristic evaluations, with an average of 0.28 for the usability studies and 0.31 for the heuristic evaluations.

This information provides a starting point for estimating sample sizes for usability evaluations. It could also be valuable to bring together a sample of the usability work done inside IBM (both comparative and formative) and, without necessarily revealing confidential information, create a database of estimated variances and *p*s for use in initial sample size estimation. With such a database, we might be able to begin to more quantitatively address issues of sample sizes for international testing and different or non-traditional areas of evaluation.

Substitute Audiences

The essential problem with using substitute audiences is the problem of generalization. Almost all of the standard usability resources state the importance of sampling from the population that will use the system under study, and point out that there is some risk associated with failing to do so (Chapanis, 1988; Holleran, 1991; Nielsen, 1997).

Nonprobability sampling

Market researchers (Parasuraman, 1986) define certain types of nonprobability-sampling methods, including convenience sampling (using anyone who is convenient to the researcher, such as colleagues or secretaries) and judgement sampling (the researcher exerts some effort in selecting a sample that is appropriate according to some criteria, but isn't sampling from the target population). According to Parasuraman:

> Even though convenience and judgement samples are unlikely to provide perfect representations of the ideal population for a study, they can be used under certain conditions. Such samples may be appropriate when the basic purpose of a research study is to generate some initial insights rather than to make any generalizations. Exploratory-research projects … often involve the use of convenience sampling or, if time and other resources permit, judgement sampling. (p. 502-503)

A related issue is that, strictly speaking, the application of statistical procedures to data depend on those data coming from a randomly selected sample (Campbell and Stanley, 1963). Again speaking strictly, the appropriate way to sample randomly from a population is to enumerate all the members of that population and to select from the population using a procedure that gives all members of the population an equal chance of selection.

Given that we are never likely to achieve these strict standards, does this mean we should give up on statistically qualified measurement? We certainly should not, but we should realize the effect that nonrandom sampling likely has on the measurement variance we get in a study versus the true population variance. The most likely effect is that the variance we see is to some unknown extent less than the true population variance, which means that our measurements aren't quite as precise as we think they are[29]. However, the fact that we have a literature with decades of experimentation and measurement (with much of the measurement appearing to be valid and experiments that are replicable) supports the hypothesis that nonprobability sampling usually works pretty well. This is especially true if you have good reasons to believe that a study's participants match the members of the target population reasonably well.

[29] I don't have a citation for this claim, but have based it on the reasoning that a convenience sample would probably be more homogeneous than the population at large. On the other hand, if a researcher deliberately introduced an excessive amount of participant heterogeneity into the convenience sample, the variance observed in the study could well be greater than the true population variability.

Matching participants to populations

"In designing studies that are to be used to predict behavior in a specific situation, the guiding principle can be summed up in one word: *similarity.* The study should be as similar as possible to the real situation. This means that subjects, apparatus, tasks, dependent variables and the test environment should simulate or match those of the application as closely as possible." (Chapanis, 1988, p. 263)

As Chapanis (1988) pointed out, there are many aspects to achieving a generalizable study, one of which is the similarity of participants to the target population. If we have not been able to acquire participants from the target population, though, then how can we approach the problem of measuring the similarity between the participants we can acquire and the target population?

One step is to develop a taxonomy of the variables that affect human performance (where performance should include the behaviors of indicating preference and other choice behaviors). Gawron, Drury, Czaja and Wilkins (1989) produced a human performance taxonomy during the development of a human performance expert system. They reviewed existing taxonomies and filled in some missing pieces. They structured the taxonomy as having three top levels: environment, subject (person), and task[30]. The resulting taxonomy takes up 12 pages in their paper, and covers many areas which would normally not concern a usability practitioner working in the field of computer system usability (for example, ambient vapor pressure, gravity, acceleration, etc.). Some of the key human variables in the Gawron et al. (1989) taxonomy that could affect human performance with computer systems are:

- Physical Characteristics
  - Age
  - Agility
  - Handedness
  - Voice
  - Fatigue
  - Gender
  - Body and body part size

- Mental State
  - Attention span
  - Use of drugs (both prescription and illicit)
  - Long-term memory (includes previous experience)
  - Short-term memory
  - Personality traits

[30] Note the relationship of these to the variables mentioned above by Chapanis (1988) as the important variables defining similarity for generalization. Also, although task and environmental representativeness are also very important (Holleran, 1991; Lewis, 1994a; Nielsen, 1997), the focus of discussion in this paper is on the human.

  - Work schedule

- Senses
  - Auditory acuity
  - Tone perception
  - Tactual
  - Visual accommodation
  - Visual acuity
  - Color perception

These variables certainly give us something to work with as we attempt to describe how participants and target populations are similar or different. The Gawron et al. (1989) taxonomy does not provide much detail with regard to some individual differences that some researchers have hypothesized affect human performance or preference with respect to the use of computer systems: personality traits and computer-specific experience.

Aykin and Aykin (1991) performed a comprehensive review of the published studies to that date that involved individual differences in human-computer interaction (HCI). Table 12 lists the individual differences for which they found HCI studies, the method used to measure the individual difference, and whether there was any indication from the literature that manipulation of that individual difference in an experiment led to a crossed interaction[31].

[31] In statistical terminology, an interaction occurs whenever an experimental treatment has a different magnitude of effect depending on the level of a different, independent experimental treatment. A crossed interaction occurs when the magnitudes have different signs, indicating reversed directions of effects. As an example of an uncrossed interaction, consider the effect of turning off the lights on the typing throughput of blind and sighted typists. The performance of the sighted typists would probably be worse, but the presence or absence of light shouldn't affect the performance of the blind typists. As an extreme example of a crossed interaction, consider the effect of language on task completion for people fluent only in French or English. When reading French text, French speakers would outperform English speakers, and vice versa.

Table 12. Results of Aykin and Aykin (1991) Review of Individual Differences in HCI

Individual Difference	**Measurement Method**	**Crossed Interactions**
Level of experience	Various methods	No
Jungian personality types	Myers-Briggs Type Indicator	No
Field dependence/ independence	Embedded Figures Test	Yes – field dependent participants preferred organized sequential item number search mode, but field independent subjects preferred the less organized keyword search mode (Fowler, Macaulay and Fowler, 1985)
Locus of control	Levenson test	No
Imagery	Individual Differences Questionnaire	No
Spatial ability	VZ-2	No
Type A/Type B personality	Jenkins Activity Survey	No
Ambiguity tolerance	Ambiguity Tolerance Scale	No
Sex	Unspecified	No
Age	Unspecified	No
Other (reading speed and comprehension, intelligence, mathematical ability)	Unspecified	No

For any of these individual differences, the lack of evidence for crossed interactions could be due to the paucity of research involving the individual difference or could reflect the probability that individual differences will not typically cause crossed interactions in HCI. It seems to be more likely that a change made to support a problem experienced by a person with a particular individual difference will either help other users or simply not affect their performance.

For example, John Black (personal communication, 1988) cited the difficulty that field dependent users had working with one-line editors at the time (decades ago) when that was the typical user interface to a mainframe computer. Switching to full-screen editing resulted in a performance improvement for both field dependent and independent users – an uncrossed interaction because both types of users improved, with the performance of field dependent users becoming equal to (thus improving more than) that of field independent users. Landauer (1997) cites another example of this, in which Greene, Gomez and Devlin (1986) found that young people with high scores on logical reasoning tests could master database query languages such as SQL with little training, but older or less able people could hardly ever master these languages. They also determined that an alternative way of forming queries, selecting rows from a truth table, allowed almost everyone to make correct specification of queries, independent of their abilities. Because this redesign improved the performance of less able users without diminishing the performance of the more able, it was an uncrossed interaction. In a more recent study, Palmquist and Kim (2000) found that field dependence affected the search performance of novices using a web browser (with field independent users searching more efficiently), but did not affect the performance of more experienced users.

If there is a reason to suspect that an individual difference will lead to a crossed interaction as a function of interface design, then it could make sense to invest the time (which can be considerable) to categorize users according to these dimensions. Another situation in which it could make sense to invest the time in categorization by individual difference would be if there were reasons to believe that a change in interface would greatly help one or more groups without adversely affecting other groups. (This is a strategy that one can employ when developing hypotheses about ways to improve user interfaces.) It always makes sense to keep track of user characteristics when categorization is easy (for example, age or sex).

Aykin and Aykin (1991) reported effects of users' levels of experience, but did not report any crossed interactions related to this individual difference. They did report that interface differences tended to affect the performance of novices, but had little effect on the performance of experts. It appears that behavioral differences related to user interfaces (Aykin and Aykin, 1991) and cognitive style (Palmquist and Kim, 2000) tend to fade with practice. Nonetheless, user experience has been one of the few individual differences to receive considerable attention in HCI (Fisher, 1991; Mayer, 1997; Miller, Stanney, and Wooten, 1997; Smith, Caputi, Crittenden, Jayasuriya, and Rawstorne, 1999).

According to Mayer (1997), relative to novices, experts have:

- better knowledge of syntax
- an integrated conceptual model of the system
- more categories for more types of routines
- higher level plans.

Fisher (1991) emphasized the importance of discriminating between computer experience (which he placed on a novice-experienced dimension) and domain expertise (which he placed on a naïve-expert dimension).

LaLomia and Sidowski (1990) reviewed the scales and questionnaires developed to assess computer satisfaction, literacy and aptitudes. None of the instruments they surveyed specifically addressed measurement of computer experience.

Miller, Stanney, and Wooten (1997) published the Windows Computer Experience Questionnaire (WCEQ), an instrument specifically designed to measure a person's experience with Windows 3.1. The questionnaire took about five minutes to complete and was reliable (Coefficient alpha = .74; test-retest correlation = .97). They found that their questionnaire was sensitive to three experiential factors: general Windows experience, advanced Windows experience, and instruction. They did not provide the answer choices or a key for the items' answers, but the questions were:

- Used Windows?
- Taken Windows courses?
- Taken Windows application courses?
- Total hours of Windows use?
- Hours per week of Windows use?
- Types of applications used?
- Functions can perform.
- What is a computer icon?
- What does Control-Esc do?
- What does Alt-Esc do?
- What does Alt-Tab do?

It is unfortunate that the authors have not yet generally released a Windows-95 version of the questionnaire, because the Windows 3.1 version is quick to administer and appears to be relevant to important experiential characteristics for computer usability assessment and generalization[32]. Such a tool would be useful for providing a way to quantitatively establish a match between participants and populations on the dimension of Windows computer experience.

Smith et al. (1999) distinguish between subjective and objective computer experience. The paper is relatively theoretical and "challenges researchers to devise a reliable and valid measure" (p. 239) for subjective computer experience, but does not offer one.

One user characteristic not addressed in any of the cited literature is one that becomes very important when designing products for international use – cultural characteristics. For example, it is extremely important that in adapting an interface for use by members of another country that all text is accurately translated. It is also important to be sensitive to the possibility that these types of individual differences might be more likely than others to result in crossed interactions. There is little research on the issue, so we can't be sure.

How to improve generalizability

One piece of advice appearing in a number of sources (Chapanis, 1988; Holleran, 1991; Landauer, 1997; Nielsen, 1997) is to build heterogeneity of participants, tasks, and measures into the study (what Landauer (1997, p. 208) refers to as "robustness in variation"). If the study has a single group, then that group should consist of people who match the target

[32] In the paper, the authors said they were working on a Windows 95 version of the questionnaire. I have contacted Kay Stanney and received the revised questionnaire, which appears to have reasonably good psychometric properties. I am still waiting to receive an answer key and to get responses to a few questions that I have. If anyone is interested in receiving a copy of the revised questionnaire to use, just let me know (jimlewis@us.ibm.com, July 7, 2000).

population on important characteristics but who are otherwise as dissimilar as possible from each other[33]. If the study has multiple groups, then the members of the groups should be reasonably homogeneous (minimizing within-group variance), but the groups should be dissimilar, consistent with the strategy of capturing the breadth of the target population (enhancing generalizability)[34].

The issue of not having access to the desired target population is something that many researchers face (for example, psychologists and market researchers as well as usability practitioners). Hopefully, the concurrent work on logistical issues will make it easier for IBM's UCD workers to get access to the desired populations. Otherwise, we will need to rely on careful record-keeping for the participants we do study, and will need to exercise careful, professional judgement on the extent to which we can generalize results. For example, in a problem-discovery study, to what extent is an observed problem associated with a physical human characteristic (such as buttons too small to press accurately), a cultural characteristic (such as the use of a culturally inappropriate symbol for an icon or culturally dissimilar e-commerce payment expectations), a knowledge-domain characteristic, etc.?

When to use multiple groups

Unless a product has different planned interfaces for different groups (for example, user vs. system administrator interfaces), then there doesn't seem to be much reason to parse the results (problems) of a problem discovery study as a function of user group.

For comparison studies, having multiple groups (for example, males and females or experts and novices) allows the assessment of potential interactions that might otherwise go unnoticed. Ultimately, the decision for one or multiple groups must be based on expert judgement and a few guidelines.

For example, consider sampling from different groups if you have reason to believe:

- There are potential and important differences among groups on key measures (Dickens, 1987)
- There are potential interactions as a function of group (Aykin and Aykin, 1991)
- The variability of key measures differs as a function of group
- The cost of sampling differs significantly from group to group

Gordon and Langmaid (1988) recommending the following approach to defining groups:

1. Write down all the important variables.

[33] Note that this is the opposite of what one would recommend for minimizing variance, but researchers would typically agree that generalizability is more important than minimal variance. Note that this is one of the tips given for maximizing *p* in a problem-discovery study.

[34] Chapanis (1988) also recommends a strategy of replicating earlier studies with variations in participants, variables, or procedures. For industrial work supporting the timely design of products, this will probably be a less frequently used approach.

2. If necessary, prioritize the list.
3. Design an ideal sample.
4. Apply common sense to collapse cells.

For example, suppose you start with 24 cells, based on the factorial combination of six demographic locations, two levels of experience, and the two levels of gender. Ask yourself whether you expect to learn anything new and important after completing the first few cells, or are you wasting your company's (or client's) money? Can you learn just as much from having one or a few cells that are homogeneous within cells and heterogeneous between cells with respect to an important variable, but are heterogeneous within cells with regard to other, less important variables?

Good generalizability and appropriate TUE coverage

Applying the concept of "robustness in variation" to usability studies with a focus on the total user experience (TUE) means including (to as great an extent as possible) scenarios that touch on all the different aspects of the user experience, from reactions to the marketing material to simulating calls to support. It also suggests that the participants in usability studies should come from diverse rather than homogeneous backgrounds, with the backgrounds as consistent as possible with those of the targeted user populations[35].

Parting words from Alphonse Chapanis

It seems appropriate to end this section with some quotations from Chapanis (1988):

> Under the circumstances, we have to fall back on the experience, sophistication, and good judgment of the investigator. Familiarity with the literature and experience with what works and what does not work are invaluable aids to the investigator in deciding how similar is similar enough. (p. 264)
>
> What I have just said may sound as though I think studies will always yield the same results when one builds heterogeneity into them, or when one replicates them. That's not true, of course. In fact, it's more the exception than the rule. Different groups of subjects often do respond differently, different tasks usually produce different kinds of behavior, and variations in the environment often cause people to perform differently. But that's no reason for abandoning heterogeneity as an approach and going back to single-variable, tightly controlled kinds of studies. And, if anything, it's all the more reason for replicating. Knowing that one can't generalize some findings across subjects, tasks, or environments is itself a generalization that is worth having. (p. 266)

[35] Note that this approach likely results in increased variance, which implies a larger sample size for studies whose calculation of measurement precision includes variance. Because this approach increases *p*, though, it implies a smaller sample size requirement for problem discovery studies.

References

Aaker, D. A., and Day, G. S. (1986). *Marketing research*. New York, NY: John Wiley.

Alreck, P. L., and Settle, R. B. (1985). *The survey research handbook*. Homewood, IL: Richard D. Irwin, Inc.

Aykin, N. M., and Aykin, T. (1991). Individual differences in human-computer interaction. *Computers and Industrial Engineering*, *20*, 373-379.

Banks, S. (1965). *Experimentation in marketing*. New York, NY: McGraw-Hill.

Boehm, B. W. (1981). *Software engineering economics*. Englewood Cliffs, NJ: Prentice-Hall.

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

Brown, F. E. (1980). *Marketing research: A structure for decision making*. Reading, MA: Addison-Wesley.

Campbell, D. T., and Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Chicago, IL: Rand McNally.

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

Churchill, Jr., G. A. (1991). *Marketing research: Methodological foundations*. Ft. Worth, TX: Dryden Press.

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

Dickens, J. (1987). The fresh cream cakes market: The use of qualitative research as part of a consumer research programme. In U. Bradley (ed.), *Applied Marketing and Social Research* (pp. 23-68). New York, NY: John Wiley.

Dumas, J., Sorce, J., and Virzi, R. (1995). *Expert reviews: How many experts is enough?* In Proceedings of the Human Factors and Ergonomics Society 39th Annual Meeting (pp. 228-232). Santa Monica, CA: Human Factors and Ergonomics Society.

Fisher, J. (1991). Defining the novice user. *Behaviour and Information Technology, 10,* 437-441.

Fowler, C. J. H., Macaulay, L. A., and Fowler, J. F. (1985). The relationship between cognitive style and dialogue style: an exploratory study. In P. Johnson and S. Cook (eds.), *People and Computers: Designing the Interface* (pp. 186-198). Cambridge, UK: Cambridge University Press.

Fujino, Y. (1980). Approximate binomial confidence limits. *Biometrika*, *67*, 677-681.

Gawron, V. J., Drury, C. G., Czaja, S. J., and Wilkins, D. M. (1989). A taxonomy of independent variables affecting human performance. *International Journal of Man-Machine Studies, 31*, 643-672.

Gordon, W., and Langmaid, R. (1988). Qualitative market research: A practitioner's and buyer's guide. Aldershot, UK: Gower.

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.

Greene, S. L., Gomez, L. M., and Devlin, S. J. (1986). A cognitive analysis of database query production. In *Proceedings of the 30th Annual Meeting of the Human Factors Society* (pp. 9-13). Santa Monica: CA: Human Factors Society.

Harris, R. J. (1985). *A primer of multivariate statistics*. Orlando, FL: Academic Press.

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

Holleran, P. A. (1991). A methodological note on pitfalls in usability testing. *Behaviour and Information Technology, 10*, 345-357.

Jacobsen, N. E., Hertzum, M., and John, B. E. (1998). The evaluator effect in usability studies: Problem detection and severity judgments. In *Proceedings of the Human Factors and Ergonomics Society 42nd Annual Meeting* (pp. 1336-1340). Santa Monica, CA: Human Factors and Ergonomics Society.

Kraemer, H. C., and Thiemann, S. (1987). *How many subjects? Statistical power analysis in research*. Newbury Park, CA: Sage.

LaLomia, M. J., and Sidowski, J. B. (1990). Measurements of computer satisfaction, literacy, and aptitudes: A review. *International Journal of Human-Computer Interaction*, *2*, 231-253.

Landauer, T. K. (1997). Behavioral research methods in human-computer interaction. In M. G. Helander, T. K. Landauer, and P. V. Prabhu (eds.), *Handbook of Human-Computer Interaction* (pp. 203-227). Amsterdam: Elsevier.

Lewis, C., and Norman, D. A. (1986). Designing for error. In D. A. Norman and S. W. Draper (eds.), *User-Centered System Design: New Perspectives on Human-Computer Interaction* (pp. 411-432). Hillsdale, NJ: Erlbaum.

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

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

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

Lewis, J. R. (1993). Multipoint scales: Mean and median differences and observed significance levels. *International Journal of Human-Computer Interaction, 5,* 383-392.

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

Lewis, J. R. (1994b). *Small-sample evaluation of the Paulson-Takeuchi approximation to the exact binomial confidence interval* (Tech. Report 54.872). Boca Raton, FL: International Business Machines Corp.

Lewis, J. R. (1995). IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use. *International Journal of Human-Computer Interaction, 7,* 57-78.

Lewis, J. R. (1996). Binomial confidence intervals for small-sample usability studies. In A. F. Ozok and G. Salvendy (eds.), *Advances in Applied Ergonomics* (pp. 732-737). West Lafayette, IN: USA Publishing.

Lewis, J. R. (1997). *A general plan for conducting human factors studies of competitive speech dictation accuracy and throughput* (Tech. Report 29.2246). Raleigh, NC: International Business Machines Corp.

Lewis, J. R. (1999). Trade-offs in the design of the IBM computer usability satisfaction questionnaires. In H. Bullinger and J. Ziegler (eds.), *Human-Computer Interaction:*

Ergonomics and User Interfaces – Vol. I (pp. 1023-1027). Mahwah, NJ: Lawrence Erlbaum.

Lewis, J. R. (2000a). *Evaluation of Problem Discovery Rate Adjustment Procedures for Sample Sizes from Two to Ten* (Tech Report in press). Raleigh, NC: International Business Machines Corp.

Lewis, J. R. (2000b). *Overestimation of p in problem discovery usability studies: How serious is the problem?* (Tech Report in press). Raleigh, NC: International Business Machines Corp.

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

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

Lewis, J. R. (2000e). *Validation of Monte Carlo estimation of problem discovery likelihood* (Tech. Report in press). Raleigh, NC: International Business Machines Corp.

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

Lord, F. M. (1953). On the statistical treatment of football numbers. *American Psychologist*, 8, 750-751.

Massaro, D. W. (1975). *Experimental psychology and information processing*. Chicago, IL: Rand McNally.

Mayer, R. E. (1997). From novice to expert. In M. G. Helander, T. K. Landauer, and P. V. Prabhu (eds.), *Handbook of Human-Computer Interaction* (pp. 781-795). Amsterdam: Elsevier.

Miller, L. A., Stanney, K. M., and Wooten, W. (1997). Development and evaluation of the Windows Computer Experience Questionnaire (WCEQ). *International Journal of Human-Computer Interaction*, *9*, 201-212.

Myers, J. L. (1979). *Fundamentals of experimental design*. Boston, MA: Allyn and Bacon.

Nielsen, J. (1997). Usability testing. In G. Salvendy (ed.), *Handbook of Human Factors and Ergonomics* (pp. 1543-1568). New York, NY: John Wiley.

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

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

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

Nunnally, J.C. (1978). *Psychometric theory*. New York, NY: McGraw-Hill.

Palmquist, R. A., and Kim, K. S. (2000). Cognitive style and on-line database search experience as predictors of web search performance. *Journal of the American Society for Information Science*, *51*, 558-566.

Parasuraman, A. (1986). Nonprobability sampling methods. In *Marketing Research* (pp. 498-516). Reading, MA: Addison-Wesley.

Smith, B., Caputi, P., Crittenden, N., Jayasuriya, R., and Rawstorne, P. (1999). A review of the construct of computer experience. *Computers in Human Behavior*, *15*, 227-242.

Steele, R. G. D., and Torrie, J. H. (1960). *Principles and procedures of statistics*. New York, NY: McGraw-Hill.

Stevens, S. S. (1951). Mathematics, measurement, and psychophysics. In S. S. Stevens (Ed.), *Handbook of Experimental Psychology* (pp. 1-49). New York: John Wiley.

Sudman, S. (1976). *Applied sampling*. New York: NY: Academic Press.

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.

Virzi, R. A. (1997). Usability inspection methods. In M. G. Helander, T. K. Landauer, and P. V. Prabhu (eds.), *Handbook of Human-Computer Interaction* (pp. 705-715). Amsterdam: Elsevier.

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

Wickens, C. D. (1998). Commonsense statistics. *Ergonomics in Design, 6(4),* 18-22.

Wright, P. C., and Monk, A. F. (1991). A cost-effective evaluation method for use by designers. *International Journal of Man-Machine Studies, 35,* 891-912.

Appendix A. BASIC Program for Approximate Binomial Confidence Intervals

```
10 ' Approximate binomial confidence limits:  Level 2.0, 8/29/90, J. R. Lewis
30 ' This BASIC program uses the Paulson-Takeuchi approximation described in
40 ' Fujino, Y.  (1980).  Approximate binomial confidence limits.
50 ' Biometrika, Volume 67, Page 679, to calculate approximate 2-sided
60 ' 90-, 95-, and 99-percent binomial confidence limits.
70 '
80 Z(1)=1.645      ' Z value for 90-percent confidence limits.
90 Z(2)=1.96       ' Z value for 95-percent confidence limits.
100 Z(3)=2.575     ' Z value for 99-percent confidence limits.
110 CLS
120 PRINT "Approximate 2-sided binomial confidence intervals (90%, 95%, 99%)"
130 PRINT
140 PRINT "Enter the observed number of occurrences (x) and the number of"
150 PRINT "opportunities for occurrence (n, the sample size) separated "
160 PRINT "by a comma.  The results displayed are proportions.  "
170 PRINT
180 PRINT "(Move the decimal place over two positions (i.e., multiply by 100)
190 PRINT "to convert to percentages.)  "
200 PRINT
210 PRINT
220 INPUT "Enter x,n: ",X,N:PRINT:PRINT
230 FOR CNT=1 TO 3
240 LET U=Z(CNT)
250 XPRIME=X:IF XPRIME=N THEN P2=1:IF P2 = 1 THEN GOTO 280
260 GOSUB 470
270 P2=PTPROB
280 ' Get lower limit pprime by replacing x by n-x and
290 ' carry out calculation as before ; then pprime = 1 - ptprob.
300 IF X=0 THEN PPRIME=0:IF PPRIME=0 THEN GOTO 340
310 XPRIME=N-X
320 GOSUB 470
330 PPRIME=1-PTPROB
340 PRINT
350 IF CNT=1 THEN PRINT "90% confidence interval: ";
360 IF CNT=2 THEN PRINT "95% confidence interval: ";
370 IF CNT=3 THEN PRINT "99% confidence interval: ";
380 PRINT USING "#.###";PPRIME;:PRINT " - ";:PRINT USING "#.###";X/N;
390 PRINT " - ";:PRINT USING "#.###";P2
400 NEXT CNT
410 PRINT:PRINT:PRINT "(Use Print Screen if hardcopy is required.)"
420 PRINT:PRINT:INPUT "Do you want to do more calculations? (Y/N)",A$
430 IF LEFT$(A$,1)="y" THEN LET A$="Y"
440 IF LEFT$(A$,1)="Y" THEN 110
450 SYSTEM
460 ' Binomial confidence interval subroutine
470 A=1/(9*(XPRIME+1)):APRM=1-A
480 B=1/(9*(N-XPRIME)):BPRM=1-B
490 F=((APRM*BPRM+U*SQR(APRM^2*B+A*BPRM^2-A*B*U^2))/(BPRM^2-B*U^2))^3
500 PTPROB=(XPRIME+1)*F/((N-XPRIME)+(XPRIME+1)*F)
510 RETURN
```