Behavior Research Methods
2010, 42 (1), 3-22
doi:10.3758/BRM.42.1.3

Improved stopping rules for the design of efficient small-sample experiments in biomedical and biobehavioral research

Douglas A. Fitts
University of Washington, Seattle, Washington

Sequential stopping rules (SSRs) should augment traditional hypothesis tests in many planned experiments, because they can provide the same statistical power with up to 30% fewer subjects without additional education or software. This article includes new Monte-Carlo-generated power curves and tables of stopping criteria based on the *p* values from simulated *t* tests and one-way ANOVAs. The tables improve existing SSR techniques by holding alpha very close to a target value when 1–10 subjects are added at each iteration. The emphasis is on small sample sizes (3–40 subjects per group) and large standardized effect sizes (0.8–2.0). The generality of the tables for dependent samples and one-tailed tests is discussed. SSR methods should be of interest to ethics bodies governing research when it is desirable to limit the number of subjects tested, such as in studies of pain, experimental disease, or surgery with animal or human subjects.

In this article, I propose a simple and efficient sequential stopping rule (SSR) that should augment the usual null hypothesis significance test in many experiments in the biomedical and biobehavioral sciences. Sequential or adaptive methods have long been available to avoid an inflation of alpha (Wald, 1947), but they require mathematical sophistication and are therefore not easily accessible to many researchers. Stopping rules and adaptive designs are well known in clinical trials in which an interim assessment of the results before the end of a trial can save lives and in which adjustments must be made in order to avoid an inflation of alpha (Bretz, Koenig, Brannath, Glimm, & Posch, 2009). In this article, I improve a simple SSR technique proposed by others (Botella, Ximénez, Revuelta, & Suero, 2006; Frick, 1996; Ximénez & Revuelta, 2007). It can augment the usual null hypothesis test in ordinary planned experiments as long as the data can be collected in stages. It requires no additional education or statistical software. Alpha is controlled at its nominal level. The widespread use of this technique can save money and, in animal experiments, thousands of lives each year.

The null hypothesis significance test has many faults (see the Should We Be Doing It This Way? subsection in the Discussion section), and one of them is inefficiency. In a typical planned experiment, the investigator does a power analysis, conducts an experiment, and analyzes the data once, for better or worse. Small errors in the estimation of parameters in the power analysis can yield large errors in the power of the test, and if the obtained *p* value is not less than the desired alpha, the investigator may be left with little to say after all the trouble. Or, if the experiment was overpowered, resources and subjects may have been wasted in order to detect a larger-than-expected effect. In either case, human or animal subjects may have been used needlessly in the experiment. This inflexible procedure, in which an experiment is stopped after testing a fixed number of subjects, is called the *fixed stopping rule*. It is the way the null hypothesis test was designed.

Adding sample size to an already completed experiment in order to increase power will increase the Type I error rate (alpha) unless extraordinary measures are taken (e.g., Timmesfeld, Schäfer, & Müller, 2007), but unfortunately, many investigators are not aware of this. More than a few investigators submitting Institutional Animal Care and Use Committee (IACUC) protocols to our office describe this erroneous procedure as a good thing, because it is efficient and saves animals. They say that they conduct the experiment with a few animals in a pilot study, test at the $\alpha = .05$ level, and then make a decision as to whether to publish, add additional animals, or abandon the project. In fact, they are not necessarily saving animals, because they are increasing the probability that the animals will have been used to support a conclusion that is unfounded. The investigators are correct about one thing, however, and that is that the procedure is efficient if the null hypothesis is false. They are often able to reject the null hypothesis with many fewer animals than would be predicted for the same power with a power analysis.

The technique proposed here is derived from Frick (1998), who described a simple SSR that applies to ordinary planned experiments in which the goal is to decide whether a treatment has an effect and, if so, in which di-

D. A. Fitts, dfitts@u.washington.edu

rection. The SSR, called COAST (composite open adaptive stopping rule), allows an investigator to conduct null hypothesis tests in stages, analyzing the data at several points along the way, without inflating alpha beyond .05. The investigator collects data from a few subjects and conducts a *t* test or ANOVA. If the *p* value is less than a lower criterion of .01, the investigator declares the effect significant at the $\alpha = .05$ level and stops testing. If the *p* is greater than an upper criterion of .36, the investigator stops the experiment and retains the null hypothesis. If *p* is between these two values (uncertain), additional subjects are added and the process is repeated. The SSR is highly efficient and may use 30% fewer subjects to generate the same statistical power as a traditional fixed stopping rule. It is an excellent choice for experiments in which an indefinite number of subjects can be added to an experiment a few at a time until a decision is reached one way or the other about the null hypothesis. However, if the number of potential subjects is limited, the experimenter may need to stop early, before a decision is reached about the null hypothesis, and this can reduce alpha and the efficiency of the technique. That is, the observed alpha may be .02 instead of .05, and as many subjects may be required as with a fixed stopping rule if many more than 1 or 2 subjects are added at each iteration.

A derivative rule, CLAST (composite limited adaptive stopping rule; Botella et al., 2006), was designed for researchers who have a more limited pool of potential subjects. The investigator selects a sample size on the basis of the fixed stopping rule via a power analysis. Testing begins with half of that fixed sample size, and subjects can be added 1 or a few at a time until a maximum sample size of 1.5 times the fixed-sample *n* is reached. Botella et al. agreed with Frick (1998) that the lower criterion should always be .01, primarily because it is a customary alpha used by researchers, and the authors empirically selected an upper criterion using Monte Carlo simulations. For a one-tailed, dependent-samples *t* test, the upper criterion was .25; for twofold contingency tables, it was .35; for a general fixed-effects ANOVA with four groups, it was .50; and for linear contrasts among four means, it was .45 (Ximénez & Revuelta, 2007).

The sample sizes tested using a dependent-samples *t* test with COAST or CLAST were rather large (>16) and the effect sizes rather small, as were appropriate to the experiments in social and cognitive psychology for which they were designed. The later application of CLAST to ANOVAs included some beginning sample sizes as low as 2 per group, but the effect sizes tested were so small that the method was not seriously validated for the smaller sample sizes and larger effects that are common in experimental biomedical and biobehavioral research. A savings in sample size would happen best with large samples, because at some point with small samples one would reach a basement so that further reductions are difficult.

In the present article, I validate the method with small sample sizes and large effect sizes and present a method to set both the lower and upper criteria of an SSR for a *t* test or one-way ANOVA so that the Type I error is maintained very near the selected alpha regardless of the number of subjects to be added per iteration (*n added*). I have abandoned the reliance on a fixed lower criterion of $\alpha = .01$, and instead find a combination of both criteria to produce an experimentwise alpha near .005, .01, .05, or .10. This new version of the method provides good power and efficiency with large effects across a selection of sample sizes and allows the investigator to design an experiment adding 1–10 subjects per group per iteration. The technique is useful when data can be collected in stages with a few subjects at a time and is particularly appropriate when subjects are rare, expensive, or ethically constrained (e.g., operated humans or animals, pain studies, death as an endpoint, etc.).

Although the description of the derivation of the power curves and tables of lower and upper criteria is a bit tedious, the method for designing experiments is simple and straightforward. To find the appropriate criteria, one first looks up the desired effect size and experimentwise alpha in a graph to determine the available models that can produce the desired power. One then chooses the desired number of subjects to be added at each iteration (called *n added*) and consults a table for the selected model to determine the lower and upper criteria to be used in the experiment.

A caution on terminology is in order to prevent confusion. *Lower and upper criteria* are the probability values that allow one to declare significance (lower criterion) or to stop the experiment because the *p* is too high (upper criterion). The *lower and upper bounds* are the sample size per group at the first iteration (lower bound) and the size beyond which the experiment will not proceed in later iterations (upper bound). The *fixed stopping rule* is the rule in which a fixed number of subjects is tested before stopping and analyzing the data for better or worse. A *fixed-criteria SSR* is a rule (such as COAST or CLAST) that allows subjects to be added sequentially, either individually or in clusters, and in which the same fixed lower and upper criteria are used for each type of test regardless of the lower and upper bounds or the *n added* at each iteration. All SSRs are at least as efficient as the fixed stopping rule and are usually more efficient (see below). In the *variable-criteria SSR* proposed here, customized criteria are used for each set of lower and upper bounds and each level of *n added*. These are decided on at the beginning of the experiment, and the criteria are obtained from a table. The variable-criteria SSR is generally a bit more efficient than the fixed-criteria sequential stopping rules, and, importantly, it holds alpha constant for different upper and lower bounds and different *n added* per iteration.

METHOD

These Monte Carlo studies were conducted using custom programs in the C programming language that were executed on a Dell XPS 400 computer equipped with dual Intel Pentium D CPU 2.80GHz processors, 1.00 GB of RAM, and mirrored 169-GB hard drives. Data were sampled using a pseudorandom number generator based on Ran2() (Press, Teukolsky, Vetterling, & Flannery, 1992),

which returns uncorrelated uniform deviates between 0 and 1.0 with a period greater than 2×10^{18}. The function was seeded on the first call using the system clock so that each sequence would be different. Because of the seeding and the large period of the generator, it is improbable that any long sequence of numbers was correlated or repeated in these simulations. In most cases, these random numbers were transformed to a unit normal distribution, and samples of 2^{20} of the generated numbers were accumulated and analyzed for normality in each run. The normal deviates were then transformed linearly using the desired means and standard deviations to create the generated samples. In some cases, skewed data were generated by taking the absolute value of the random normal deviate before a linear transformation with the mean and standard deviation to produce a markedly positively skewed (third *z*-score moment ~1.0) and slightly leptokurtic distribution.

The experimental designs include both independent- and dependent-samples *t* tests with either a one- or two-tailed hypothesis (four variations) and a one-way ANOVA with four independent groups. The first to be described is the independent-samples *t* test with a two-tailed alpha. The generated *p* values are identical to those for a one-way between-groups ANOVA for fixed effects with two groups.

Two-Tailed, Independent-Samples *t* Test

All simulations were conducted 100,000 times except for the contrived example in Table 1B (to reduce the size of the table). The following parameters were included in the simulations.

The combinations of lower/upper bounds (models) were 3/9, 3/15, 4/10, 4/18, 5/12, 5/19, 6/12, 6/18, 7/14, 7/21, 8/24, 8/32, 9/27, 9/36, 10/30, and 10/40.

Table 1A
Illustration of a Theoretical Distribution of 10,000 Observations After Four Sequential *t* Tests With *n*s Equal to 10, 11, 12, and 13 per Group Assuming That All Observations Are Independent

	Test 1	Test 2	Test 3	Test 4	Cumulative *f*	Cumulative *p* (α)
n tested	10	11	12	13		
S_1	500	500	500	500	500	.05
U_1S_2	0	155	155	155	655	.0655
U_2S_3	0	0	48.05	48.05	703.05	.0703
U_3S_4	0	0	0	14.8955	717.9455	**.0718 (EPR)**
Uncertain	3,100	961	297.91	92.3521	810.2976	.0810
U_3NS_4	0	0	0	190.6624	1,000.96	.1001
U_2NS_3	0	0	615.04	615.04	1,616	.1616
U_1NS_2	0	1,984	1,984	1,984	3,600	.36
NS_1	6,400	6,400	6,400	6,400	10,000	1.0
Total	10,000	10,000	10,000	10,000		

Note—The lower criterion was .05, the upper criterion was .36, the lower bound was $n = 10$/group, the upper bound was $n = 13$/group, and the *n added* per group with each sequential test was 1. S_i, U_i, and NS_i, the numbers that were significant (S, $p \leq .05$), uncertain (U, $.05 < p \leq .36$), or not significant (NS, $p > .36$) on the *i*th test (e.g., U_1S_2 is the number that were uncertain on the first test and significant on the second test); Cumulative *f*, cumulative frequency, includes the total number determined to be significant in all experiments, ~718, or a cumulative alpha of .0718 (in bold). In Test 1 with $n = 10$, 5% of the 10,000 experiments (500) were significant, 31% (3,100) were uncertain, and 64% (6,400) were not significant. In Test 2, the 3,100 uncertain in Test 1 were tested again after adding a subject to each group ($n = 11$), and 5% were significant (155), 31% were uncertain (961), and 64% were not significant, etc. The cumulative frequencies illustrate that the outcomes of all 10,000 experiments are accounted for.

Table 1B
Illustration of the Results From a Monte Carlo Simulation of 10,000 Observations of the Same Design in Which the Cumulative Alpha for All Significant Tests, .0846 (in Bold), Is Greater Than the Value in the Model That Assumes Independence

	Test 1	Test 2	Test 3	Test 4	Cumulative *f*	Cumulative *p* (α)
n tested	10	11	12	13		
S_1	503	503	503	503	503	.0503
U_1S_2	0	135	135	135	638	.0638
U_2S_3	0	0	112	112	750	.075
U_3S_4	0	0	0	96	846	**.0846 (EPR)**
Uncertain	3,097	2,321	1,826	1,449	2,295	.2295
U_3NS_4	0	0	0	281	2,576	.2576
U_2NS_3	0	0	383	383	2,959	.2959
U_1NS_2	0	641	641	641	3,600	.36
NS_1	6,400	6,400	6,400	6,400	10,000	1.0
Total	10,000	10,000	10,000	10,000		

Note—Alpha, which must be estimated from simulations, is greatly inflated with this experimental strategy of testing repeatedly with $\alpha = .05$. For a simulation, the empirical proportion of rejections (EPR) is an estimate of alpha if the null hypothesis is true and an estimate of power if the null hypothesis is false.

The ranges of *n added* per iteration were 1–6 in smaller models and up to 1–10 in the largest model. The number of iterations was determined by the lower and upper bounds and the number of *n added* as follows. In the first iteration, both samples contained the number of subjects at the lower bound (e.g., 7 for the two models with 7 as a lower bound and either 14 or 21 as an upper bound). In the second and successive iterations, both groups were augmented by *n added.* The process was stopped when the addition of *n added* subjects would exceed the upper bound. Thus, the 7/21 model required more iterations than the 7/14 model because of the larger upper bound.

The standardized effect sizes included 0 (null), 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, and 2.0. The standardized effect size for an independent-samples *t* test is the difference between the means divided by the pooled standard deviation. Cohen (1988) labeled 0.8 as a large effect, but it is the smallest effect in this set.

The four levels of experimentwise alpha were .005, .01, .05, and .10.

Estimation of Alpha

In a simulation in which the point null hypothesis was true (effect size = 0.0), alpha was estimated as the empirical proportion of rejections (EPR; Botella et al., 2006). The EPR is the proportion of experiments in which the obtained *p* was less than or equal to alpha after all iterations, and the method is demonstrated in Tables 1A and 1B and described below.

For the initial simulation, the lower and upper criteria were estimated on the basis of a rough guess, and the EPR for those criteria was determined in each of eight independent simulations of 100,000 *t* tests using the same criteria for all simulations. A 95% confidence interval for the grand EPR, the observed alpha, was determined on the basis of the standard error of the eight independent means. The percentage deviation of the grand mean EPR from the nominal alpha was also calculated. If the nominal alpha (e.g., $\alpha = .05$) was not contained within the 95% confidence interval, or if the absolute value of the percentage deviation of the grand mean from the nominal alpha was greater than 1.0%, the criteria were adjusted either upward or downward. The data set was again probed with the revised criteria to produce a new set of means and a new standard deviation of the means. The lower and upper criteria for the simulation were thus adjusted by successive approximations until two conditions were both met: (1) the nominal alpha fell within the 95% confidence interval for the mean EPR, and (2) the percentage deviation of the mean EPR from the nominal alpha was 1% or less. For example, for the $\alpha = .05$ level, the value .05 had to be included in the 95% confidence interval for the mean EPR based on the eight simulations, and the mean EPR had to fall between .0495 and .0505 ($\pm 1\%$). The successful identification of a pair of criteria in this fashion typically required about 10–30 computer-assisted approximations from the data set. The lower and upper criteria for the final run that satisfied both conditions are reported in Table 2 and were used in subsequent simulations for the estimation of power for various effect sizes when the null hypothesis was not true. The calculated EPR for those simulations in which the null hypothesis was false then served as an estimate of power for the selected effect size.

These criteria are not unique. Other combinations of lower and upper criteria could satisfy both conditions as well. However, it was also necessary to set an upper limit on the number of significant digits for the lower and upper criteria. For example, it might be possible to set a constant upper criterion, such as .45, and have all of the adjustment for alpha made with the lower criterion (the opposite of COAST and CLAST). To do so would require as many as five or six significant digits for the lower criterion, and such false precision is highly undesirable. For that reason, I established two general rules that were almost always obeyed: (1) The criteria should be limited to three significant digits, and (2) the upper criterion for *p* should fall within the range of .2–.5. There are a few exceptions to these general rules, and they were broken only when it was otherwise impossible to meet both of the previous conditions for stability in the criteria.

The use of eight independent replications was necessary, primarily for the more stringent alphas of .01 and especially .005, because the events were rare enough even with 100,000 samples that multiple simulations were required to achieve stability. Stability was easier to achieve for the $\alpha = .05$ and .10 levels with fewer simulations.

To begin a simulation, a single pair of lower and upper bounds was selected on the basis of pilot simulations, the effect size was set to zero (point null hypothesis true), and an initial 100,000 experiments were randomly sampled in two independent groups from populations with identical means and standard deviations, with the *n* in each group equal to the lower bound. A two-tailed *t* test was performed between the groups, and the resulting *p* value was saved for each of the 100,000 tests. One subject was then added to each existing group, and the 100,000 *t* tests were performed again with the augmented sample size. The process was repeated, adding 1 subject per group at each iteration until the total sample size in each group was equal to the upper bound. This produced a first file of 100,000 *p* values at each of the sample sizes from the lower bound to the upper bound under the null hypothesis. The process was repeated at each of the different levels of *n added* to produce additional files. For example, in the 7/21 model, there were 15 columns of *p* values in the first file in which *n added* was 1 (from 7 to 21, inclusive). In the second simulation, 2 subjects were added at each iteration, so there were 8 columns in the file, representing *n*s of 7, 9, 11, 13, 15, 17, 19, and 21. For 3 subjects added, there were 5 columns, representing *n*s of 7, 10, 13, 16, and 19. Note that the process was stopped at 19 because the addition of 3 subjects would exceed the upper bound of 21. Table 2 gives the number of levels of *n added* for each model; seven levels were used with the 7/21 model, so one file was created at each level of *n added* for 7 total files. The entire process was replicated eight independent times, for a total of 56 total files in the null hypothesis condition of the 7/21 example. These files represented the basic data set for the null hypothesis for each model, and this data set was then probed using different criteria in order to determine a set of criteria for each level of *n added* that satisfied the two conditions for accuracy in the null hypothesis condition.

Determination of EPR

For each file, the first column represented the number of subjects at the lower bound, and the last column represented the number of subjects tested when the addition of *n added* subjects would exceed the upper bound in the next iteration (see the examples in the previous paragraph). Each line represented a single complete experiment after all iterations. All lines of the file were sorted into order on the basis of the first column of *p* values. The *p* values in the first column that were less than or equal to the lower criterion were counted and eliminated in all columns. Values of *p* greater than the upper criterion were also eliminated in all columns. The remaining *p* values in the second column thus represented only the experiments in the *uncertain* range (*p* greater than the lower criterion and less than or equal to the upper criterion) of the first column. The remaining lines of the file were then sorted on the basis of the values in the second column in order to determine the number of values less than or equal to the lower criterion or above the upper criterion. The *uncertain* ranges of each successive column were sequentially analyzed in this fashion for each augmented *n* until the upper bound was reached or until a decision about the null hypothesis had been reached for all experiments.

The EPR is difficult to calculate mathematically, and that is why it is approximated in simulations (Botella et al., 2006; Frick, 1998). Tables 1A and 1B demonstrate the technique for determining the total EPR for a small example with the null hypothesis true (effect size = 0), a lower criterion of .05, an upper criterion of .36, lower and upper bounds of 10 and 13 per group, an *n added* of 1 per iteration, and 10,000 experiments. Table 1A gives the expected number at each iteration as if all of the tests were independent and easy to

calculate. Table 1B gives the result of a simulation. Notice how, in the theoretical computation in Table 1A, the size of the *uncertain* region rapidly diminishes as 64% of each iteration is eliminated as *not significant*. In the simulation in Table 1B, the *uncertain* region begins very near the theoretical value but, instead of diminishing, stays large throughout testing. When 10 subjects per group have already been tested and the mean difference is in the *uncertain* range, the addition of a single additional subject per group will only very rarely cause a wide change to *significant* or *not significant*. Instead, the observations are correlated, and the *uncertain* range stays large. This produces a larger pool of potential Type I errors at each iteration, and as a result, the estimated alpha is increased from the theoretical value of .0718 in Table 1A (already a bad inflation of alpha) to .0846 in the simulation in Table 1B.

Determination of Lower and Upper Criteria

Lower and upper criteria for four levels of alpha were determined independently by successive approximations at each level of *n added* for each model (i.e., combination of lower and upper bounds) in the null hypothesis condition until the two conditions in the Estimation of Alpha section (above) were met. Note that the values at each level of *n added* were determined from a new simulation instead of by re-using the original data from the data set with an *n added* of 1 per iteration. This prevents any bias or correlation in the criteria based on the vagaries of a particular data set. Each pair of criteria was based on the average EPR derived from the eight independent simulations of 100,000 *t* tests in the basic data set when the null hypothesis was true.

Estimation of Power and Mean Sample Size

Using the stable criteria from the null hypothesis condition just described, a simulation of 100,000 *t* tests was conducted with each of seven standardized true effect sizes, *d*, equal to 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, and 2.0 (Cohen, 1988). For the 7/21 combination for example, this resulted in a total of 49 files written (seven true effect sizes times seven different *n added* values). The EPR calculated for these files served as an estimate of the power of the test at each of the levels of effect size, alpha, and *n added*. In addition to the EPR, the means and standard deviations of the sample sizes for those iterations at which the samples became significant were calculated.

The total number of models tested in the main experiment was 16, 1 for each combination of lower and upper bounds. Additional simulations were conducted as necessary, also with 100,000 samples each, to determine the behavior of the method under fixed-criteria rules borrowed from SSRs such as COAST or CLAST (Botella et al., 2006; Frick, 1998; Ximénez & Revuelta, 2007) or with skewed distributions. These simulations using COAST or CLAST criteria sometimes deviated from the published rules for using the methods. When using COAST criteria, the process was stopped at an upper bound rather than proceeding indefinitely until a decision about the null hypothesis had been reached for all experiments. When using CLAST criteria, an *n* from a fixed stopping rule was not always the exact midpoint of the lower and upper bounds. Thus, the comparisons here show how the fixed sequential stopping criteria would behave outside the formal constraints suggested in these publications (Botella et al., 2006; Frick, 1998; Ximénez & Revuelta, 2007). In one case, there is a formal comparison of the variable-criteria SSR with CLAST, as was intended by Ximénez and Revuelta (2007). This is the only example actually labeled CLAST (see Figure 2). The others are referred to as *fixed-criteria SSRs*.

Dependent-Samples *t* Tests

For dependent samples (one-sample, matched-sample, or correlated samples *t* tests), pairs of scores were sampled from two populations of scores in which the distributions were unit normal and the correlation between the populations was .50. The difference scores for pairs of scores sampled in this fashion have a population mean of 0.0 and a population standard deviation of 1.0. The scores were transformed linearly to produce effects of desired sizes without affecting the underlying distribution or correlation. Other details of the simulations, including all of the models of lower and upper bounds in Table 2, are the similar to those for independent samples, except that new criteria were not derived. Instead, I used the criteria from the independent-samples tests in Table 2 to determine whether new tables were necessary. Figures of the estimated power of the tests were constructed.

One-Tailed (Directional) Tests

A limited number of one-tailed tests was conducted in order to compare them with the two-tailed tests to determine the degree to which a table of criteria for a two-tailed test can be used to determine one-tailed probabilities. These included the models 3/9, 4/18, 5/19, 6/18, 7/21, 8/24, 9/27, and 10/30.

One-Way ANOVA With Four Groups

All 16 models from the independent-samples *t* test were tested using the criteria from Table 2 with four independent groups instead of just two groups in a one-way ANOVA. As with the independent-samples *t* test, eight replications of 100,000 experiments were conducted with the null hypothesis true (the four population means were identical). The data were plotted to demonstrate the stability of alpha at nominal levels of .005, .01, .05, and .10. Power values were estimated for eight large effect sizes with the standardized effect size *f* varying from .4 to .75 in increments of .05 (Cohen, 1988).

RESULTS

The Variable Criteria

The lower and upper criteria for all models tested are displayed in Table 2. To read the table, one must first select an experimentwise alpha, a model of lower and upper bounds, and a number of subjects to add to the uncertain groups at each iteration (*n added*). For example, for an alpha of .05, a lower bound of 5 subjects per group, an upper bound of 19 subjects per group (model 5/19), and an *n added* of 4 subjects per group, the lower and upper criteria are read from Table 2 as .0260 and .200, respectively. These lower and upper criteria from Table 2 constrain alpha for the entire experiment very close to the selected alpha. See the Recommendations for Use subsection in the Discussion section.

As was noted in the Method section, the pairs of criteria were determined by successive approximations. These approximations proceeded by adjusting the lower criterion until the grand mean EPR was in the general neighborhood of the nominal alpha, and then by adjusting the upper criterion until the conditions were met (95% confidence interval contains nominal alpha, and grand mean EPR is within 1% of nominal alpha). One could liken these to the adjustments on a microscope for coarse focus (lower criterion) and fine focus (upper criterion). As can be seen in Figure 1, changing the upper criterion from .25 to .50 has only a very small effect on the observed EPR. On the other hand, small changes in the lower criterion produce large changes in the observed EPR. This is one reason that the upper criteria appear to move in haphazard fashion from one pair of criteria to another and that the lower criterion sometimes requires several significant digits to put the EPR in the correct range. Identifying pairs of criteria that change smoothly from model to model or stay within a

Table 2
Lower and Upper Criteria for Each Combination of Alpha and *n Added* for Each Model of Lower and Upper Bounds

Lower/Upper Bound Model	*n Added*	Alpha .00500 Lower	.00500 Upper	.01000 Lower	.01000 Upper	.0500 Lower	.0500 Upper	.1000 Lower	.1000 Upper
3/9	1	.00110	.200	.00250	.200	.0150	.430	.0500	.219
	2	.00150	.250	.00320	.330	.0200	.330	.0550	.260
	3	.00180	.300	.00400	.240	.0220	.450	.0550	.350
	4	.00260	.300	.00520	.470	.0290	.450	.0630	.440
	5	.00270	.250	.00530	.430	.0300	.390	.0630	.450
	6	.00280	.200	.00530	.450	.0300	.400	.0640	.450
3/15	1	.00080	.290	.00200	.200	.0150	.260	.0300	.470
	2	.00110	.300	.00230	.400	.0150	.450	.0450	.320
	3	.00130	.340	.00300	.280	.0250	.200	.0600	.250
	4	.00160	.310	.00370	.250	.0280	.200	.0590	.290
	5	.00190	.370	.00390	.410	.0250	.350	.0600	.330
	6	.00200	.290	.00490	.200	.0260	.330	.0600	.350
4/10	1	.00120	.190	.00260	.200	.0158	.450	.0400	.350
	2	.00160	.200	.00330	.450	.0200	.380	.0500	.310
	3	.00190	.250	.00400	.250	.0250	.270	.0550	.340
	4	.00260	.200	.00530	.450	.0300	.360	.0630	.450
	5	.00260	.470	.00530	.450	.0300	.370	.0640	.450
	6	.00260	.300	.00530	.450	.0300	.425	.0650	.450
4/18	1	.00086	.350	.00200	.200	.0150	.260	.0350	.360
	2	.00100	.450	.00250	.219	.0150	.410	.0450	.300
	3	.00130	.390	.00300	.310	.0250	.190	.0450	.390
	4	.00156	.425	.00400	.150	.0300	.160	.0550	.300
	5	.00200	.200	.00400	.350	.0300	.200	.0600	.320
	6	.00200	.260	.00400	.390	.0250	.350	.0600	.330
5/12	1	.00120	.200	.00270	.200	.0190	.210	.0500	.229
	2	.00172	.200	.00360	.360	.0210	.440	.0500	.340
	3	.00190	.280	.00400	.430	.0250	.275	.0550	.350
	4	.00269	.200	.00530	.450	.0290	.450	.0630	.450
	5	.00260	.200	.00530	.450	.0290	.460	.0630	.470
	6	.00260	.380	.00530	.450	.0300	.400	.0630	.470
	7	.00260	.300	.00530	.450	.0340	.210	.0630	.490
5/19	1	.00088	.400	.00180	.420	.0140	.310	.0340	.390
	2	.00100	.497	.00240	.275	.0190	.220	.0450	.300
	3	.00140	.350	.00290	.460	.0200	.310	.0450	.400
	4	.00170	.200	.00350	.275	.0260	.200	.0510	.350
	5	.00200	.200	.00438	.200	.0260	.280	.0560	.360
	6	.00200	.200	.00380	.490	.0250	.350	.0610	.310
	7	.00200	.250	.00400	.390	.0300	.210	.0550	.410
6/12	1	.00140	.200	.00300	.420	.0200	.230	.0550	.200
	2	.00179	.600	.00370	.400	.0219	.350	.0500	.360
	3	.00200	.500	.00420	.450	.0250	.290	.0550	.370
	4	.00273	.200	.00540	.380	.0300	.360	.0630	.510
	5	.00273	.200	.00540	.380	.0300	.450	.0640	.440
	6	.00270	.200	.00540	.440	.0300	.430	.0640	.450
6/18	1	.00094	.400	.00200	.500	.0165	.240	.0390	.330
	2	.00120	.320	.00270	.260	.0180	.300	.0440	.330
	3	.00140	.330	.00300	.320	.0200	.300	.0490	.320
	4	.00160	.250	.00375	.200	.0230	.270	.0520	.330
	5	.00190	.500	.00400	.350	.0250	.300	.0590	.320
	6	.00190	.500	.00420	.300	.0260	.300	.0590	.330
7/14	1	.00130	.450	.00290	.200	.0200	.200	.0550	.200
	2	.00189	.500	.00390	.210	.0250	.200	.0500	.400
	3	.00210	.200	.00430	.450	.0250	.360	.0550	.380
	4	.00270	.100	.00560	.400	.0313	.350	.0640	.510
	5	.00270	.200	.00560	.450	.0310	.400	.0640	.500
	6	.00273	.500	.00550	.450	.0300	.430	.0640	.470
	7	.00273	.200	.00550	.430	.0300	.430	.0640	.460
7/21	1	.00094	.500	.00200	.400	.0150	.300	.0350	.400
	2	.00120	.220	.00250	.350	.0200	.210	.0450	.310
	3	.00140	.500	.00300	.490	.0200	.319	.0450	.410
	4	.00160	.250	.00340	.350	.0250	.200	.0500	.370

Table 2 (Continued)

Lower/Upper Bound Model	*n Added*	Alpha .00500 Lower	.00500 Upper	.01000 Lower	.01000 Upper	.0500 Lower	.0500 Upper	.1000 Lower	.1000 Upper
	5	.00200	.200	.00406	.500	.0250	.310	.0550	.380
	6	.00200	.210	.00410	.480	.0250	.350	.0600	.310
	7	.00200	.240	.00400	.430	.0300	.200	.0550	.410
8/24	1	.00090	.200	.00200	.320	.0150	.300	.0350	.390
	2	.00110	.500	.00250	.350	.0160	.420	.0440	.310
	3	.00130	.500	.00300	.260	.0200	.280	.0500	.290
	4	.00150	.220	.00312	.400	.0200	.340	.0500	.320
	5	.00160	.410	.00343	.400	.0250	.210	.0500	.380
	6	.00200	.250	.00438	.275	.0250	.350	.0600	.310
	7	.00200	.200	.00400	.500	.0250	.350	.0600	.320
	8	.00200	.280	.00430	.300	.0300	.200	.0600	.320
8/32	1	.00078	.400	.00187	.238	.0150	.250	.0350	.350
	2	.00094	.310	.00210	.300	.0150	.340	.0360	.400
	3	.00110	.300	.00250	.270	.0190	.260	.0400	.400
	4	.00120	.430	.00280	.250	.0190	.300	.0500	.290
	5	.00140	.450	.00330	.235	.0250	.200	.0500	.340
	6	.00156	.219	.00310	.355	.0300	.150	.0450	.450
	7	.00170	.230	.00390	.200	.0250	.250	.0500	.410
	8	.00180	.200	.00390	.230	.0260	.245	.0550	.340
9/27	1	.00094	.210	.00200	.290	.0130	.450	.0450	.250
	2	.00100	.480	.00230	.370	.0150	.450	.0500	.245
	3	.00120	.450	.00280	.260	.0200	.250	.0600	.210
	4	.00150	.210	.00310	.400	.0200	.350	.0450	.440
	5	.00170	.260	.00375	.220	.0240	.230	.0550	.300
	6	.00170	.250	.00375	.220	.0250	.240	.0500	.400
	7	.00210	.280	.00410	.500	.0250	.350	.0550	.400
	8	.00200	.400	.00410	.450	.0250	.350	.0550	.400
	9	.00200	.300	.00410	.450	.0250	.350	.0550	.430
9/36	1	.00078	.400	.00187	.238	.0150	.250	.0350	.350
	2	.00094	.280	.00200	.355	.0150	.330	.0370	.380
	3	.00100	.280	.00210	.480	.0190	.230	.0530	.240
	4	.00120	.300	.00270	.290	.0200	.250	.0450	.350
	5	.00130	.400	.00300	.266	.0230	.220	.0420	.450
	6	.00150	.300	.00310	.450	.0200	.370	.0450	.430
	7	.00170	.300	.00375	.270	.0250	.250	.0530	.340
	8	.00160	.350	.00375	.240	.0250	.250	.0500	.410
	9	.00170	.280	.00375	.240	.0250	.260	.0540	.350
10/30	1	.00090	.200	.00200	.270	.0150	.280	.0350	.380
	2	.00110	.200	.00230	.390	.0150	.440	.0400	.350
	3	.00120	.400	.00270	.390	.0190	.280	.0440	.350
	4	.00132	.310	.00280	.360	.0200	.280	.0460	.350
	5	.00150	.290	.00320	.320	.0200	.350	.0490	.340
	6	.00160	.450	.00350	.450	.0230	.290	.0490	.420
	7	.00200	.280	.00410	.500	.0240	.410	.0530	.450
	8	.00200	.290	.00410	.450	.0240	.420	.0530	.450
	9	.00200	.320	.00410	.430	.0250	.350	.0540	.430
	10	.00190	.340	.00420	.300	.0250	.360	.0550	.420
10/40	1	.00078	.350	.00180	.240	.0150	.250	.0350	.350
	2	.00094	.290	.00200	.300	.0150	.320	.0400	.330
	3	.00100	.330	.00240	.240	.0190	.230	.0430	.330
	4	.00125	.260	.00280	.210	.0200	.240	.0450	.330
	5	.00130	.240	.00280	.270	.0200	.270	.0490	.310
	6	.00140	.220	.00290	.340	.0210	.280	.0490	.330
	7	.00150	.230	.00310	.370	.0230	.250	.0490	.360
	8	.00164	.380	.00380	.200	.0240	.270	.0500	.400
	9	.00180	.200	.00375	.275	.0250	.250	.0500	.400
	10	.00180	.200	.00370	.280	.0250	.260	.0500	.420

Note—To conduct an experiment, first identify the available models from the appropriate power curves in Figures 6, 8, or 10. Select the number of added subjects (*n added*) desired for each iteration of the experiment, then look up the model here and identify its lower and upper criteria. Test the number of subjects at the lower bound and calculate *p*. If *p* is less than or equal to the lower criterion, reject the null hypothesis at your selected level of alpha; if *p* is greater than the upper criterion, stop the experiment and retain the null hypothesis; otherwise, add *n added* subjects and retest. Repeat this procedure until the upper bound is reached. An effect in the uncertain range at the upper bound is not significant. The observed alpha will remain stable for the experiment at the nominal level.

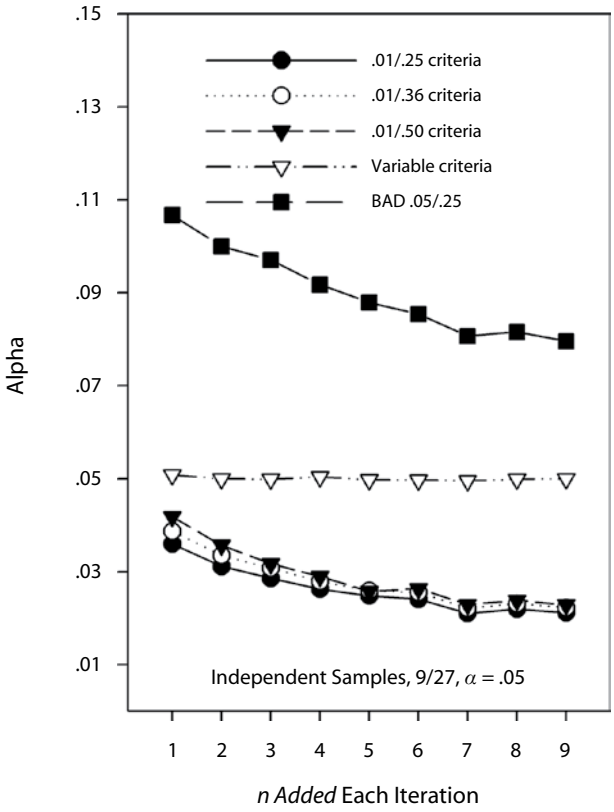

Figure 1. Estimates of alpha under the null hypothesis in the variable-criteria SSR and various fixed-criteria SSRs at a lower bound of 9 and an upper bound of 27 subjects with an overall experimentwise alpha of .05. Fixed criteria derived from COAST and CLAST SSRs (.01/.25, .01/.36, .01/.50) do limit alpha below .05, but they diverge sharply from .05 when more than 1 subject is added per iteration of the experiment. The variable-criteria SSR proposed here holds alpha constant at .05. The BAD case represents an example of testing experiments sequentially at the .05 level. Alpha is greatly inflated by the BAD strategy. As the number of subjects added per iteration between 9 and 27 increased from 1 to 9, the number of tests decreased from 19 at 1 per iteration to only 3 at 9 per iteration, so the inflation of alpha was directly related to the number of tests conducted. Clearly, an alteration of the upper criterion alone cannot constrain alpha near .05. The general pattern of the figure is representative of all combinations of upper and lower bounds tested.

certain narrow range without exceeding three significant digits in the lower bound is a formidable and maybe impossible task.

Comparison With Fixed-Criteria SSRs

Figure 1 gives an example of the estimated alpha determined this way for the 9/27 combination of lower/upper bounds across nine levels of *n added*. The results generalize to all other sets of lower and upper bounds. Alpha for the variable-criteria technique is held constant for all *n added* during the experiment. When simulations were analyzed using fixed-criteria for the lower and upper bounds of, say, .01 and .36 (as was recommended in COAST; Frick, 1998), .01 and .25 (as was recommended for the *t* test in CLAST; Botella et al., 2006), or .01 and .50 (for CLAST with ANOVA; Ximénez & Revuelta, 2007), the alpha was held below .05, as was advertised. However, as was recognized by Frick (1998), Botella et al., and Ximénez and Revuelta, these fixed-criteria techniques lose efficiency because of a serious reduction of alpha as the number of *n added* increases. If the null hypothesis is false, the first few subjects may be consumed just to bring the power back to the level of alpha.

Clearly, the problem of a deflation of alpha in this circumstance cannot be solved by varying the upper criterion alone. Holding alpha constant requires a combination of variable lower and upper criteria. The other curve in the figure, labeled "BAD," is for the inappropriate strategy, using .05 as a lower criterion in sequential testing at the .05 level.

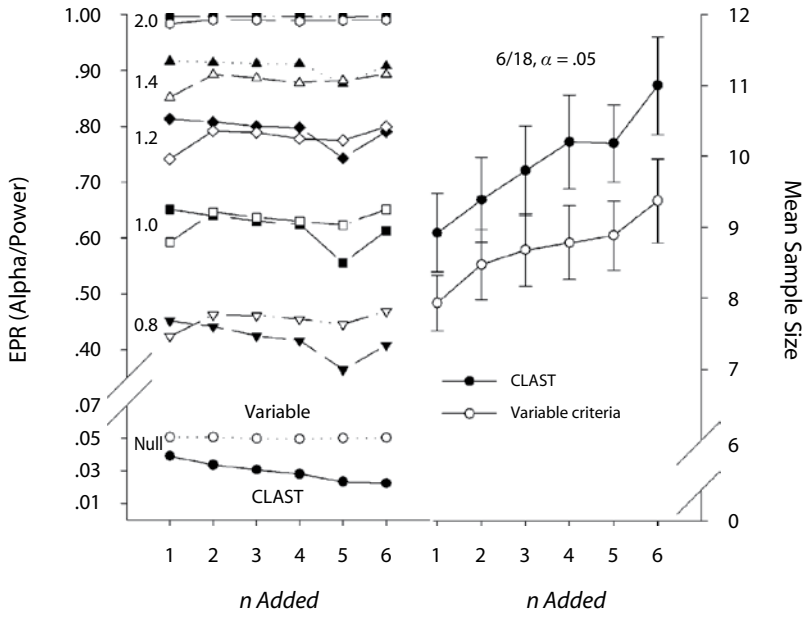

Figure 2. Comparison between the variable-criteria SSR and CLAST with a two-tailed alpha of .05. The CLAST test used the lower and upper criteria of .01 and .50 for an independent-groups ANOVA (Ximénez & Revuelta, 2007). According to the CLAST rule, sampling began at 6 subjects and terminated at 18 subjects per group. The different simulations included increments of sample size from 1 to 6 subjects per group per iteration (i.e., maxima of 13 separate tests at 1 subject added per iteration and 3 tests at 6 subjects added per iteration). Left panel: The variable-criteria SSR held alpha constant under the null effect condition and matched CLAST for power at seven levels of effects (effects 0.8–2.0; levels 1.6 and 1.8 omitted for clarity). Right panel: The mean of the sample sizes at the time of the rejection of the null hypothesis when averaged over the seven effect levels for the variable-criteria SSR and CLAST and displayed across six levels of *n added*. Both tests lost some efficiency with increased *n added* per iteration. The power of CLAST in the left panel came at the expense of additional sample size compared with the variable-criteria SSR. For sample size comparison, the fixed stopping rule with a standardized effect size of 1.2 requires 12 subjects per group for a two-tailed alpha of .05 and a power of ~0.8. Error bars represent ±1 standard error of the mean.

Figure 2 illustrates a direct comparison of the variable-criteria SSR with CLAST according to the rules for CLAST for ANOVAs (Ximénez & Revuelta, 2007). The CLAST rule requires that we begin testing with 50% fewer subjects and stop testing with 50% more subjects, so our lower and upper bounds are 6 and 18. The best criteria discovered for ANOVAs were .01 and .50, so we use those limits for CLAST (although the published experiment used four groups instead of two). The simulation was conducted 100,000 times for each of eight effect sizes from null through 2.0 and for levels of *n added* from 1 to 6 for a total of 48 independent simulations. As in Figure 1, the variable-criteria rule held the estimated alpha nearly constant at .05 when the null hypothesis was true, but, also as in Figure 1, the estimated alpha for CLAST was considerably less than .05 and diverged seriously at higher levels of *n added*. Other than that, the CLAST and variable-criteria SSRs had nearly the same observed power at all of the different effect sizes and *n added* level (the effect sizes 1.6 and 1.8 are omitted for clarity in the figure). The right side of the figure gives the mean and standard error of the mean of the sample size at the rejection of the null hypothesis for CLAST and for the variable-criteria SSR at all six levels of *n added*, averaged over all seven effect sizes (0.8–2.0). Although both tests lost some sample size efficiency at higher levels of *n added*, the variable-criteria SSR typically required fewer subjects than did CLAST. Thus, the equality of power between the two techniques in the left side of the figure came at the expense of slightly higher sample sizes in the CLAST case. This is a direct result of having an observed alpha much lower than the nominal alpha in CLAST. For comparison purposes, a power analysis for the fixed stopping rule with an anticipated standardized effect size, *d*, of 1.2 indicates that the optimal sample size for a two-tailed alpha of .05 and a power of .8 is 12 subjects per group. Both SSR techniques were considerably more efficient than the fixed stopping rule.

The level of observed power, or the EPR, for two different strategies in Figure 2 were fairly constant across different levels of *n added* for a given effect size, and this

was generally true for all of the fixed and variable SSRs, although there was often a slight increase in power with the largest levels of *n added*. In order to simplify the presentation, Figures 3, 4, and 5 present data that are averaged across all levels of *n added*.

Figure 3 displays the mean sample size at the rejection of the null hypothesis averaged across all levels of *n added* at each level of effect size tested for each of the simulations with the same 9/27 lower/upper bounds given in Figure 1. The curve labeled "Fixed *N*" in Figure 3 is the sample size required by the fixed stopping rule to produce the same power as was achieved by the variable-criteria strategy. The fixed *N* size was similar for all SSR tests because the powers of the tests were similar at various effect sizes, as is demonstrated for the 9/27 model in Figure 4. The principal message is that the average sample sizes when using either fixed or variable SSRs were less than or equal to those when using the fixed stopping rule for the same observed power. The best strategy in terms of sample size efficiency is the "BAD" test, but, as we have found, it gains its efficiency at the expense of a greatly inflated alpha. The next best is the variable-criteria SSR described here, followed by the fixed-criteria strategies of .01/.25, .01/.36, and .01/.50.

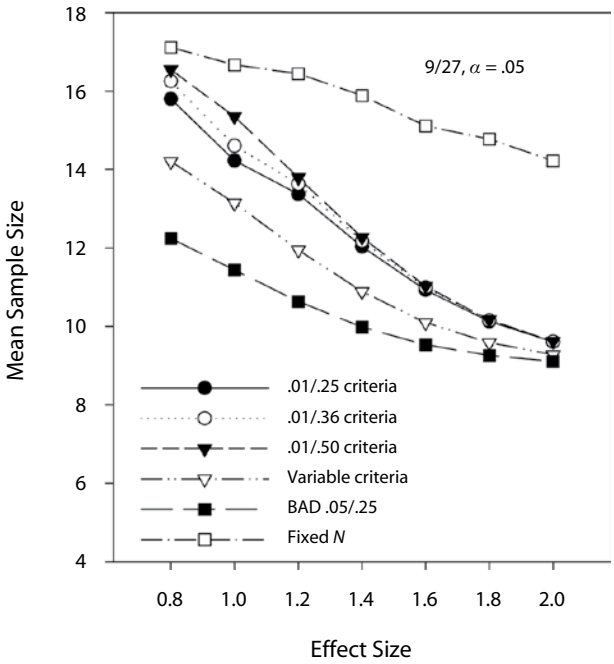

Figure 3. Mean sample size at the rejection of the null hypothesis averaged across all levels of *n added* at each level of effect size for each of the simulations with the 9/27 lower/upper bounds given in Figure 1. Sample size for a fixed-*N* power analysis at comparable power to the variable-criteria SSR is at the top. The variable-criteria SSR was more efficient in terms of sample size at all levels of effect size than the fixed-criteria SSRs. The BAD strategy was most efficient because of its hugely inflated alpha (see Figure 1). The results are representative of all of the combinations of lower/upper bounds in terms of the relative performance of the different strategies.

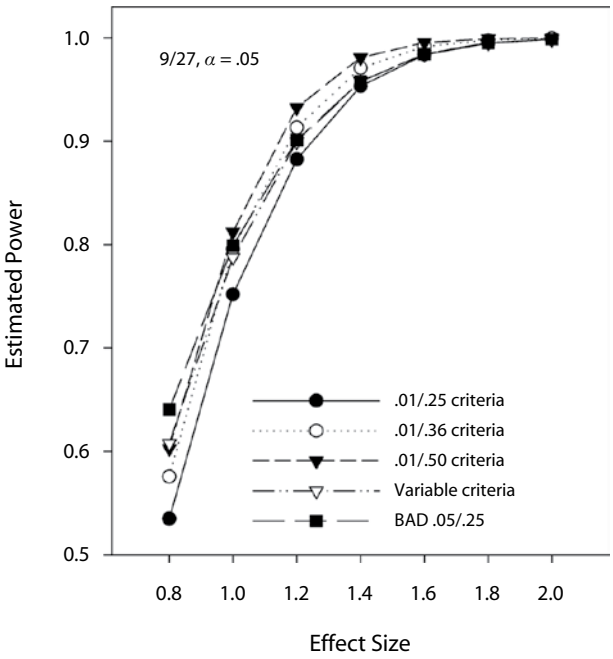

Figure 4. Estimated power averaged across all levels of *n added* at each level of effect size for each of the simulations with the 9/27 lower/upper bounds given in Figures 1 and 3. The different strategies are similar in terms of power across the different effect sizes. With the exception of the BAD test, all fixed-criteria SSRs required slightly larger sample sizes to achieve a similar level of power compared with the variable-criteria SSR (see Figure 3).

Sample Size Efficiency and Power

As was expected, the efficiency of SSRs for reducing the sample size required to produce significance at a given level of power is limited by the lower bound. At very low initial lower bounds (3 or 4), it becomes difficult to reduce the sample size from the fixed stopping rule (there are basement effects). Figure 5 displays the mean and standard deviation of the sample size at the time of the rejection of the null hypothesis in our variable-criteria SSR tests with all of the larger upper bounds, as well as the sample size that would be required in order to produce the same power with a fixed stopping rule. The SSR allows greater sample size efficiency with larger bounds and larger power (larger effect sizes). Nevertheless, the variable-criteria SSR was at least as efficient as the fixed stopping rule in all circumstances tested.

Figure 6 displays power curves for the two-tailed, independent-groups *t* tests for all tested sets of lower and upper bounds at the .005, .01, .05, and .10 levels of experimentwise significance. The smaller upper bounds are on the left side, and the larger upper bounds are on the right side. Although all of the other examples in this article have focused on the .05 level of experimentwise alpha, the general pattern of results holds very well for either the more conservative or the more liberal alphas. The power curves can be used to select models to produce significance with high sample-size efficiency relative to the fixed stopping rule. For example, if the anticipated effect size is 1.2 for

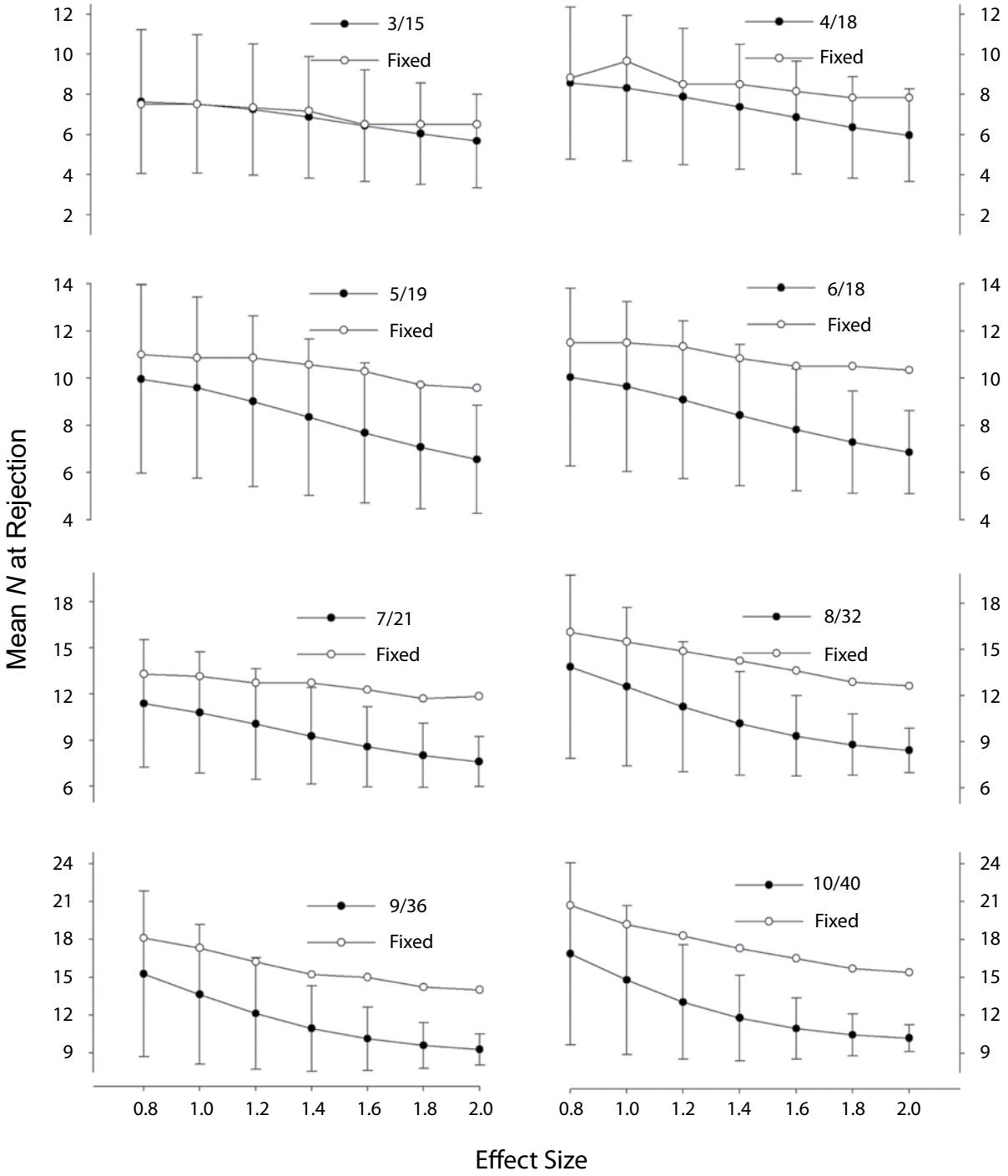

Figure 5. Means of the sample sizes at the time of the rejection of the null hypothesis in independent-samples, two-tailed *t* tests when using the variable-criteria SSR at several different lower/upper bounds, as well as the sample size that would be required to produce the same power with a fixed stopping rule (Fixed). The variable-criteria SSR allows greater sample size efficiency with larger bounds and larger effect sizes (i.e., larger powers). The variable-criteria SSR was at least as efficient as the fixed stopping rule in all circumstances tested. Error bars represent ±1 *SD*.

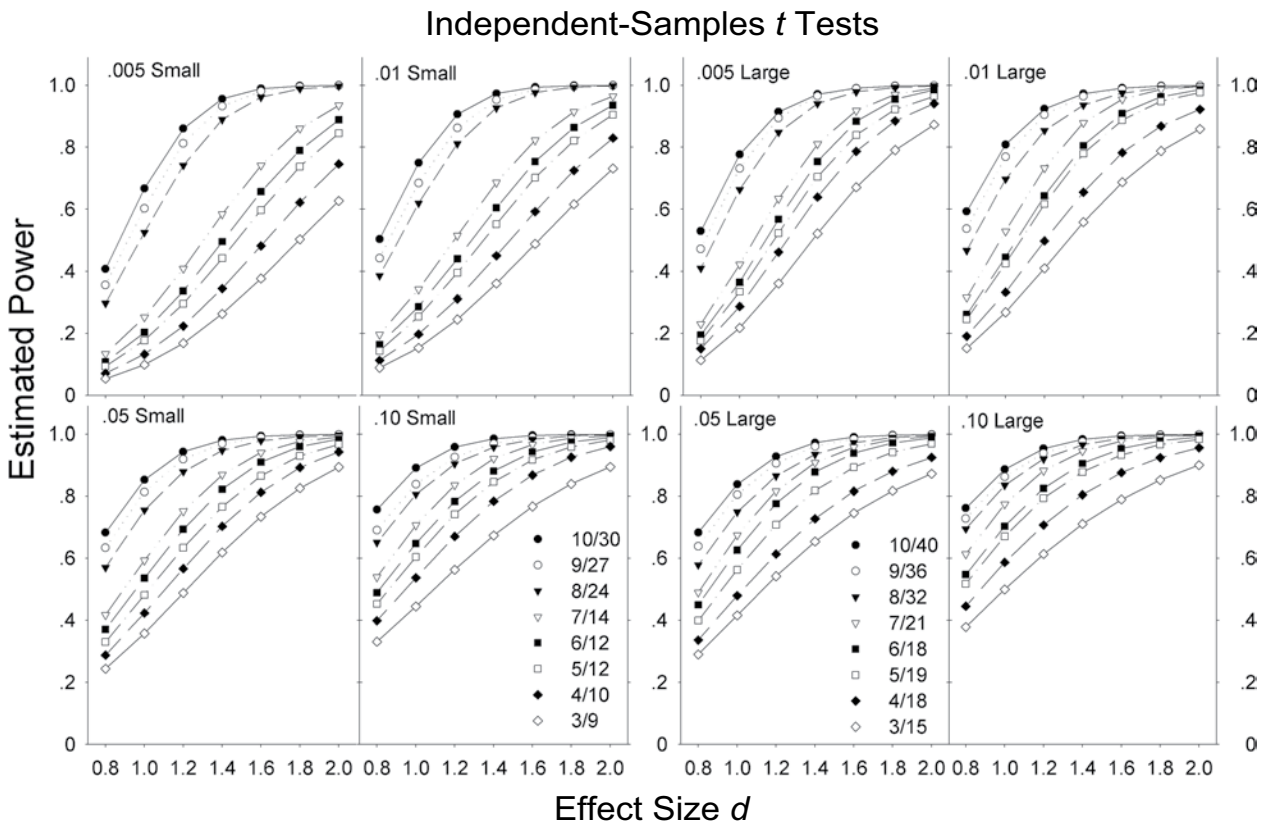

Figure 6. Power curves for independent-samples *t* tests for different levels of experimentwise alpha (.005, .01, .05, .10) at all lower bounds tested in combination with the smaller (left side) and larger (right side) of the two upper bounds for each. The different models are given as lower/upper bound (e.g., 10/30). Given a standardized effect size and a desired alpha and level of power, one can determine the models (combinations of bounds) that will yield at least that level of power in a two-tailed test. Each point is the average of 100,000 independent simulations at each level of *n added* for that model, so it is approximate for a given level of *n added*. The criteria to use for the selected model at various levels of *n added* are found in Table 2.

an experiment, the desired power is .80, and the desired experimentwise alpha is .01, we can look at the two .01 graphs for a power of .80 and observe that the 8/24 model, with the smaller upper bound, or the 8/32 model, with the larger upper bound, will provide about that level of power. We then select the *n added* per iteration for the experiment, identify the lower and upper criteria for that level of *n added* in Table 2, and then apply these rules consistently during the experiment. For the 8/32 strategy, for instance, the tabled lower and upper criteria are .0025 and .27 for an experimentwise alpha of .01 and an *n added* value of 3. Thus, beginning with a sample size of 8, adding 3 subjects per group per iteration to experiments for which a decision has not yet been made about the null hypothesis, and using no more than 32 subjects per group, the alpha will remain constant at .01 and the power of the overall experiment will be approximately .80. Space does not allow the printing of efficiency curves for all levels of significance, but the particular example cited would produce a significant result with a mean of 14–15 subjects per group compared with an *n* of 18 for similar power with the fixed stopping rule.

One-Tailed and Dependent-Sample *t* Tests

Because the criteria in Table 2 were derived from simulations using two-tailed, independent-samples *t* tests, it is of interest to know whether the same criteria can be used in one-tailed tests or in dependent-samples *t* tests or whether we must generate a separate Table 2 for each type of test. Figure 7 displays the estimated alphas from simulations with an effect size of zero (null hypothesis true) for dependent-samples or one-tailed tests using the criteria from Table 2. Alpha was estimated from 100,000 independent simulations of various sample-size models on the basis of criteria from Table 2. The two-tailed tests included all 16 models of lower and upper bounds represented in Table 2. The one-tailed tests included the following 8 models: 3/9, 4/18, 5/19, 6/18, 7/21, 8/24, 9/27, and 10/30. Thus, the simulations included independent-samples (between-subjects, B) and dependent-samples (within-subjects, W) *t* tests with either a one-tailed or a two-tailed hypothesis. The 16 new simulations for the two-tailed, independent-samples tests (B2tail) strikingly validate the criteria from Table 2 for the same type of test. The criteria from Table 2 also apply very well to a one-tailed test with independent samples (B1tail). When applied to within-subjects tests (W2tail, W1tail), there was a tiny but consistent increase in alpha. Because the error is very small, I decided against publishing a separate version of Table 2 for within-subjects tests. The criteria in Table 2 can be used with any of these types of *t* test, although the investigator should be aware of the small positive bias in alpha when Table 2 is used for within-subjects tests.

When conducting a one-tailed test, it is important to note that the experimenter should use the actual *p* from the one-tailed test and the criteria for the desired actual alpha from the table. Using the criteria for a two-tailed alpha of .10 will not yield correct results for a one-tailed test at the .05 level.

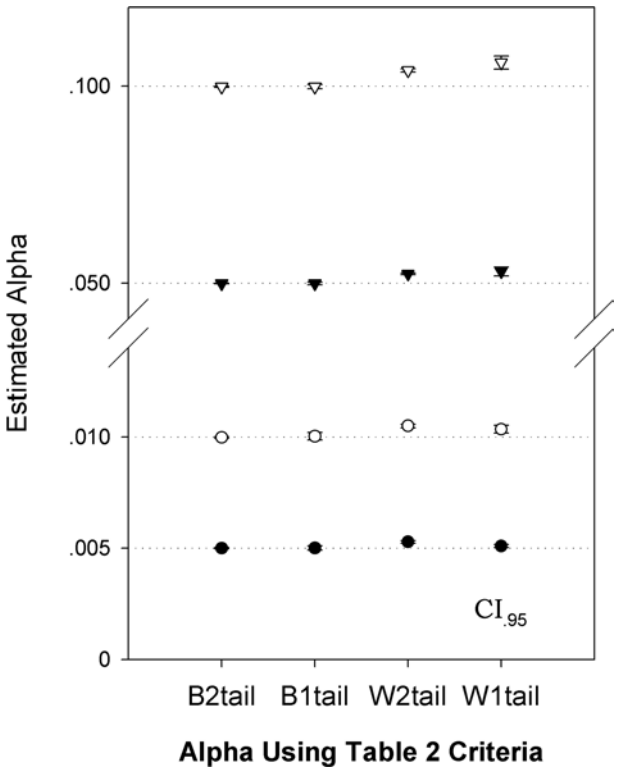

Figure 7. Estimation of alpha from simulations of dependent-samples or one-tailed *t* tests using criteria from Table 2. Alpha was estimated from 8–16 independent simulations of various sample-size models, and the data are displayed as the mean and a 95% confidence interval for the mean. The confidence interval in most cases is smaller than the height of the symbol. The two-tailed tests included all 16 models of lower and upper bounds represented in Table 2. The one-tailed tests included the following eight models: 3/9, 4/18, 5/19, 6/18, 7/21, 8/24, 9/27, and 10/30. The simulations were repeated for each model to include independent- (between-subjects, B) and dependent- (within-subjects, W) samples *t* tests with either a one-tailed or two-tailed hypothesis. The 16 new simulations for two-tailed, independent-samples *t* tests (B2tail) validate the criteria from Table 2. The criteria from Table 2 also apply well to a one-tailed test with independent samples (B1tail). When applied to within-subjects tests (W2tail, W1tail) there was a tiny but consistent increase in alpha. The criteria in Table 2 can be used with any of these types of *t* test, although the investigator should be aware of the small bias in alpha when using within-subjects or matched-sample tests.

After conducting the 16 new simulations for the null hypothesis with two-tailed, dependent-samples tests in Figure 8 using the criteria of Table 2, I also conducted simulations to determine the estimated power of the tests using the same criteria with the same effect sizes, *d*, used in the independent-samples experiments (i.e., 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, and 2.0). These power curves are displayed in Figure 8.

Skewed Distributions

In all of the preceding analyses, I have used data that were normal in the underlying distribution. The *t* test is known to be robust with respect to a violation of the assumption of normality (Boneau, 1960). That is, alpha

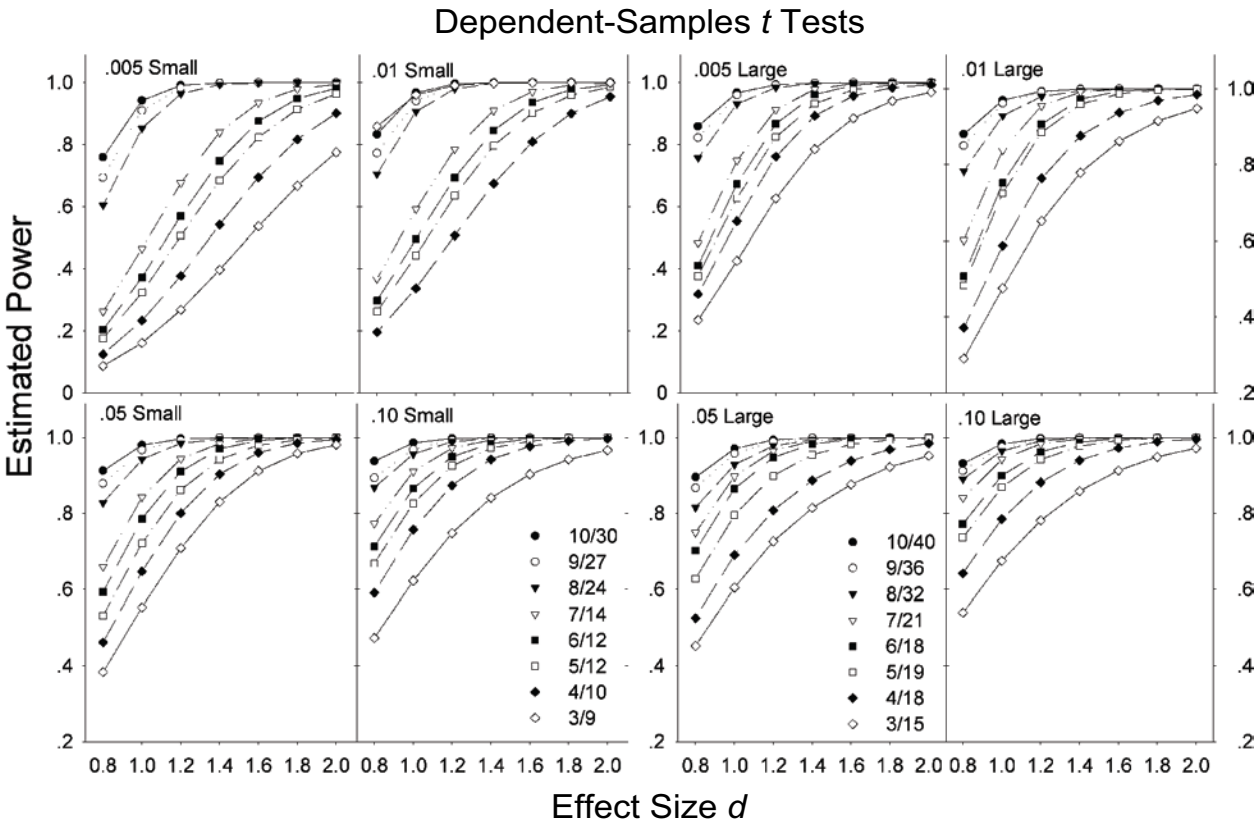

Figure 8. Power curves for dependent-samples *t* tests for different levels of experimentwise alpha (.005, .01, .05, .10) at all lower bounds tested in combination with the smaller (left side) and larger (right side) of the two upper bounds for each. The different models are given as lower/upper bound (e.g., 10/30). Given a standardized effect size and a desired alpha and level of power, one can determine the models (combinations of bounds) that will yield at least that level of power in a two-tailed *t* test. Each point is the average of 100,000 independent simulations at each level of *n added* for that model, so it is approximate for a given level of *n added*. The criteria to use for the selected model at various levels of *n added* are found in Table 2.

Table 3
Simulated Data From a Highly Positively Skewed Distribution Analyzed and Compared With the Normal Distribution Using the Lower and Upper Criteria for the .05 Level of Significance for the Model in Table 2 With Lower/Upper Bounds of 7/14

Effect	*n*	Normal			Skewed		
Size	*Added*	EPR	Mean *N*	*SD*	EPR	Mean *N*	*SD*
Null	1	.0486	8.7	2.0	.0575	8.7	2.0
	2	.0508	8.7	2.0	.0479	8.8	2.0
	3	.0497	9.1	2.3	.0479	9.1	2.3
	4	.0503	8.2	1.8	.0477	8.3	1.9
	5	.0498	8.6	2.3	.0495	8.7	2.4
	6	.0507	9.4	2.9	.0542	9.4	2.9
	7	.0510	9.7	3.4	.0476	9.7	3.4
1.0	1	.5763	8.7	2.0	.6135	8.8	2.1
	2	.5670	8.7	2.0	.5863	8.6	2.0
	3	.6038	9.2	2.4	.6190	9.2	2.4
	4	.5185	8.4	1.9	.5350	8.3	1.9
	5	.5502	8.9	2.4	.5815	8.9	2.4
	6	.6152	9.9	3.0	.6254	9.8	3.0
	7	.6465	10.6	3.5	.6520	10.4	3.5
1.4	1	.8528	8.1	1.7	.8720	8.1	1.7
	2	.8460	8.1	1.7	.8502	8.1	1.7
	3	.8791	8.5	2.1	.8819	8.5	2.1
	4	.8063	8.0	1.7	.8086	7.9	1.7
	5	.8303	8.4	2.2	.8460	8.3	2.2
	6	.8919	9.1	2.9	.8861	9.0	2.8
	7	.9074	9.6	3.4	.8964	9.4	3.3
1.8	1	.9685	7.5	1.1	.9716	7.5	1.2
	2	.9667	7.5	1.2	.9633	7.5	1.2
	3	.9812	7.8	1.6	.9790	7.7	1.6
	4	.9510	7.5	1.4	.9446	7.5	1.3
	5	.9615	7.7	1.7	.9622	7.7	1.8
	6	.9837	8.1	2.3	.9804	8.1	2.3
	7	.9884	8.3	2.7	.9823	8.3	2.7

Note—Neither the empirical proportion of rejections (EPR, an estimate of alpha or power) nor the sample size efficiency was markedly affected by the skewness. It is probably safe to use the variable-criteria SSR with skewed distributions if that is the only deviation from the assumptions of ANOVA.

and power remain relatively constant even when the data analyzed are from a highly skewed distribution. To test the effect of a highly skewed distribution on the behavior of the variable-criteria stopping rule, 100,000 tests were conducted with a skewed distribution at the .05 level of significance using the lower and upper criteria for the model with a lower bound of 7 and an upper bound of 14 (7/14). Only selected data are displayed in Table 3, to conserve space, but the data were consistent at all levels of effect size. The alpha and power of the procedure were only marginally affected by skewness of the samples. It is probably safe to use the variable-criteria SSR with skewed data as long as that is the only violation of the assumptions for the *t* test. Other violations, such as heterogeneous variances and unequal sample sizes, and combinations of violations, remain to be examined for this procedure. Nevertheless, equal or nearly equal sample sizes in each group are highly recommended.

One-Way ANOVA With Four Independent Groups

Figure 9 shows the empirical proportion of rejections when the null hypothesis was true for simulations of a one-way ANOVA with four independent groups for all 16 models of lower and upper bounds and all levels of *n added* listed in Table 2. The lower bound is on the abscissa. Above each lower bound, at each level of alpha, the results are plotted in side-by-side clusters of points with the smaller of the upper bounds on the left of the cluster and the larger upper bounds on the right of the cluster. All levels of *n added* are plotted for each model in Table 2, although the points are mostly printed on top of one another near the nominal alpha. Each individual point is the average of eight independent simulations with 100,000 experiments. The power is so great in this demonstration that a point is significantly different from the nominal alpha if the point does not touch the line. Thus, for some levels of *n added*, the observed alpha demonstrated a consistent bias away from alpha, particularly at the .10 level. Because the results were particularly stable and fairly accurate for the .05 and .005 levels, and because the level of control of alpha is far superior to that demonstrated with the COAST and CLAST alternatives, I have decided that separate tables like Table 2 for the ANOVA are not

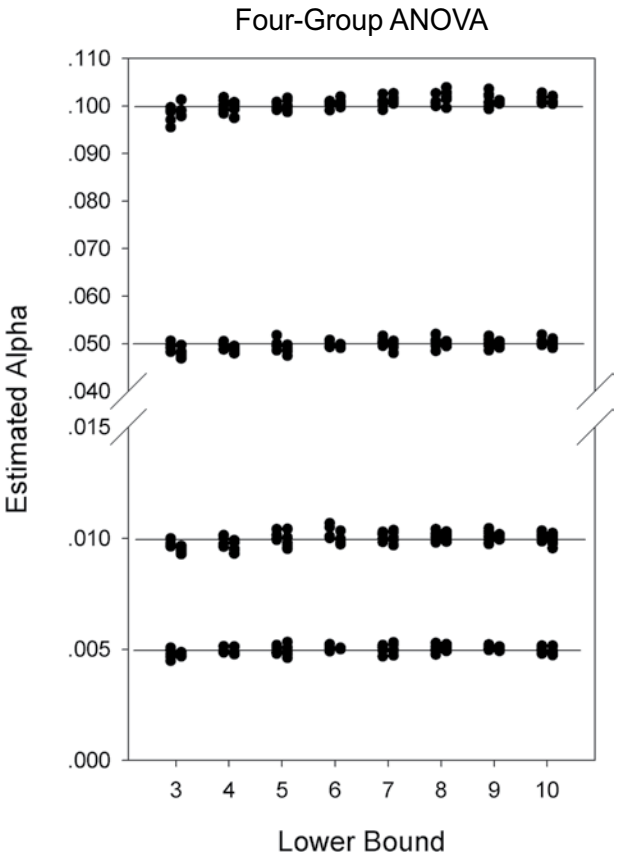

Figure 9. Alphas estimated from the empirical proportion of rejections when the null hypothesis was true in a four-group one-way ANOVA (all population means equal) using the criteria from Table 2. Each circle is the mean of eight simulations of 100,000 ANOVAs. The left half of each cluster is the smaller of the upper bounds, and the right half of each cluster is the larger of the upper bounds. Each column represents the proportion of rejections at the .005, .01, .05, or .10 level at all levels of *n added* displayed in Table 2 (from 6 to 10 points each). Note that the criteria in the table were based on a two-independent-groups *t* test, and the same criteria here generalize very well to a four-independent-groups ANOVA.

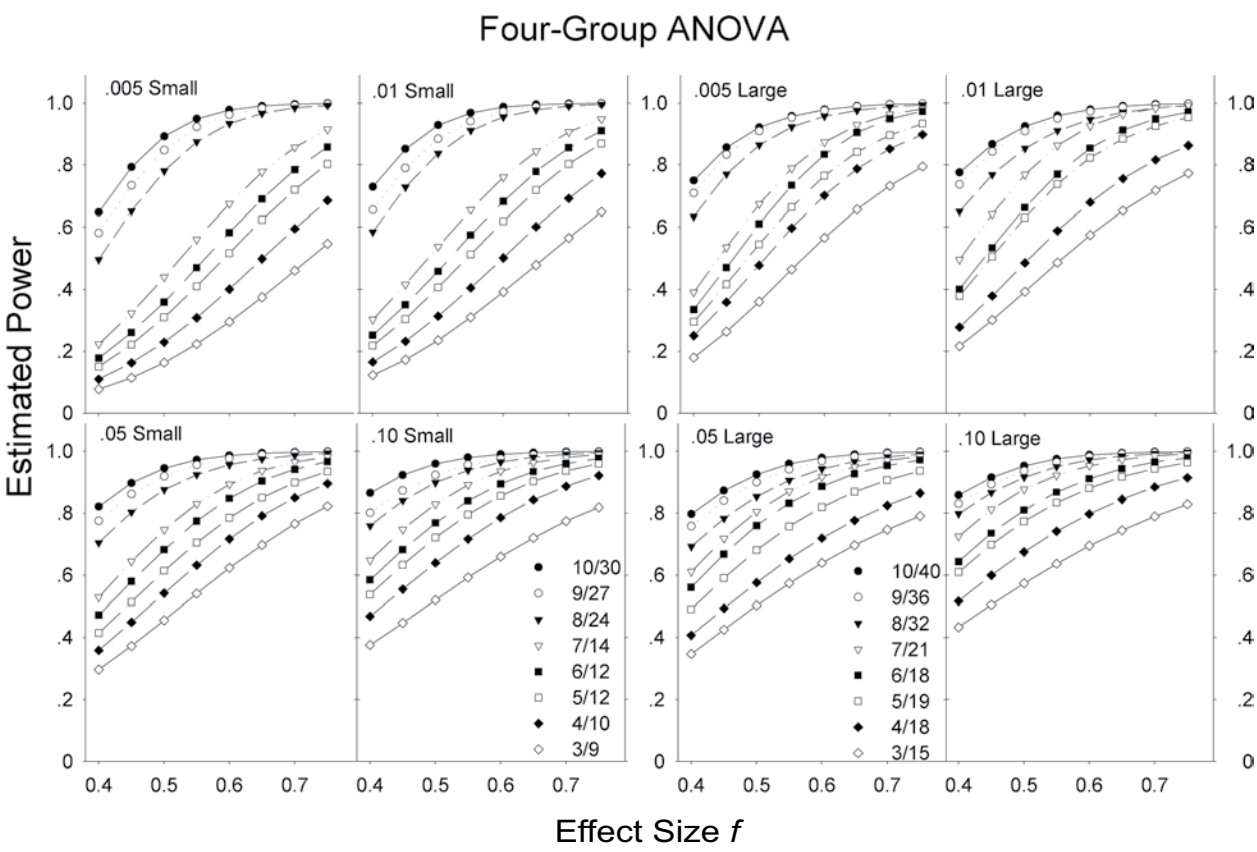

Figure 10. Power curves for a completely random four-group one-way ANOVA for different levels of experimentwise alpha (.005, .01, .05, .10) at all lower bounds tested in combination with the smaller (left side) and larger (right side) of the two upper bounds for each. The different models are given as lower/upper bound (e.g., 10/30). Given a standardized effect size and a desired alpha and level of power, one can determine the models (combinations of bounds) that will yield at least that level of power. Each point is the average of 100,000 independent simulations at each of 6–10 levels of *n added* for that model, so the power is only approximate for any given level of *n added*. The criteria to use for the selected model at various levels of *n added* are found in Table 2.

necessary. Future researchers with a more mathematical-theoretical approach may discover a way to solve these problems for exact criteria and could replace Table 2 with a single mathematical function that could then be made available for computers. In the meantime, it appears safe to use these methods with two to four independent groups of subjects on the basis of power curves in Figures 6, 8, and 10 and the criteria in Table 2. The power curves for the four-group ANOVA are displayed in Figure 10.

DISCUSSION

The fixed stopping rule for a null hypothesis test dominates research design and analysis in the biomedical and biobehavioral sciences. With the fixed stopping rule, an investigator conducts a power analysis and calculates the sample sizes for an experiment, and then the data are collected for all subjects and analyzed for better or worse.

In those same experiments, an SSR could produce significance on average with up to 30% fewer subjects than the fixed stopping rule. SSRs are efficient because experiments are conducted in iterations that allow stopping of the study when an obtained p is less than a lower criterion or greater than an upper criterion. If the initial power analysis is inaccurate, the SSR is forgiving, because the effects of a larger size can be detected and stopped with smaller sample sizes, and the effects of smaller size can be detected with larger sample sizes without inflating alpha.

Previous studies of the merits of SSRs used the same lower and upper criteria for all experiments of a type (t test, correlation, etc.), and this fixed-criteria strategy caused alpha to be drastically reduced and efficiency to be lost when adding more than 1 or 2 subjects per iteration. The present results with the variable criteria improve these SSRs by holding alpha very close to a target value of .005, .01, .05, or .10 when 1–10 subjects are added at each iteration. This variable-criteria SSR produces power equal to the fixed-criteria SSRs with even fewer subjects, because alpha—the lower limit of power when the null hypothesis is false—is not degraded in the variable-criteria SSR as it is with COAST or CLAST.

On average, the tests of significance using the proposed variable criteria will produce significance before the upper bound is reached as long as the actual effect size is similar to the predicted effect size. Thus, the upper bound is not a good predictor of the number of subjects that will be

used in a successful experiment. This upper bound will ordinarily be reached only if the effect size is very small or absent, or if, for reasons of bad luck (Type II error), the result remains in the *uncertain* zone throughout testing until the upper bound is reached. There is no requirement that the experimenter test all of the subjects all the way to the upper bound. As Frick (1998) pointed out, the primary consequence of stopping early is that the observed experimentwise alpha will be reduced.

In models with the same upper bounds and different lower bounds (e.g., 4/18 vs. 6/18 or 5/12 vs. 6/12) the observed power may be a bit higher when using the larger lower bound. This may be because initial tests with smaller power (lower n) can cause the investigator to retain the null hypothesis (i.e., commit a Type II error) in a few cases before there is sufficient power in the experiment to generate a significant result. Therefore, when given a choice between models with similar upper bounds, the model with the larger lower bound may be preferred.

In models with the same lower bounds and different upper bounds (e.g., 10/30 and 10/40) the model with the larger upper bound is always more powerful. However, if subjects are rare enough, or if the ethical consequences of the experiment are severe enough, it may be obvious at the beginning of the experiment that the number of subjects at the larger upper bound will never be used. Stopping very early in such conditions could cause alpha to be seriously reduced. For that reason, I have included two sets of upper bounds for each lower bound. Alpha will be held near the nominal level for all of the models, so investigators working with limited numbers of subjects will not be forced to use the criteria for the larger upper bound when it is obvious from the beginning that alpha will be reduced by early stopping.

One good way to determine the preferred upper bound is to estimate the size of the effect that would be implied by a power analysis with the number of subjects at the upper bound. If the size of effect is trivial or uninteresting, the experiment is overpowered, and the smaller upper bound should be selected.

Recommendations for Use

SSRs are appropriate for some but not all experimental circumstances of null hypothesis significance testing. They are good when the goal is to determine whether or not a treatment has a significant effect and when the investigator is satisfied with a *yes* or *no* answer to that question (Frick, 1998). The efficiency of the method allows an answer with considerably fewer subjects on average than the fixed stopping rule. SSRs would not be employed for experiments in which the goal of the experiment is to determine the size of an effect with the most accuracy, although it may be possible to develop such an application. The width of a confidence interval for an effect size in an experiment is determined in part by the sample size, and, all other factors being equal, larger sample sizes give smaller confidence intervals. Sample size planning for such experiments can include the techniques suggested by Maxwell, Kelley, and Rausch (2008). SSRs may nevertheless be useful in such research, especially when it is necessary to determine whether an effect exists before one devotes a large amount of resources to determine the size of the effect. SSRs are excellent for pilot studies.

SSRs are good for research in which the experiments can be conducted in stages by adding 1 or a few subjects at a time and when the results are known soon after the subjects are tested. The method is recommended in experiments in which subjects may be difficult to obtain or expensive, when the procedure requires considerable surgical or other time-consuming preparation, or when the use of a large number of subjects as a matter of course is ethically objectionable, such as in studies of pain. The method would not often be useful in experiments in which the results will not be known for months or years, because every iteration of the experiment would then require months or years. However, SSRs can be useful for relatively long experiments in laboratories that are capable of conducting multiple experiments at a time. Instead of devoting all of the laboratory's resources at once to a single large experiment using the fixed stopping rule, smaller replications of experiments can be planned to overlap, so that various iterations of several experiments are conducted simultaneously and the laboratory is continuously busy and productive. On average, all of these experiments will require fewer subjects and resources if they use the SSR approach.

Steps for Using the Variable-Criteria SSR

To use the variable-criteria SSR described in this article for an experiment involving a t test, one needs to follow several simple steps. As with a fixed stopping rule, the investigator must first select an alpha and the desired power for the entire experiment and estimate the size of the effect in standardized units. This is done using previous data or by establishing a minimum effect size that would be considered interesting or important. Many power analysis programs calculate this easily from provided means, standard deviations, and correlations. Then do the following:

1. Examine the power curves in Figure 6 for an independent-samples t test, in Figure 8 for a dependent-samples t test, or in Figure 10 for a one-way ANOVA for four groups, using the alpha, desired power, and anticipated effect size to determine the available models (combinations of lower and upper bounds) that are capable of producing the desired amount of power under the selected conditions.

2. Select any of these models on the basis of the needs and constraints of the experiment. Using larger lower bounds generally provides more power. One should use a lower bound that is large enough to convince reviewers and readers in the field that an effect is real if it is reported as significant, because one will have to stop testing if the p with that number of subjects is less than the lower criterion.

3. Decide how many subjects to add per iteration and look up the selected model in Table 2. This determines the lower and upper criteria to use in the experiment.

4. Test the number of subjects per group at the lower bound. If p is less than or equal to the lower criterion, stop testing and reject the null hypothesis at the selected experimentwise alpha. If p is greater than the upper criterion, stop testing and retain the null hypothesis. Otherwise, add the number of subjects that you previously determined in Step 3, and reanalyze with the augmented sample size. For independent-samples t tests, this number will be added to each group. Repeat this procedure until you have rejected the null hypothesis, retained the null hypothesis, or reached the upper bound. If adding *n added* subjects to the sample size would exceed the upper bound, and if the p value still in the *uncertain* region, one must retain the null hypothesis. There is not sufficient evidence (or power) to declare the result significant.

5. Sometimes an investigator might need to stop an experiment while the result is still in the *uncertain* region and the upper bound of sample size has not yet been attained. From the rule just stated, this result cannot be declared significant because the p at the end of the last iteration was not less than or equal to the lower criterion. The alpha in this circumstance could easily be estimated from the data in the simulations, and it will always be less than the nominal experimentwise alpha for the procedure. Stopping early without a significant result cannot inflate the Type I error rate, because a Type I error can be made only when the result is significant.

Note that alpha is inflated by the intention of the investigator to add additional subjects when the result is not quite significant rather than by the actual addition of subjects to an experiment when the result is not quite significant. If the investigator begins an experiment with the intention of using the variable-criteria SSR with an alpha of, say, .05, and finds that the resulting p is less than .01 after the first test at the lower bound, the result is significant at $p \leq .05$ (not .01; Frick, 1998). The final obtained p should not be reported without explanation, because it is no longer an accurate indication of the probability of an event as extreme or more extreme than the obtained result.

Planning for Replicability

If the p in an isolated experiment is actually very near .05, the investigator is wise to consider it an interesting but unconfirmed observation. As was noted by Greenwald, Gonzalez, Harris, and Guthrie (1996), an obtained t with a p of exactly .05 is the best estimate of the mean of all t values in all possible t tests with the same conditions and power. That means that one half of all exact replications would be expected to have larger ts (i.e., significant), and the other half would be expected to have smaller ts (not significant). Investigators generally want better replicability than that, so they repeat the experiment once or twice to confirm that the results will have a good chance of being repeated by other investigators.

It is difficult to know in a rational way how many replications of an experiment are necessary and how many are a waste of resources or a needless duplication. If the experiments involve pain in animal or human subjects, for example, it is important to have a rational basis for deciding that an experiment is likely to be replicated so the testing can be limited to the fewest subjects necessary. Greenwald et al. (1996) demonstrated that a p value of .005 provides an 80% chance that an exact replication will be significant at the .05 level, and 80% is a value that many investigators consider to be acceptable power. Therefore, instead of conducting multiple independent replications of an experiment with an alpha of .05, a single experiment can be conducted with an experimentwise alpha of .005 using the variable-criteria SSR described here to improve efficiency and save subjects. Because each iteration of the SSR is conducted independently, it is similar to an independent replication (and can be conducted with different batches of animals or drugs or on different days with different experimenters). Differences between iterations can even be removed if the *n added* is large enough by considering the iterations as blocks in an ANOVA. If the null hypothesis is rejected with a p less than or equal to the lower criterion for the selected model with an experimentwise alpha of .005, one can consider the experiment to have at least an 80% chance of being repeated with a significant finding at the .05 level in an exact replication with identical power if the observed effect is equal to the true effect.

Using an alpha of .005 under the fixed stopping rule would also indicate that the results were replicable. However, it is wasteful of resources to use an alpha of .005 routinely with a fixed stopping rule in experiments in which the treatments may turn out to have little or no effect. For example, the sample size for significance at the .005 level for a fixed stopping rule with an anticipated standardized effect size of 1.0 is 29 per group. That is a lot of subjects for a biomedical or biobehavioral researcher to invest in determining that a particular avenue of research is not fruitful. However, by conducting the experiment using the variable-criteria SSR, the experiment can be conducted efficiently in stages, without affecting alpha. Experiments with very small or no effects will be stopped on average long before 29 subjects per group are tested, and most experiments in which a true effect of ~1.0 exists will use fewer than 29 subjects per group to reject the null hypothesis at the .005 level. For example, the variable-criteria SSR using the 10/40 model with 6 added subjects per iteration provides ~80% power to reject the null hypothesis at the .005 level and requires on average 24 subjects per group. The lower and upper criteria in Table 2 for this model are .0014 and .220. This means that if the true effect size is zero, 100% − 22% = 78% of the trials would be stopped after only a single iteration with 10 subjects.

An investigator wishing to demonstrate repeatability of a result would likely plan two replications of an experiment, each at the .05 level, with the second experiment contingent on a significant result in the first. The sample size for a single replication of the experiment just illustrated would require 17 subjects per group with the fixed stopping rule and an alpha of .05. Conducting two independent replications at the .05 level would require 34 subjects per group. As was mentioned above, the same goal can be achieved with 24 subjects per group on average using a variable-criteria SSR and an alpha of .005.

Another method to address replicability is to use the obtained p from an experiment to estimate the probabil-

ity that an exact replication, using the same conditions and power, will have a result in the same direction (i.e., a positive or negative mean difference, not necessarily significant; Killeen, 2005). Planning an experiment using the variable-criteria SSR and an alpha of .01 provides about a 97% chance that an exact replication will produce a result in the same direction (Cumming, 2005).

Should We Be Doing It This Way?

Most investigators who use null hypothesis significance tests would be surprised to learn that their workhorse statistical routines have been roundly and thoroughly criticized, to the point that many enlightened statisticians have called for a replacement of the technique altogether (e.g., Goodman, 1999; Greenwald et al., 1996; Killeen, 2005, 2006; Loftus, 1996; Meehl, 1967). Few statisticians have defended the method, yet it persists as the dominant paradigm for analysis of biomedical data. Some have offered interesting explanations for why it remains popular (Frick, 1996; Greenwald et al., 1996).

Statisticians critical of the null hypothesis test will wonder why I propose to help investigators do something wrong more efficiently instead of re-educating investigators to do it correctly. The fact is that most biomedical scientists have a strong background in the content of their area but not in the techniques for analyzing that content. In an informal analysis of the requirements for a few PhD-granting departments and programs at my institution that frequently use animals in research, no graduate-level class in statistics is currently required by the Pharmacology, Immunology, Physiology and Biophysics, Neurobiology and Behavior, Biology, or Speech and Hearing Sciences programs; one class is required by the Bioengineering, Pathology, Oral Biology, and Aquatic and Fishery Sciences programs; and two graduate classes and two statistical computing laboratories are required by the Psychology program. This may account for the increased awareness of the problem in psychology journals (see the References section). The remaining PhD students must pick up their statistical knowledge from a combination of old undergraduate classes, electives (if they took them), biostatistics consultants, and their mentors. The only mention of the word *statistics* in a curricular context in the handbook for our medical school's MD program occurs in the description of a class on clinical epidemiology and evidence-based medicine, which is taught partly in a journal club format. Excellent biomedical statisticians do exist, but they do not represent the average biomedical scientist whose grant applications and IACUC protocols I read every day. Re-educating the workforce about alternative methodologies will require more than one journal article, as is demonstrated by the many that have already had little effect. However, a simple change in the usual statistical technique may lead to a savings of thousands of research animals every year.

One of the criticisms of null hypothesis testing is that large, overpowered experiments can detect significant, but unimportant, effects. The proper use of the variable-criteria SSR encourages investigators to seek effects of a meaningful size without overpowering the experiment and wasting subjects. Investigators should become more aware of Bayesian and other alternative methods for the analysis of laboratory data, report confidence intervals for effect sizes, include estimates of population standard deviations in publications to assist meta-analyses, and avoid the misconceptions and misinterpretations of null hypothesis significance test as outlined by many authors (reviews: Cumming, 2005; Frick, 1996; Goodman, 1999; Greenwald et al., 1996; Killeen, 2005, 2006; Loftus, 1996; Meehl, 1967).

Recommendations for IACUCs and Other Regulatory Agencies

Regulatory agencies such as the United States Department of Agriculture (which enforces the Animal Welfare Act), the Office of Laboratory Animal Welfare at the National Institutes of Health, and the Association for the Assessment and Accreditation of Laboratory Animal Care International all require a rational basis for the determination of the number of animals that are requested and used by principal investigators on IACUC protocols. A power analysis with a fixed stopping rule is often suggested as one way to meet this goal.

SSRs are useful augmentations for many of these tests, because they can be much more efficient and result in the use of fewer humans or animals in research without a loss of statistical power or an inflation of alpha. A sequential approach is intuitive and gives investigators a range of sample sizes to work with in order to determine in an iterative fashion whether an avenue of research is worth pursuing.

In the justification for the number of animals to be used in an experiment, the investigator should report the model for the variable-criteria SSR that best fits the anticipated results. The IACUC should then approve the number of animals for the experiment at the upper bound. Experiments in which the null hypothesis is true will tend to be stopped early, before that many subjects have been tested. Also, the investigator will not be penalized for being imperfectly precognizant of the size of the effect. The only circumstance in which an investigator changing to a valid SSR will use more subjects on average than with the previous method is when the previous method is the BAD and inappropriate strategy of testing multiple sequential tests at the .05 level to determine significance (see Figure 1).

AUTHOR NOTE

I thank Geoffrey R. Loftus and Anthony G. Greenwald for helpful and encouraging comments on an early draft of the manuscript. I thank Nona K. Phillips, Virginia G. Batterson, and the rest of the Office of Animal Welfare for their tolerance and forbearance during the creation of this research. Correspondence concerning this article should be addressed to D. A. Fitts, Office of Animal Welfare and IACUC, University of Washington, Box 357160, Seattle, WA 98195 (e-mail: dfitts@u.washington.edu).

REFERENCES

BONEAU, C. A. (1960). The effects of violations of assumptions underlying the *t* test. *Psychological Bulletin*, **57**, 49-64. doi:10.1037/h0041412

BOTELLA, J., XIMÉNEZ, C., REVUELTA, J., & SUERO, M. (2006). Opti-

mization of sample size in controlled experiments: The CLAST rule. *Behavior Research Methods, Instruments, & Computers*, **38**, 65-76.

BRETZ, F., KOENIG, F., BRANNATH, W., GLIMM, E., & POSCH, M. (2009). Adaptive designs for confirmatory clinical trials. *Statistics in Medicine*, **28**, 1181-1217. doi:10.1002/sim.3538

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

CUMMING, G. (2005). Understanding the average probability of replication: Comment on Killeen (2005). *Psychological Science*, **16**, 1002-1004. doi:10.1111/j.1467-9280.2005.01650.x

FRICK, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, **1**, 379-390. doi:10.1037/1082-989X.1.4.379

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

GOODMAN, S. N. (1999). Toward evidence-based medical statistics. 1: The *P* value fallacy. *Annals of Internal Medicine*, **130**, 995-1004.

GREENWALD, A. G., GONZALEZ, R., HARRIS, R. J., & GUTHRIE, D. (1996). Effect sizes and *p* values: What should be reported and what should be replicated? *Psychophysiology*, **22**, 175-183. doi:10.1111/j.1469-8986.1996.tb02121.x

KILLEEN, P. R. (2005). An alternative to null-hypothesis significance tests. *Psychological Science*, **16**, 345-353.

KILLEEN, P. R. (2006). Beyond statistical inference: A decision theory for science. *Psychonomic Bulletin & Review*, **13**, 549-562.

LOFTUS, G. R. (1996). Psychology will be a much better science when we change the way we analyze data. *Current Directions in Psychological Science*, **5**, 161-171.

MAXWELL, S. E., KELLEY, K., & RAUSCH, J. R. (2008). Sample size planning for statistical power and accuracy in parameter estimation. *Annual Review of Psychology*, **59**, 537-563. doi:10.1146/annurev.psych.59.103006.093735

MEEHL, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science*, **34**, 103-115.

PRESS, W. H., TEUKOLSKY, S. A., VETTERLING, W. T., & FLANNERY, B. P. (1992). *Numerical recipes in C: The art of scientific computing* (2nd ed.). Cambridge: Cambridge University Press.

TIMMESFELD, N., SCHÄFER, H., & MÜLLER, H.-H. (2007). Increasing the sample size during clinical trials with *t*-distributed test statistics without inflating the Type I error rate. *Statistics in Medicine*, **26**, 2449-2464.

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

XIMÉNEZ, C., & REVUELTA, J. (2007). Extending the CLAST sequential rule to one-way ANOVA under group sampling *Behavior Research Methods, Instruments, & Computers*, **39**, 86-100.

(Manuscript received June 10, 2009;
revision accepted for publication August 19, 2009.)