Behavior Research Methods
2006, 38 (1), 65-76

Optimization of sample size in controlled experiments: The CLAST rule

JUAN BOTELLA, CARMEN XIMÉNEZ, JAVIER REVUELTA, and MANUEL SUERO
Universidad Autónoma de Madrid, Madrid, Spain

Sequential rules are explored in the context of null hypothesis significance testing. Several studies have demonstrated that the fixed-sample stopping rule, in which the sample size used by researchers is determined in advance, is less practical and less efficient than sequential stopping rules. It is proposed that a sequential stopping rule called CLAST (*composite limited adaptive sequential test*) is a superior variant of COAST (*composite open adaptive sequential test*), a sequential rule proposed by Frick (1998). Simulation studies are conducted to test the efficiency of the proposed rule in terms of sample size and power. Two statistical tests are used: the one-tailed *t* test of mean differences with two matched samples, and the chi-square independence test for twofold contingency tables. The results show that the CLAST rule is more efficient than the COAST rule and reflects more realistically the practice of experimental psychology researchers.

In the context of research with controlled experiments, it is considered appropriate to establish the sample size (N) by using the so-called power curves. N is determined as a function of the effect size, the significance level (α), and the desired power of the test (see, e.g., Allison, Silverstein, & Gorman, 1996; Hays, 1988; Kirk, 1995; R. G. O'Brien & Muller, 1993; Winer, 1971). The experiment is run with the sample, the data are analyzed, and a statistical test is applied to make a decision about the null hypothesis (H_0). This procedure, which, following Frick (1998), we will call the *fixed-sample rule* (FSR), has several shortcomings, the principal one being that it is inefficient because it does not stop early enough when statistical significance is either nearly assured or unlikely. For this reason, researchers frequently fail to use it and sometimes violate it when they do.

There is another type of sampling rule, called the *sequential sampling rule* (SSR), in which the number of subjects is not fixed in advance. The SSR allows researchers to test their hypotheses as the data are accumulated so that decisions can be made as the study progresses. Special SSRs have been proposed to allow multiple testing while the Type I error rate is maintained at a reasonable level under H_0. Despite their early introduction by Kimball (1950) and Fiske and Jones (1954), SSRs have received little attention in the field of psychology (see, e.g., Vos, 2001). Mathematical sophistication is one reason for their limited use. For these reasons, Frick (1998) proposed the COAST (*composite open adaptive sequential test*) SSR, which is simple and easy to implement and which will be used as a test reference here. However, outside of the behavioral sciences, SSRs have been widely used in clinical trials research (Lachin, 1981; P. C. O'Brien & Fleming, 1979).

In psychology, the usual goal of experimental research is to determine the existence of an effect (typically defined in terms of a causal relation between two variables). Statistical testing allows decision-making as to whether or not H_0 should be rejected. Researchers consider this process to be more efficient, since fewer resources are invested to reach a conclusion. One procedure is more efficient than another when it achieves the same result (i.e., when it has the same power) with fewer resources.

Following the work of Frick (1998), in the present article we analyze sequential sampling in terms of efficiency. We propose an alternative SSR to both the FSR and Frick's SSR. Our proposed rule is also sequential, but it reflects more realistically the practice of researchers in experimental psychology.

The Fixed-Sample Rule

The advantages and disadvantages of FSR are due to its most important characteristic: its strictness. Among its advantages are its clarity, precision, and fairness when used properly. The reader of a research paper has precise information about what the researcher has done and is normally safe in assuming that, given the procedures used, the probability of Type I error is equal to the nominal significance level, α. The main disadvantage of FSR has to do with the inefficiency of the procedure. A rule is more efficient when it achieves the same power with fewer observations while maintaining the Type I error rate at a comparable level. If N is determined in advance and the decision is

This research was partially supported by Grants BSO2003-08908 from the Ministerio de Educación y Ciencia of Spain and 06/HSE/0005/2004 from the Comunidad de Madrid, Spain. We thank two anonymous reviewers for their insightful comments and suggestions. Thanks are also due Jim Juola for his comments on an earlier version of the manuscript. Correspondence concerning this article should be addressed to J. Botella, Facultad de Psicología, Departamento de Psicología Social y Metodología, Universidad Autónoma de Madrid, Cantoblanco s/n, 28049 Madrid, Spain (e-mail: juan.botella@uam.es).

made with the same Type I error rate and power as in another procedure in which a smaller N is used, then the former procedure is less efficient than the latter.

The strictness of the FSR maintains that if the p value is not less than the α value (e.g., .05), H_0 will not be rejected. If the p value is small but statistically nonsignificant (e.g., .06), the researcher might decide that using a larger N could change the decision about H_0. Moreover, a nonnull effect is usually the most desirable result. When researchers obtain p values that are larger than α but still small (e.g., $.05 < p < .15$), they might describe the results as "marginally significant" or use other descriptions that convey uncertainty. In such a case, strict application of the FSR leads to the conclusion that there is no effect, but the implication is that collecting a few more observations than planned would have led to a p value of less than .05 (the reader should notice the implicit conclusion that there is an effect, even though the formal conclusion states the opposite).

On some occasions, the opposite situation occurs: The p value is so high (e.g., .80) that the researcher thinks the same decision (i.e., not to reject H_0) could have been reached with a much smaller N. In this situation, the researcher might have used more observations than needed to reach the conclusion. The low efficiency of the FSR is recognized by many researchers. We suspect that sometimes this rule is not used in practice.

Sequential Rules

The main alternative to the FSR—SSRs—was first proposed by Wald (1947). He developed SSRs in which additional sampling is conditionalized on an analysis of the previous data. That is, a hypothesis test conducted over the collected data is the basis for deciding whether (1) more observations should be added to the sample or (2) the experiment should be ended and a final decision made about H_0. General overviews of SSRs in statistical contexts can be found in Siegmund (1985, 1994), Wetherill and Glazebrook (1986), Ghosh and Sen (1991), and Lai (2001).

SSRs differ in the number of observations that are incorporated into the sample after each analysis. In the simplest scenario, data are analyzed after each observation is incorporated into the sample. Other SSRs assume that data are incorporated in groups, and the analyses are made over certain proportions of the final sample (see, e.g., Lan & DeMets, 1989; P. C. O'Brien & Fleming, 1979). These rules, generally referred to as *group sequential methods*, have rarely been used in psychology (Jennison & Turnbull, 2000). The expected number of observations in the final sample decreases as the number of observations added in each step of the sequence decreases. For this reason, sequential methods based on adding single observations should be more efficient than group sequential methods, since finer decisions can be made with the former type of methods.

In the remainder of this article, we use the term *N2* to denote the number of observations incorporated into the sample after each analysis. Here, we assume that $N2 = 1$ unless otherwise noted. However, in many applied contexts, group SSRs might be more appropriate, and we defer them for future research.

We suspect that some researchers begin by setting a minimum sample size required for an experiment. If the p value is significant ($p \leq \alpha$), they stop the process and reject H_0. If the value of p is nonsignificant and large, they end the experiment and conclude that there is no effect. However, if the p value is small but not statistically significant, they continue to incorporate observations into the sample. As Frick (1998) noted, it appears that personal criteria determine whether or not an experiment is a pilot study. If the pilot data are discouraging, the study will not be run, whereas if the results are significant the pilot study becomes an experiment of record. As we will see, a consequence of this procedure is that, below a nominal value of .05, the true probability of a Type I error can typically amount to more than twice that value.

The most important problem with this practice is that, since researchers do not follow the FSR, it is difficult to know specifically which stopping rules they use. These rules are also likely to be subjective and idiosyncratic, and researchers ignore any effects they might have on the interpretation of the results. But if these stopping rules are known, their properties could be studied in terms of efficiency and to guarantee that the Type I error rate is maintained at a reasonable level under H_0.

The present work focuses on a very specific scenario, which was explored by Frick (1998). It has the following characteristics: (1) The goal is to use some hypothesis test to decide if there is an effect, (2) it is easy to add subjects to the sample, and (3) the researcher seeks to work with a specific probability of Type I error that does not exceed a certain value (α) but that also needs to achieve a certain level of power ($1 - \beta$).

Characteristic 1 assumes that the goal is to make a dichotomous decision about whether or not there is an effect. If there is also a secondary goal (e.g., to determine whether or not the effect size exceeds some criterion), the procedure would be different. Specifically, the researcher can establish a minimum effect size to be considered as reaching "practical significance" (something other than statistical significance) and, assuming a specific level of power, calculate the proper sample size. Of course, the procedure would also be different if the goal were to achieve an accurate estimate of the effect size.

Characteristic 2 indicates that there are no limitations of availability for sampling new observations. In some situations, the sample has some characteristics that make it difficult to incorporate new subjects. For example, if the individuals in the sample have received long-term clinical interventions, then sample size is necessarily determined in advance.

Characteristic 3 is the most usual setting for researchers. However, it must be assumed that this procedure maintains the Type I error rate at a reasonable level while retaining adequate power of the test.

The consequences of adopting an SSR are analyzed below. Let us suppose that a researcher proceeds as is

$$P(S_{N1}) + P(S_{N1+N2}, U_{N1}) + P(S_{N1+2\cdot N2}, U_{N1+N2}) = \alpha_1 + P(S_{N1+N2}|U_{N1})P(U_{N1}) + P(S_{N1+2\cdot N2}|U_{N1+N2}, U_{N1})P(U_{N1+N2}|U_{N1})P(U_{N1}) = \alpha_1 + P(U_{N1})P(S_{N1+N2} \cup S_{N1+2\cdot N2}|U_{N1}).$$

shown in the upper panel of Figure 1 (here, we refer to a one-tailed right-sided statistical test, but the procedure can be generalized to left-sided tests).

The SSR can be specified by the following steps:

1. Set the minimum sample size needed ($N1$), run the experiment with that sample size, and analyze the data, classifying the result into one of the following three categories as a function of the p value and the α values that limit the uncertainty region (α_l and α_u):

Significant (S) if $p \leq \alpha_l$;

Uncertain (U) if $\alpha_l < p \leq \alpha_u$;

Nonsignificant (NS) if $p > \alpha_u$.

2. Proceed as follows on the basis of the obtained p value:

If S, stop the experiment and reject H_0.

If U, incorporate $N2$ subjects into the sample and reanalyze the data.

If NS, stop the experiment without rejecting H_0.

3. Classify the result with $N1 + N2$ subjects into one of the three categories listed in Step 1 and then proceed again with Step 2. The same process is repeatedly applied until the value of the test statistic falls outside the U region or a cumulative value of N exceeds a certain maximum sample size (N_{max}).

Although this process can be repeated infinitely, in practice researchers are not willing to test an infinite number of subjects (Whitehead & Brunier, 1990). The real sample size is usually that which the researcher estimates is the largest that does not have an excessively high power value and does not detect nonrelevant effects. This sample size is denoted here by N_{max}.

Let us consider the consequences of applying this rule for the actual probability of a Type I error—that is, assuming that H_0 is true. The probability of a Type I error with the $N1$ or the $N1 + N2$ sample is

$$P(S_{N1}) + P(S_{N1+N2}, U_{N1}) = \alpha_1 + P(S_{N1+N2}|U_{N1})P(U_{N1}),$$

where $P(S_{N1}) = \alpha_1$ is the probability of a Type I error for the initial sample (that with $N1$ observations).

If the test statistics of the $N1 + N2$ sample fall into the uncertainty region (U), then additional $N2$ observations should be incorporated into the sample and the data must be reanalyzed. The probability of rejecting H_0 when it is true in one of the three tests is given by the equation at the top of the page.

Suppose that the sequential test is repeated until the sample size reaches N_{max}. Then, the number of samples taken after $N1$ is $J = (N_{max} - N1)/N2$. Since $P(U_{N1}) = \alpha_u - \alpha_l$, the probability of making a Type I error in one of the $J + 1$ samples can be written as in Equation 1 at the bottom of the page, where

$$P(S_{N1+N2} \cup \cdots \cup S_{N1+J\cdot N2}|U_{N1})$$

gives the probability of rejecting H_0 conditional on an uncertain result on the initial sample. Therefore, the probability of Type I error must be $\alpha_l \leq P(\text{Type I error}) \leq \alpha_u$. However, there are no analytical solutions for setting the values of α_l and α_u that lead to a prespecified probability for the Type I error. In this article, those values will be found using Monte Carlo simulations.

As an illustration of the quantitative effect of repeated testing on Type I error rate, we introduce the result of a simulation for a one-tailed t test of two matched samples with an α value of .05 and a maximum p value of .36 ($\alpha_l = .05$; $\alpha_u = .36$). The initial sample size ($N1$) was 15. With H_0 true, and assuming as a secondary stopping criterion a sample size of $N_{max} = 45$, we obtained a proportion of rejections of H_0 of .132—that is, more than twice the proportion expected on the basis of the nominal significance level, α.[1] However, this does not imply an unsolvable problem. To solve this problem, it suffices to know how much the probability of a Type I error increases, then control it and take it into account in the conclusions. For instance, to avoid an excessive increase in the probability of a Type I error, it is worthwhile to set a conservative α value for the sequential tests (in the above example, a value smaller than the desired .05 significance level).

The maximum value of p for stopping the process should not be too high. For example, assume that a re-

$$P(\text{Type I error}) = \alpha_l + \sum_{i=1}^{J} P(S_{N1+i\cdot N2}, U_{N1+(i-1)\cdot N2}) = \alpha_l + (\alpha_u - \alpha_l)P(S_{N1+N2} \cup \cdots \cup S_{N1+J\cdot N2}|U_{N1}), \quad (1)$$

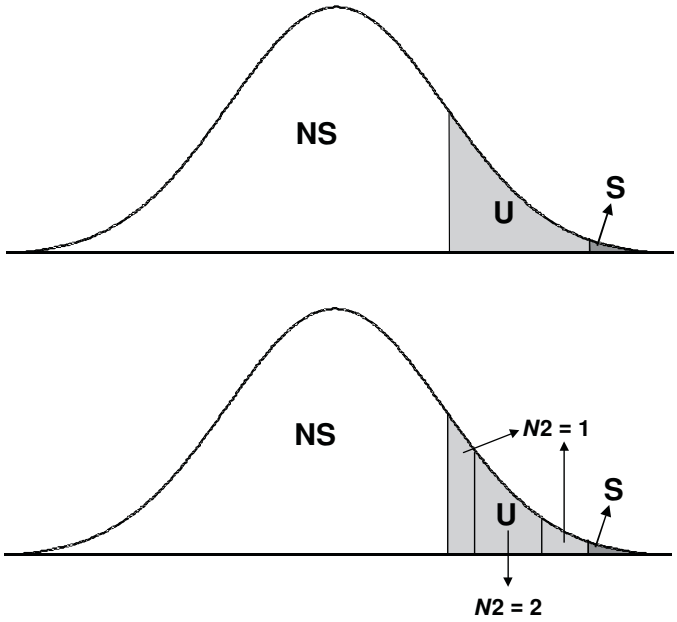

Figure 1. Graphic representation of the CLAST (upper curve) and CLAST-T (lower curve) rules. In both curves, the three decisions are designated as NS (nonsignificant), S (significant), and U (uncertain) (see text).

searcher decides to apply the sequential procedure with a minimal sample size of 10 subjects and incorporates subjects, one by one, as a function of the results of the data analyses. The process will be stopped when the p value is either less than or equal to .05 or greater than .36, or when sample size reaches 25 subjects. An assumption about the conditional probabilities must also be made. It is obvious that the probability that the p value falls in the U region is larger than it was in the previous tests. The reason is that subjects involved in previous tests are also included in the present test. Thus, these tests are not independent, and their correlations will increase the α value. The computation of the probability of a Type I error is a difficult task. But assume for a moment that the tests were independent, even though this would lead to an underestimation of the Type I error rate. In this case, Equation 1 would give a real probability of Type I error of .072—that is, on the incorrect and conservative assumption that the tests are independent, the Type I error rate would increase by 44%.

The effects of the SSR on the power of the test are more difficult to compute because $(1 - \beta)$ depends on the sample size defined in each test, which is different in each sequential analysis. The best way to deal with this problem is to estimate β from the empirical rejection rate when H_0 is false. The value of α_l adopted for the first test determines the value of β in the same test. α_l will be small in comparison with the β value, which results from the FSR. This constitutes an advantage relative to the power because it allows an increase without exceeding the FSR value. In fact, in each test there is an opportunity to obtain a p value smaller than α_l or greater than α_u. We need to establish which values of α_l and α_u produce a failure-to-reject rate closer to the β value that results from the FSR. The solution is to set a conservative α_l value. This decreases the probability of rejecting H_0 in each test, but if the U region is large enough, the test statistic may fall in the U region and the incorporation of new observations may lead to rejection of H_0.

Another important task is to decide the reference criterion used to assess efficiency. It is necessary to estimate the expected value of the number of subjects [E(N)] for the rule. E(N) is estimated from the mean of observations in the 10,000 samples of each simulation. But which N is used for comparison with the estimated N? Frick (1998) used the mean sample size obtained in the simulations with the SSR (N_{SEQ}). That is, he obtained N_{SEQ} and the empirical proportion of rejections (EPR) of H_0 as an estimate of the power of the SSR. Then, the N associated to the FSR (i.e., N_{FSR} with power equivalent to EPR) is computed and the efficiency is assessed in terms of the relationship between the two sample sizes. More specifically, he computed the percentage of savings, in terms of number of subjects, with a given SSR—that is,

$$\frac{N_{FSR} - N_{SEQ}}{N_{FSR}} \times 100. \qquad (2)$$

Here, we propose an alternative procedure that we believe is a better representation of what researchers do in practice. As we mentioned earlier, a simple solution to

some of the problems reviewed here consists of adopting a stricter decision criterion for rejecting H_0 than the one adopted with the FSR. Assume, for example, that we seek an alternative to the FSR with an α value of .05. As seen above, the Type I error rate with the SSR increases in comparison with the significance level applied in each subsequent test. Setting a more restrictive criterion in each sequential test, the EPR could reach values closer to the α value of the FSR criterion. Frick (1998) has shown that when a .01 value is adopted as the lower bound in each test, the final Type I error rate is stabilized at around .05. Furthermore, the .01 value has the advantage of being one of the most widely used in practice.

The COAST Sequential Rule

Since COAST is the best known SSR in the context of psychology, we will analyze it in some detail. It was proposed by Frick (1998). The rule is clear and easy to apply, but some of its specifications appear arbitrary. COAST consists of the process described above, setting $\alpha_l = .01$ and $\alpha_u = .36$. However, COAST specifies no value for $N1$ or for N_{max}. In a Monte Carlo simulation study, Frick applied COAST to a two-tailed *t* test of means with matched samples. The results indicated that (1) the proportion of rejections is very close to .05 and (2) COAST requires about 30% fewer subjects than does FSR when the sample size for FSR is applied to achieve the same power.

Our Proposal: The CLAST Sequential Rule

The present investigation is a continuation of Frick's (1998) method and is based on the following ideas.

1. We agree with Frick (1998) in arbitrarily setting the value of α_l at .01. It is easy to remember and, together with .05, is one of the most popular values used for statistical tests in psychology.

2. However, there is no clear reason for setting α_u at .36. This figure could be based on empirical findings that are not fully explained by Frick (1998). All of his tests were two-tailed. In practice, directional hypotheses are probably as frequently used as nondirectional hypotheses, and it is not obvious that Frick's conclusions will also apply to one-tailed tests.

3. The only stopping criterion for COAST is based on the *p* value, which must fall outside the U region to stop the process. This is not realistic because in practice there is always an N_{max}. If the *p* value is in the U region after N_{max} is reached, the researcher makes a decision with the available information.

We have run some simulations with the COAST rule using 10,000 replications. In some replications, the sequential procedure required 40 individuals in order for a conclusion to be reached. However, in some cases COAST required 400 subjects! Frick (1998) does not report on how he handled such situations, but it is clear that, in some cases, sample size may be too large to be achieved in practice.

Arnold and Harvey (1998) introduced the *data monitoring* approach to experimental design. Data monitoring is an SSR that assumes a fixed value for N_{max}. However, data monitoring differs from COAST in that no α_u value is used and H_0 is held only if sample size reaches N_{max} without being rejected. Arnold and Harvey conducted a simulation to determine the significance level that should be set for each hypothesis test in order to obtain an overall significance level for the complete sequence of tests. One problem of data monitoring is that the lack of α_u produces a loss in efficiency in terms of number of subjects. Furthermore, it is too complex to be used in practice.

Our proposal combines the COAST rule with data monitoring. We assume a value for α_u and a fixed value for N_{max}. The resulting SSR is no longer open, and we refer to it as *CLAST* (*composite limited adaptive sequential test*).

4. Frick (1998) does not propose any way to set $N1$ in the COAST rule. In the CLAST rule, we propose that both $N1$ and N_{max} be set as a function of N_{FSR}.

5. We wonder whether complex SSRs improve performance of simple sequential tests such as COAST and CLAST. Specifically, the U region can be split into several regions (see the bottom panel of Figure 1). It is unlikely that the *p* value will fall outside of U after a new subject is incorporated if it is a central value within U. A multiple SSR may establish that more than one observation must be incorporated simultaneously on those occasions. In contrast, when the *p* value lies close to the boundaries of U, subjects could be added one by one. The goal of the multiple SSR is to reduce the number of tests. In the following, we will refer to it as *CLAST-T* (the final T standing for "triangular").

6. The performance of the SSR may be affected by the test statistic. Specifically, it is likely that performance is impaired when the test statistic has an asymmetric distribution. We evaluate this hypothesis by using a chi-square independence test, which, to our knowledge, has not been applied before in conjunction with the COAST rule or its variants.

The CLAST rule could be specified in the following steps (the CLAST-T rule differs from it only in Step 3):

1. Set the sample size needed for the experiment according to the FSR. This size is determined by the classical procedure and is denoted here by N_{FSR}. $N1$ and N_{max} are computed as $\pm 50\%$ of N_{FSR}, $N1$ being $0.5 \cdot N_{FSR}$ and N_{max} being $1.5 \cdot N_{FSR}$).

2. Run the experiment with the $N1$ sample size and analyze the data, classifying each of them into one of the following three categories as a function of the *p* value and the α values that limit the uncertainty region (α_l and α_u): S if $p \leq \alpha_l$; U if $\alpha_l < p \leq \alpha_u$; NS if $p > \alpha_u$.

3. For the CLAST rule, proceed as follows on the basis of the obtained *p* value:

If S, stop the experiment and reject H_0.

If U, incorporate $N2 = 1$ subject into the sample and reanalyze the data.

If NS, stop the experiment without rejecting H_0.

For the CLAST-T rule, the procedure is the same as for CLAST, but the central region of U is identified by setting

two values, α_{cl} and α_{cu}, such that $\alpha_l < \alpha_{cl} < \alpha_{cu} < \alpha_u$. Then, the value of $N2$ is obtained as follows:

If $\alpha_l < p \leq \alpha_{cl}$, then $N2 = 1$.

If $\alpha_{cl} < p \leq \alpha_{cu}$, then $N2 = 2$.

If $\alpha_{cu} < p \leq \alpha_u$, then $N2 = 1$.

4. The result with the $N1 + N2$ sample is again classified in one of the three categories of Step 2, and the process is reiterated.

5. If N_{max} is achieved in Step 4, then the following rule of classification is applied to end the process:

S if $p \leq \alpha_i$,

NS if $p > \alpha_i$,

where α_i is the intended value for the Type I error rate in the whole sequential test.

SIMULATION STUDY

We conducted two simulation studies to evaluate the efficiency of CLAST and CLAST-T rules and compare it to that of COAST. The first study is a pilot simulation intended to determine the optimal values of α_u for these rules. In the second simulation, we keep α_u constant to the value obtained in the pilot, and the rules are compared using different sample and effect sizes. Both simulations are conducted separately for the *t* test and the chi-square test.

Method

The study was conducted using Monte Carlo simulations. The programs were written in R language (Venables & Smith, 2001), a free distribution version of the well-known S language (Becker, Chambers, & Wilks, 1988). The simulation consists of the steps described above.

Test statistics. The simulation was run with two different test statistics: the one-tailed *t* test of mean differences with two matched samples and the chi-square test of independence for twofold contingency tables. The *t* test is based on the *D* variable (see Appendix A), which is distributed $N(\mu, 1)$. The hypotheses for the one-tailed test are H_0: $\mu_D = 0$ and H_1: $\mu_D > 0$. The test statistic (T) follows a Student's *t* distribution with $N - 1$ degrees of freedom. Effect size is defined as the standardized mean difference (see Appendix A). The computation of the effect size for the chi-square statistic is described in Appendix B.

SSRs. The SSRs differ in several respects. The COAST rule (Frick, 1998) is based solely on the *p* value, and $N2$ is fixed at 1. The CLAST rule is a modification of COAST that requires computation of the value of the *N* that would have been used with the FSR (referred to as N_{FSR}). This value is used to set $N1$ and N_{max}. Moreover, CLAST has two stopping criteria: one based on the *p* value and the other based on the determination of N_{max}. As we have mentioned, the rule is called *CLAST* since it is not completely open because of the maximum sample size. The CLAST-T rule is similar to CLAST, but the uncertainty region is divided into three regions, which determine the number of observations ($N2$) that must be incorporated into the sample (one observation if the statistic falls into the regions that are closer to S and NS, and two observations if it falls in the central part). The rule is called *CLAST-T* because $N2$ is computed using a triangular function.

Parameters of the simulations. Several aspects of the simulations are common to all conditions. First, the values of $N1$ and N_{max} depend on N_{FSR}. Specifically, $N1 = N_{FSR}/2$ and $N_{max} = N_{FSR} + N_{FSR}/2$. Second, $N2 = 1$ for all rules except for CLAST-T, which computes $N2$ as mentioned earlier. Third, the values of N_{FSR} have been set from the range most frequently employed by experimental psychologists. Regarding the *t* test, N_{FSR} equals 16, 20, 24, 30, and 40, and the effect sizes are 0, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90, and 1.00 (δ values; see Appendix A). Of course, H_0 is true only when δ equals 0. Regarding the chi-square test, the values of N_{FSR} are 30, 40, 50, 60, 70, and 80, and the effect sizes are 0.10, 0.30, and 0.50 (*w* values; see Appendix B). We have used different values than in the *t* test because (1) $N1$ values smaller than 30 frequently generate empty cells and (2) the effect size measure uses a different scale than the *t* test.

Following the results of the pilot simulation, we set $\alpha_u = .25$ for the *t* test and $\alpha_u = .35$ for the chi-square test for the second simulations. Recall that these values were selected to determine the value of α_u that yields a cumulative Type I error rate as close as possible to .05. The results of the pilot simulations are presented below for each statistical test. The values of α_{cl} and α_{cu} for CLAST-T were set at .1 and .2. The value of α_i is .05. The simulation was repeated 10,000 times for each condition.

Data analysis. The data gathered from each condition are the mean and *SD* of the sample size (N) and the EPR of H_0, which, under H_0 false, is an estimate of the power of the SSR. The efficiency of the procedure is computed from the sample size using Equation 2 and from the power using Equation 3 below. The efficiency of the power is obtained by computing power under the FSR (which assumes a sample size of N_{FSR}). The power of the SSR (EPR) is compared with the power of the FSR using the equation

$$\frac{\text{Power}\left(N_{FSR}\right) - \text{EPR}}{\text{Power}\left(N_{FSR}\right)} \times 100. \quad (3)$$

If the efficiency is negative, the power of the SSR is higher than that of the corresponding FSR, and the reverse occurs if the efficiency is positive.

Table 1
Results of the Simulations for H_0 True ($\delta = 0$) and Several Upper Bounds (α_u) of CLAST in the One-Tailed Test of Mean Differences With Two Matched Samples

				N			
α_u	N_{FSR}	$N1$	N_{max}	*M*	*SD*	EPR	*N* Efficiency
.20	16	8	24	9.27	3.52	.047	42.06
	20	10	30	11.39	4.11	.042	43.05
	24	12	36	13.61	4.78	.046	43.29
	30	15	45	17.03	5.99	.046	43.23
	40	20	60	22.47	7.66	.044	43.83
.25	16	8	24	9.68	4.04	.051	39.69
	20	10	30	12.06	5.00	.048	39.40
	24	12	36	14.35	5.91	.049	40.17
	30	15	45	17.86	7.29	.049	40.63
	40	20	60	23.54	9.38	.050	40.88
.30	16	8	24	10.23	4.70	.055	36.44
	20	10	30	12.71	5.83	.053	36.30
	24	12	36	15.07	6.80	.053	37.25
	30	15	45	18.62	8.18	.052	36.97
	40	20	60	24.88	11.06	.052	38.00
.35	16	8	24	10.82	5.25	.060	32.25
	20	10	30	13.38	6.45	.056	32.85
	24	12	36	16.02	7.71	.058	33.29
	30	15	45	19.97	9.63	.056	33.83
	40	20	60	26.23	12.51	.054	34.15

Note—N_{FSR}, sample size with the fixed-sample rule; EPR, empirical proportion of rejections of H_0 (here it is an estimation of Type I error rate); *N* Efficiency, percentage of savings in number of observations following Equation 2.

Results

Results are presented separately for each statistical test. Our primary goal was to keep the Type I error rate as close as possible to the intended significance level, and our secondary goal was to maximize efficiency. In the set of simulations of the CLAST-T rule, we did not find systematic changes in comparison with that of the CLAST rule. The efficiency values are very similar to those of CLAST in terms of both number of observations and power, and the changes do not fluctuate except for random error. Thus, we will not describe the results related to the CLAST-T rule.

Table 2
Results of Simulations of CLAST With $\alpha_u = .25$ and Under Different Effect Sizes in the One-Tailed Test of Mean Differences With Two Matched Samples

				Power	*N*				
δ	N_{FSR}	$N1$	N_{max}	of N_{FSR}	*M*	*SD*	EPR	*N* Efficiency	P Efficiency
0.10	16	8	24	.103	10.58	5.02	.105	33.88	−1.94
	20	10	30	.112	13.22	6.22	.112	33.90	0.00
	24	12	36	.121	15.88	7.48	.118	33.83	2.48
	30	15	45	.134	19.86	9.32	.131	33.80	2.24
	40	20	60	.153	26.82	12.64	.153	32.95	0.00
0.20	16	8	24	.189	11.68	5.75	.205	27.00	−8.47
	20	10	30	.217	14.58	7.23	.226	27.10	−4.15
	24	12	36	.244	17.66	8.60	.246	26.42	−0.82
	30	15	45	.283	22.21	10.84	.289	25.97	−2.12
	40	20	60	.344	29.79	14.22	.350	25.53	−1.74
0.30	16	8	24	.309	12.65	6.16	.335	20.94	−8.41
	20	10	30	.363	15.87	7.60	.379	20.65	−4.41
	24	12	36	.414	18.83	8.92	.427	21.54	−3.14
	30	15	45	.484	23.42	10.92	.489	21.93	−1.03
	40	20	60	.587	30.71	13.90	.583	23.23	0.68
0.40	16	8	24	.453	13.03	6.12	.468	18.56	−3.31
	20	10	30	.532	16.06	7.37	.544	19.70	−2.26
	24	12	36	.601	19.09	8.60	.619	20.46	−3.00
	30	15	45	.690	23.08	9.96	.689	23.07	0.14
	40	20	60	.800	28.85	11.87	.797	27.87	0.38
0.50	16	8	24	.604	13.28	5.90	.627	17.00	−3.81
	20	10	30	.695	15.96	6.91	.712	20.20	−2.45
	24	12	36	.768	18.45	7.76	.775	23.13	−0.91
	30	15	45	.848	21.64	8.68	.836	27.87	1.42
	40	20	60	.928	26.14	9.39	.908	34.65	2.16
0.60	16	8	24	.740	12.96	5.53	.758	19.00	−2.43
	20	10	30	.827	15.18	6.22	.825	24.10	0.24
	24	12	36	.886	17.18	6.73	.881	28.42	0.56
	30	15	45	.941	19.57	6.81	.926	34.77	1.59
	40	20	60	.981	23.65	6.78	.968	40.88	1.33
0.70	16	8	24	.847	12.34	4.97	.852	22.88	−0.59
	20	10	30	.915	14.02	5.24	.909	29.90	0.66
	24	12	36	.954	15.72	5.50	.938	34.50	1.68
	30	15	45	.982	17.91	5.08	.968	40.30	1.43
	40	20	60	.997	21.79	4.38	.991	45.53	0.60
0.80	16	8	24	.920	11.59	4.35	.917	27.56	0.33
	20	10	30	.964	13.07	4.42	.955	34.65	0.93
	24	12	36	.984	14.46	4.18	.975	39.75	0.91
	30	15	45	.996	16.73	3.71	.988	44.23	0.80
	40	20	60	.999	20.83	2.78	.998	47.93	0.10
0.90	16	8	24	.963	10.74	3.71	.956	32.88	0.73
	20	10	30	.987	12.12	3.46	.976	39.40	1.11
	24	12	36	.996	13.52	3.09	.989	43.67	0.70
	30	15	45	.999	15.89	2.49	.997	47.03	0.20
	40	20	60	1.000	20.35	1.66	.999	49.13	0.10
1.00	16	8	24	.985	10.05	3.05	.977	37.19	0.81
	20	10	30	.996	11.43	2.70	.990	42.85	0.60
	24	12	36	.999	12.92	2.28	.997	46.17	0.20
	30	15	45	1.000	15.47	1.67	.999	48.43	0.10
	40	20	60	1.000	20.10	0.71	1.000	49.75	0.00

Note—N_{FSR}, sample size with the fixed-sample rule; EPR, empirical proportion of rejections of H_0 (here it is an estimation of the correct rejections rate); *N* Efficiency, percentage of savings in number of subjects, following Equation 2; P Efficiency, percentage of decrease in power of N_{FSR}, following Equation 3.

Test of mean differences with two matched samples. To analyze the data, we sought the upper bound of the uncertainty region (α_u) that held the Type I error rate as close as possible to .05. Table 1 shows the results of the pilot simulations (which assume $\delta = 0$). The table has four sections, each showing the results for a different α_u value. The EPR (empirical proportion of rejections of H_0) column represents estimates of the Type I error rate. As can be seen, the upper bound that produces cumulative rates closest to .05 is .25. This upper bound is different from the one used by Frick (1998). That is, we have modified Frick's procedure by using one-tailed tests and setting a maximum value for N as an additional stopping criterion. Thus, the maximum p value at which subjects can continue to be incorporated while the Type I error rate is held at a .05 level is .25, instead of the .36 proposed by Frick.

In comparison with the efficiency achieved with the FSR rule, that achieved with this procedure reduces the number of subjects by about 40%. Frick (1998) achieved a 30% general efficiency with the COAST rule, but, as was noted above, he computed efficiency in a different way. In order to compare the two results, we have calculated the efficiency of Frick's procedure. The efficiency of the CLAST rule is lower when calculated in this way, but it is still higher than that of the COAST rule.

Table 2 includes some simulations similar to those of Table 1 but for the case in which H_0 is false and the δ values are different. Results indicate that the EPR (here an estimate of power) is very close to the power of the corresponding FSR. This result is very similar to that found by Frick (1998). With small effect sizes (e.g., $\delta = 0.20$), the efficiency is about 27% and does not change as a function of N_{FSR}. With medium effect sizes (e.g., $\delta = 0.50$), the efficiency is a function of N_{FSR}. With a small N (e.g., $N_{FSR} = 16$), the efficiency is only 17%, but when N is medium (e.g., $N_{FSR} = 40$) the efficiency is about 35%. With a large effect size (e.g., $\delta = 0.80$), the efficiency is about 39% and is a function of N_{FSR}.

The results indicate that the power of EPR is very similar to that of N_{FSR}—that is, the procedure allows a saving of subjects without losing the power associated with N (see the right column of Table 2).

Chi-square independence test for twofold contingency tables. In this case, we also ran a pilot simulation to determine the value of α_u that yields a cumulative Type I error rate as close as possible to .05. Table 3 shows that the most appropriate value is not .25 but .35.

On the other hand, in comparison with the efficiency of FSR, the efficiency achieved with this procedure yields about a 39% savings in number of subjects. We do not have a reference criterion with which to compare this efficiency level because Frick (1998) did not provide results for this statistical test.

Table 4 includes the results of similar simulation studies, but with H_0 false and different w values. Again, the EPR of H_0 is very similar to the power found if the FSR is applied with a sample size of N_{FSR}. This result is very similar to the general result found by Frick (1998). With small effect sizes (e.g., $w = .10$), the efficiency is about 35% and does not change as a function of N_{FSR}. However, with a medium effect size (e.g., $w = .30$), the efficiency is about 28%; with the largest effect size ($w = .50$) it is about 40%; and in both cases it increases as a function of N_{FSR}. As can be seen, here the relation between effect size and efficiency does not increase monotonically but is U-shaped. The right column of Table 4 shows that the values of efficiency in terms of power are generally close to zero. Thus, the CLAST rule provides a level of power similar to that of the FSR, with a saving in number of subjects.

Table 3
Results of the Simulations for H_0 True ($w = 0$) and Several Upper Bounds (α_u) of CLAST in the Chi-Square Independence Test for Twofold Contingency Tables

				N			
α_u	N_{FSR}	$N1$	N_{max}	*M*	*SD*	EPR	*N* Efficiency
.25	30	15	45	17.18	5.85	.041	42.73
	40	20	60	22.65	7.44	.041	43.38
	50	25	75	28.26	9.14	.044	43.48
	60	30	90	33.57	10.47	.045	44.05
	70	35	105	39.27	12.35	.047	43.90
	80	40	120	44.71	14.03	.038	44.11
.30	30	15	45	17.91	6.86	.042	40.30
	40	20	60	23.53	8.69	.051	41.18
	50	25	75	29.14	10.65	.053	41.72
	60	30	90	35.25	13.06	.048	41.25
	70	35	105	40.86	14.94	.048	41.63
	80	40	120	46.31	16.55	.046	42.11
.35	30	15	45	18.69	7.76	.045	37.70
	40	20	60	24.55	10.00	.051	38.63
	50	25	75	30.84	12.51	.056	38.32
	60	30	90	36.51	14.68	.050	39.15
	70	35	105	42.35	16.95	.048	39.50
	80	40	120	48.66	19.53	.049	39.18
.40	30	15	45	19.52	8.55	.051	34.93
	40	20	60	25.70	11.06	.054	35.75
	50	25	75	31.68	13.50	.054	36.64
	60	30	90	37.76	16.08	.058	37.07
	70	35	105	44.62	19.50	.052	36.26
	80	40	120	50.14	21.16	.056	37.32
.45	30	15	45	20.49	9.36	.051	31.70
	40	20	60	27.04	12.33	.054	32.40
	50	25	75	33.12	14.84	.054	33.76
	60	30	90	39.26	17.60	.058	34.57
	70	35	105	46.11	20.77	.052	34.13
	80	40	120	52.07	23.01	.056	34.91

Note—N_{FSR}, sample size with the fixed-sample rule; EPR, empirical proportion of rejections of H_0 (here it is an estimation of Type I error rate); *N* Efficiency, percentage of savings in number of subjects, following Equation 2.

Discussion

In the present study, a variant of Frick's (1998) COAST rule is proposed. This alternative SSR, named CLAST, sets as a secondary stopping criterion a maximum sample size, which is determined by the researcher in advance. That is, the rule determines not only an initial sample size but also a maximum sample size. Both sizes are computed on the basis of the N assumed by the FSR (denoted here as

Table 4
Results of Simulations of CLAST With $\alpha_u = .35$ and Under Different Effect Sizes in the Chi-Square Independence Test for Twofold Contingency Tables

				Power	*N*				
w	N_{FSR}	*N*1	N_{max}	of N_{FSR}	*M*	*SD*	EPR	*N* Efficiency	P Efficiency
.10	30	15	45	.085	19.414	8.684	.085	35.29	0.00
	40	20	60	.097	25.787	11.431	.100	35.53	−3.09
	50	25	75	.109	32.231	14.285	.110	35.54	−0.92
	60	30	90	.121	39.031	17.511	.123	34.95	−1.65
	70	35	105	.133	45.456	20.309	.136	35.06	−2.26
	80	40	120	.145	52.432	23.825	.143	34.46	1.38
.30	30	15	45	.376	22.378	10.389	.392	25.41	−4.26
	40	20	60	.475	29.266	13.192	.478	26.84	−0.63
	50	25	75	.564	36.839	16.026	.576	26.32	−2.13
	60	30	90	.642	43.062	18.224	.635	28.23	1.09
	70	35	105	.709	49.164	20.115	.698	29.77	1.55
	80	40	120	.765	54.850	21.670	.745	31.44	2.61
.50	30	15	45	.782	21.255	8.247	.789	29.15	−0.90
	40	20	60	.885	25.933	8.972	.880	35.17	0.56
	50	25	75	.942	30.262	9.101	.941	39.48	0.11
	60	30	90	.972	34.237	8.539	.961	42.94	1.13
	70	35	105	.987	38.368	7.728	.981	45.19	0.61
	80	40	120	.994	42.480	6.604	.988	46.90	0.60

Note—N_{FSR}, sample size with the fixed-sample rule; EPR, empirical proportion of rejections of H_0 (here it is an estimation of the correct rejections rate); *N* Efficiency, percentage of savings in number of subjects, following Equation 2; P Efficiency, percentage of decrease in power of N_{FSR}, following Equation 3.

N_{FSR}). The initial *N* is half of N_{FSR}, and the maximum *N* to stop the process is 1.5 times N_{FSR}. In both the COAST and the CLAST rules, the new observations are incorporated one by one into the sample at each step of the process. A variant of the CLAST rule (named the *CLAST-T rule*), in which more than one observation can be incorporated at each step, is also tested. Those SSRs reflect more realistically the practice of experimental researchers, who obviously are not willing to incorporate an unlimited number of observations into their samples.

For the sake of simplicity, we have set *N*2 to 1 in all our simulations for the CLAST rule. As one reviewer pointed out, our simulations follow a very specific research scenario, in which the outcome for each individual is available immediately after he or she is incorporated into the study. On the other hand, in some circumstances there is a lag between enrollment in the study and the determination of the outcome—for instance, when the study involves the application of an educational program. In such circumstances, group sequential testing is more appropriate. Group sequential testing assumes $N2 > 1$, reflecting the fact that groups of individuals are incorporated into the experiment. The investigation of power and efficiency for group sequential designs is an important topic that we defer for future research. However, in practical applications *N*2 should be kept as low as possible because increasing *N*2 reduces efficiency.

The efficiency of the proposed rules, in terms of power and savings in number of observations while maintaining Type I error rate at .05, has been tested through simulation studies. We used two types of hypothesis tests: the one-tailed *t* test of mean differences with two matched samples and the chi-square independence test for twofold contingency tables. Results showed that the upper bound of the uncertainty region (denoted here by α_u), which maintains the Type I error rate at .05, is .25 in the *t* test and .35 in the chi-square independence test. The CLAST rule is more efficient than the COAST rule in terms of power and savings in number of subjects. However, the CLAST-T rule does not improve the efficiency of CLAST and complicates the procedure.

In order to provide a more direct comparison with COAST, simulations were also conducted for the two-tailed *t* test of mean differences. For the sake of brevity, we have not included the results. The conclusions are as follows: First, the value of α_u that maintains the Type I error rate at .05 is .36, as was found by Frick (1998). This is an important point that should be highlighted. The .36 value proposed by Frick is correct for two-tailed tests, but for one-tailed tests it should be replaced by .25. In short, the upper limit of the uncertainty region must be empirically determined for each statistical test and type of contrast. Second, the efficiency values, in terms of both number of observations and power, are similar to those found with the CLAST rule in the one-tailed test. Therefore, the differences between CLAST and COAST are due not to the directionality of the test, but to the fact that CLAST (but not COAST) has a maximum sample size (N_{max}).

Overall, our results coincide with those of Frick (1998) in terms of a comparison between the SSRs and the FSRs. SSRs have a clear practical advantage in experimental data analyses that involve hypothesis testing with a one-tailed *t* test of mean differences with two matched samples and with a chi-square independence test for twofold contin-

gency tables. However, the rule should include an explicit maximum number of subjects as a secondary criterion for stopping and making a decision about H_0.

The CLAST rule can be applied almost directly to any test after the proper adjustments have been made to the lower and upper limits of the uncertainty region and of $N1$ and $N2$. However, notable exceptions that deserve a more detailed analysis are those cases in which the strategy for analyzing a set of data itself involves several sequential tests. The most obvious example is the sequence "ANOVA + post hoc comparisons." Some relevant questions are whether or not the CLAST rule should be applied only to the F test with no post hoc comparisons until the final decision is made, and what the nominal value of α should be in the post hoc comparisons. More studies are needed in which these more complex scenarios are analyzed.

As Frick (1998) noted, the use of sequential stopping rules is not very common among experimental psychologists, sometimes because they believe that the only acceptable stopping rule is the fixed-sample rule and other times because they ignore the existence of the SSRs. Thus, more effort is needed to develop sequential sampling rules such as COAST and CLAST, and to make researchers aware of their benefits.

REFERENCES

Allison, D. B., Silverstein, J. M., & Gorman, B. S. (1996). Power, sample size estimation, and early stopping rules. In R. D. Franklin, D. B. Allison, & B. S. Gorman (Eds.), *Design and analysis of single-case research* (pp. 335-371). Mahwah, NJ: Erlbaum.

Arnold, D. H., & Harvey, E. A. (1998). Data monitoring: A hypothesis-testing approach for treatment–outcome research. *Journal of Consulting & Clinical Psychology*, **66**, 1030-1035.

Becker, R. A., Chambers, J. M., & Wilks, A. R. (1988). *The new S language: A programming environment for data analysis and graphics*. New York: Chapman & Hall.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Fiske, D. W., & Jones, L. V. (1954). Sequential analysis in psychological research. *Psychological Bulletin*, **51**, 264-275.

Frick, R. W. (1998). A better stopping rule for conventional statistical tests. *Behavior Research Methods, Instruments, & Computers*, **30**, 690-697.

Ghosh, B. K., & Sen, P. K. (Eds.) (1991). *Handbook of sequential analysis*. New York: Dekker.

Hays, W. L. (1988). *Statistics* (4th ed.). Philadelphia: Holt, Rinehart & Winston.

Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*. Boca Raton, FL: Chapman & Hall.

Kimball, A. W. (1950). Sequential sampling plans for use in psychological test work. *Psychometrika*, **15**, 1-15.

Kirk, R. E. (1995). *Experimental design: Procedures for the behavioral sciences* (3rd ed.). Belmont, CA: Brooks/Cole.

Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials*, **2**, 93-113.

Lai, T. L. (2001). Sequential analysis: Some classical problems and new challenges. *Statistica Sinica*, **11**, 303-408.

Lan, K. K. G., & DeMets, D. L. (1989). Changing frequency of interim analysis in sequential monitoring. *Biometrics*, **45**, 1017-1020.

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, **35**, 549-556.

O'Brien, R. G., & Muller, K. E. (1993). Unified power analysis for *t*-tests through multivariate hypotheses. In L. K. Edwards (Ed.), *Applied analysis of variance in behavioral science* (pp. 297-344). New York: Dekker.

Siegmund, D. (1985). *Sequential analysis*: *Tests and confidence intervals*. New York: Springer.

Siegmund, D. (1994). A retrospective of Wald's sequential analysis: Its relation to challenge-point detection and sequential clinical trials. In S. S. Gupta & J. O. Berger (Eds.), *Statistical decision theory and related topics* (pp. 9-33). New York: Springer.

Venables, W. N., & Smith, D. M. (2001). *An introduction to R*. Retrieved February 16, 2005 from http://www.r-project.org/.

Vos, H. J. (2001). A minimax procedure in the context of sequential testing problems in psychodiagnostics. *British Journal of Mathematical & Statistical Psychology*, **54**, 139-159.

Wald, W. (1947). *Sequential analysis*. New York: Dover.

Wetherill, G. B., & Glazebrook, K. D. (1986). *Sequential methods in statistics*. London: Chapman & Hall.

Whitehead, J., & Brunier, H. (1990). The double triangular test: A sequential test for the two-sided alternative with early stopping under the null hypothesis. *Sequential Analysis*, **9**, 117-136.

Winer, B. J. (1971). *Statistical principles in experimental design* (2nd ed.). New York: McGraw-Hill.

NOTE

1. This estimation of the proportion of Type I error is very similar to that offered by Frick (1998), who set an α level for each test equal to .05 and an upper bound of .36 with $N1 = 20$. He obtained a proportion of Type I error in two-tailed tests equal to .13 if the data were reanalyzed after each new observation was incorporated ($N2 = 1$) and without setting an N_{max} as a secondary stopping criterion.

APPENDIX A

In the one-tailed test of mean differences with two matched samples, effect size is defined as follows. Let D_i be equal to $X_{i1} - X_{i2}$, where i refers to a pair of matched observations (often called *repeated measures* in situations in which the same subject responds on two occasions). It is assumed that D is normally distributed, with mean μ_D and *SD* σ_D.

Under H_0 true, $\mu_D = 0$. Thus, D follows an $N(0, \sigma_D)$ distribution and its test statistic is

$$T = \frac{\overline{D}}{S_D / \sqrt{N}}, \tag{A1}$$

where T follows a Student's t distribution with $N - 1$ degrees of freedom.

Under H_0 false, $\mu_D \neq 0$. Thus, D follows an $N(\mu_D, \sigma_D)$ distribution and the effect size, δ, is defined as

$$\delta = \frac{|\mu_D|}{\sigma_D}. \tag{A2}$$

T follows the noncentral Student's t distribution with $N - 1$ degrees of freedom, where the noncentrality parameter λ is

$$\frac{|\mu_D|\sqrt{N}}{\sigma_D}$$

(henceforth denoted by t'). Note that λ is equal to the effect size (δ) times the square root of the sample size (N).

In the simulations of the present study, it is assumed that D follows an $N(\mu_D, 1)$ distribution. If $\mu_D = 0$, H_0 is true and therefore $\delta = 0$. If $\mu_D > 0$, H_0 is false and therefore $\delta = \mu_D$.

In the one-tailed test, the critical value ${}_{1-\alpha}t_{N-1}$ is obtained for a given α. When H_0 is false, the test statistic follows the t' distribution and the probability associated to the value of ${}_{1-\alpha}t_{N-1}$ in the t' distribution is the value of β. That is,

$${}_{1-\alpha}t_{N-1} = {}_{\beta}t'_{N-1}. \tag{A3}$$

In the simulations of the present study, (1) when effect size is zero (H_0 true), the proportion of rejections of H_0 is an estimate of α, the probability of Type I error; (2) when effect size is larger than zero (H_0 false), the proportion of rejections of H_0 is an estimate of the power $1 - \beta$, where β is the probability of Type II error.

APPENDIX B

In the chi-square independence test for twofold contingency tables, effect size is defined as follows. Let X and Y be two Bernoulli variables with values 1 and 2. The joint probability of the (X, Y) pairs is described by the probability vector $\boldsymbol{\pi}' = \{\pi_{11}, \pi_{12}, \pi_{21}, \pi_{22}\}$. Let π_{i+} and π_{+j} be the marginal probabilities of X and Y, respectively. The null hypothesis of independence is H_0: $\boldsymbol{\pi} = \boldsymbol{\pi}_0$, where each of the elements of $\boldsymbol{\pi}_0$ is given by

$$\pi_{0ij} = \pi_{i+} \cdot \pi_{+j}; \quad i = \{1, 2\}, j = \{1, 2\}.$$

If the sample size is n, the effect size is defined as

$$\begin{aligned} w &= \sqrt{\frac{\chi^2}{n}} \\ &= \sqrt{\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(\pi_{ij} - \pi_{0ij}\right)^2}{\pi_{0ij}}}, \end{aligned} \tag{B1}$$

where χ^2 is the chi-square Pearson statistic, which follows the chi-square distribution with one degree of freedom if H_0 is true (this distribution is denoted by P_{π_0}). The value of w is interpreted as the Pearson correlation between two dichotomous variables. The value of w ranges from 0 to 1. The values .1, .3 and .5 are indicative of small, medium and large effect sizes, respectively (Cohen, 1988).

The true distribution of the statistic χ^2 (denoted by P_π) is the noncentral chi-square distribution with the noncentrality parameter

$$\begin{aligned} \lambda^2 &= nw^2 \\ &= n\sum_{i=1}^{2}\sum_{j=1}^{2}\frac{\left(\pi_{ij} - \pi_{0ij}\right)^2}{\pi_{0ij}}. \end{aligned}$$

APPENDIX B (Continued)

Therefore, if c is the critical value for a given α [i.e., $P_{\pi_0}(\chi^2 > c) = \alpha$], the power is the probability of $\chi^2 > c$ under the true distribution—that is, $P_{\pi}(\chi^2 > c)$.

The goal of this appendix is to specify how the vector $\boldsymbol{\pi}$ can be computed from a fixed value of w. The only previous constraint imposed on this vector is

$$\pi_{11} + \pi_{12} + \pi_{21} + \pi_{22} = 1. \tag{B2}$$

Therefore, Equation B1 has three unknowns and infinite solutions. To solve this problem, additional constraints are imposed, which drive to a unique solution. First, the elements of $\boldsymbol{\pi}_0$ are fixed at .25. Then, Equation B1 reduces to

$$.25\, w^2 = \sum_{i=1}^{2}\sum_{j=1}^{2}\left(\pi_{ij} - .25\right)^2 \Rightarrow$$

$$.25\left(w^2 + 1\right) = \sum_{i=1}^{2}\sum_{j=1}^{2}\pi_{ij}^2.$$

Moreover, the two diagonals of the contingency table are assumed to be equal—that is,

$$\pi_{11} - \pi_{22} = 0 \tag{B3}$$

and

$$\pi_{12} - \pi_{21} = 0. \tag{B4}$$

Consequently, all the marginal probabilities are equal to .5. Then, the equation takes the form

$$.25\left(w^2 + 1\right) = 2\pi_{11}^2 + 2\pi_{12}^2 \Rightarrow$$

$$4\pi_{11}^2 - 2\pi_{11} + .5 - .25\left(w^2 + 1\right) = 0.$$

This is a polynomial equation with one unknown that allows two possible solutions. The fourth constraint is imposed in order to choose one of them:

$$\pi_{11} - \pi_{12} \geq 0. \tag{B5}$$

As an example, Table B1 shows the vectors of probabilities associated to different fixed values of w assuming the restrictions above.

Table B1
Parameter Values and Effect Sizes (w)
for the Multinomial Distribution

w	π_{11}	π_{12}	π_{21}	π_{22}
.00	.250	.250	.250	.250
.10	.275	.225	.225	.275
.20	.300	.200	.200	.300
.30	.325	.175	.175	.325
.40	.350	.150	.150	.350
.50	.375	.125	.125	.375
.60	.400	.100	.100	.400
.70	.425	.075	.075	.425
.80	.450	.050	.050	.450
.90	.475	.025	.025	.475
1.00	.500	.000	.000	.500

(Manuscript received August 24, 2004;
revision accepted for publication March 7, 2005.)