Comparative Medicine
Copyright 2011
by the American Association for Laboratory Animal Science

Vol 61, No 3
June 2011
Pages 206–218

Minimizing Animal Numbers: The Variable-Criteria Sequential Stopping Rule

Douglas A Fitts

The variable-criteria sequential stopping rule (SSR) allows an investigator to use a few subjects at a time to determine whether a planned experiment is worth pursuing without increasing the rate of false discoveries (type I errors). The SSR is appropriate whenever testing a null hypothesis if the experiment can be conducted in stages. The investigator adds a predetermined number of subjects at each stage and tests repeatedly for significance until the experiment is stopped because: (1) a significant effect is detected; (2) the effect is clearly not going to be significant; or (3) the predetermined maximal number of subjects has been reached. Two crucial features of the SSR are that it holds the probability of a type I error constant and maintains excellent power. The method is more efficient than is performing a typical significance test after a power analysis because SSR can require 30% fewer subjects to achieve the same power. The variable-criteria SSR provides a formal method for assuring the use of a minimal number of animals. This article provides practical examples of how to use the SSR in combination with a *t* test, one-way ANOVA, one-way ANOVA with a planned contrast as the focus of the stopping rule, or, in limited circumstances, multifactorial ANOVA.

Abbreviation: SSR, sequential stopping rule.

The numbers of animals approved for scientific experiments must be justified and reasonable. The *IACUC Guidebook*[16] includes power analyses and pilot studies among the methods to achieve this end. A companion article[9] showed how use of power analysis or a pilot study with a traditional null-hypothesis test for animal studies can be inefficient under some circumstances. This traditional test is called the 'fixed stopping rule' because one must stop testing after a fixed number of subjects. By contrast, some statistical methods allow investigators to test sequentially, with numerous interim 'peeks' at the data, to control errors and increase efficiency. A sequential stopping rule (SSR) allows one to conduct an experiment in stages in order to minimize the number of subjects used (see the section *Other Techniques for Sequential or Interim Analysis* for additional discussion). An SSR is powerful, flexible, and usually more efficient for animal subjects than is the fixed stopping rule. The variable-criteria SSR[7,8] is an improvement of previous methods[3,11,21] that holds the rate of false discoveries (type I errors) stable while providing excellent power. The method helps an investigator to ensure that no more than the minimal number of animals will be used and that experiments are repeatable as well as significant without wasting animals. The SSR is particularly useful when little prior information is available for conducting a power analysis under novel experimental circumstances. It legitimately allows an investigator to invest minimal resources in a new project until it is fairly certain that the experiment could succeed in its goals. This ability saves money, resources, and animals.

Some experiments are not suitable for use with an SSR. The SSR is intended for use with a null hypothesis significance test, which is appropriate when an investigator is interested in whether a difference or relationship exists and, if so, in which direction.[10,11] SSR are not helpful for determining the size of an effect with a narrow confidence interval because, to save animals, the method stops the experiment while the confidence interval is still very wide (just significant). For details of these and other limitations of the SSR, see the section *When to Use the Variable-Criteria SSR*.

This article provides guidance for biomedical researchers on how to use the variable-criteria SSR in several common, practical experimental circumstances. It describes how and when to use SSR and provides concrete examples of experiments using a *t* test, one-way ANOVA, ANOVA with planned comparisons, and multifactorial ANOVA. Finally, the article provides advice to principal investigators and IACUC on how to incorporate the variable-criteria SSR into the justification of the number of animals used in IACUC protocols.

How to Use the Variable-Criteria SSR

Requirements and terminology. The variable-criteria SSR is a set of rules for using a null hypothesis significance test in a particular way. In most cases, using the SSR will require no additional software than what is available on the typical laboratory computer. In simplest terms, the investigator begins with a few subjects in the test and then decides whether to add additional subjects or stop the test. The test can be stopped if the experiment is significant (*P* less than or equal to a lower criterion), if the experiment is quite unlikely to succeed (*P* is greater than an upper criterion), or if the experiment has used up all of the assigned subjects. If the *P* value is between the lower and upper criteria and if adding subjects will not exceed the maximum, the investigator may add a set number of subjects to each group and test again. Some investigators use a variation of this technique inappropriately by always using 0.05 as the lower criterion and by having no determined value for the upper criterion. This use is inappropriate because it increases the

Received: 20 Jul 2010. Revision requested: 24 Sep 2010. Accepted: 26 Dec 2010.
Office of Animal Welfare, University of Washington, Seattle, Washington.
Email: dfitts@u.washington.edu

rate of type I errors in the experiment.[9,11] The variable-criteria SSR provides criteria suited to the particular experimental design so that the rate of type I errors does not increase.[7,10]

A full-featured statistics program is recommended but not required. A few necessary or highly desirable computer programs or documents (described following) will be of great help. At a minimum, investigators must be able to determine the *P* value to several decimal digits. A program or table that simply reports whether the *P* is less than 0.05 will not be adequate. Calculators are available on the Internet[5] that report a *P* value to several digits given an obtained statistic and its degrees of freedom (df). Free calculators to determine power and sample size are also available on the Web, although some are better than others. One suggested calculator is G*Power.[6]

A user of the variable-criteria SSR needs to acquire a publication[7] that contains a table of stopping criteria and power curves for the significance test that will be used in the planned experiment. This table of stopping criteria can be used with either an independent- or dependent-samples *t* test with either 1- or 2-tailed *P* values. [7] The criteria were validated by using independent-groups ANOVA with as many as 20 groups[8] without any substantial drift of type I errors from the nominal level (alphas are provided for 0.005, 0.01, 0.05, and 0.10 levels). The author also can provide tables for selected nonparametric tests including the Mann–Whitney–Wilcoxon rank test, Wilcoxon signed-rank test, and Kruskal–Wallis test for multiple independent groups. Using nonparametric tests with the criteria from the ANOVA table will severely deflate alpha and power.

Once the appropriate table has been identified, additional decisions are necessary before an investigator can select the correct criteria from the table. These are a level of α (0.005, 0.01, 0.05, or 0.10), a sample size model, and the number to be added at each stage of the experiment (the '*n* added'). Each combination of these factors has been tested with simulations to produce a unique pair of criteria, the lower criterion and the upper criterion, that will hold the rate of type I errors stable at the nominal alpha for the entire test. Therefore, an understanding of these terms with respect to the variable-criteria SSR is desirable. Reference 9 provides a brief review of the null hypothesis significance testing method and the importance of controlling type I errors during sequential testing.

Alpha is the proportion of the time that the investigator will conclude that an experiment is significant when it is not. This is called a type I error, and it can be encountered only if the null hypothesis is actually true in the population. A conventional level of α is chosen by the investigator at the beginning of the experiment to be small, such as 0.05, so that these errors will be rare. In computer simulations, the programmer distinguishes between this nominal alpha set by the investigator/programmer and the observed alpha. Population parameters are known during the simulation, so the programmer knows whether a *t* test gives a correct decision or an error. The observed alpha in these simulations is the actual rate of type I errors after many thousands of tests when the population means are equal. Under conditions of sequential sampling with a criterion for significance set at the nominal level of alpha (for example, 0.05), the observed alpha can inflate to more than double that value. These simulations are used in the variable-criteria SSR to determine criteria that allow the observed alpha for the entire sequential procedure to be equal to the nominal alpha after inflation. These criteria vary with the sample size (therefore, 'variable criteria') and are included in a table.[7] If the *P* value obtained from the significance test is less than or equal to the lower criterion, one can conclude that the alternative hypothesis is more likely than the null hypothesis as an explanation for the data and that the effect is significant. When declaring significance in an experiment, it is important to know and state the rate of type I errors. Investigators who are particularly concerned about not publishing type I errors may want to use a smaller α, such as 0.01 or 0.005. These will require larger sample sizes to achieve the same level of power as a test at the 0.05 level.

The sample size model consists of a starting sample size per group, called the lower bound, and a sample size per group that will not be exceeded in the experiment, called the upper bound. The stopping criteria in a variable-criteria SSR must be determined individually by using hundreds of thousands of computer simulations for each sample size model, so only a few sample size models are available. Lower bounds are available from 3 to 10 subjects per group. For each lower bound, there is a choice of 2 upper bounds, and the largest available upper bound is 40. Few experiments in biomedical sciences exceed 40 subjects per group, so the range is appropriate for many studies in many fields. The investigator selects the model that most closely fits the needs of the experiment. As with the fixed stopping rule, larger sample size models provide greater power in the experiment. Therefore, knowing the power of the model can be helpful in deciding which model to use.

Power curves have been published[7,8] to assist in the selection of a sample size model to use for the SSR with a *t* test or ANOVA. The use of these power curves is not essential as long as the investigator has a good idea of some starting and stopping sample sizes that will bracket the ideal sample size with the desired amount of power. Unlike the fixed stopping rule, the variable-criteria SSR begins with a sample size that provides modest power and progresses to sample sizes that have strong power. The investigator stops the experiment with the smallest sample size that yields significance in the null hypothesis test. Advice on how to use the curves is provided in the examples that follow.

The power curves rely on a standardized size of effect for the *t* test or ANOVA.[4] These standardized effect sizes can easily be calculated from sample size software.[6] Sometimes means and standard deviations are not available in prior publications, and for these circumstances, formulas are available for calculating effect sizes from other types of statistics such as the *t* or *F* value and the sample size.[19]

When the null hypothesis is that the difference in the population is 0 (rather than some specific nonzero value), the calculation of the effect size for an independent groups *t* test, *d*, is simply the difference between the means divided by a pooled estimate of the population standard deviation.[4] For a matched samples *t* test, *d* is calculated as the mean of the difference scores divided by the standard deviation of the difference scores.[4] If the difference scores are not available, the SD of the differences can be calculated from a formula using the SD of the 2 sets of scores and the Pearson product–moment correlation coefficient, *r*, between the 2 sets of scores. For ANOVA involving multiple groups, Cohen[4] devised a different measure of effect size, *f*, that is a ratio of the amount of spread among the various means derived from the between-groups mean square and an estimate of the pooled within-sample SD of the scores derived from the error mean square. These effects

sizes *d* or *f* can be calculated easily by software such as G*Power,[6] given input of means and SD.

When using power calculation tools to assist with the estimation of effect sizes, one should not be misled by effect-size conventions such as 'small,' 'medium,' and 'large.' These conventions were intended to apply principally to small effects in the social sciences.[4] By comparison, many effects in biological sciences are 'huge.'

The published power curves for variable-criteria SSR[7,8] plot the power of 16 separate sample size models at various effect sizes that are most relevant to animal research projects in the biomedical and biobehavioral sciences. Separate curves are available at 4 levels of α. The investigator first estimates the minimal size of effect that would be of interest to the project and then selects the level of α to choose the correct set of curves. By inspecting the graph with the effect size on the abscissa and the power on the ordinate, one can select the nearest sample size model that provides at least the requisite level of power (for example, 0.80). The upper bound should be selected so that it is near the upper limit of the number of subjects that one would be willing to invest in such a project.

The *n* added is the number of subjects that the investigator adds to each group at each sequential stage of the testing process. The sample size for each group on the first test is the lower bound of the sample size model. One then adds *n* added subjects at each stage until the experiment is stopped by one of the criteria. The *n* added will be selected based on its convenience for the investigator. Power varies only slightly at various levels of *n* added for a given sample size model, so power is not an important consideration. By selecting an *n* added of 1, the investigator can test significance after the addition of each individual animal per group. If one has sufficient apparatus to test 4 animals per group at a time, then *n* added can be set at 4. If one desires to use *n* added as a replicate factor (see following), one can set *n* added equal to the lower bound to have equal numbers in each replication. For example, to use the 6/18 sample size model with an *n* added of 6, one would begin with 6 per group (the lower bound), test again with 12 per group, and stop after testing 18 per group (the upper bound). Note that some selections of *n* added may not allow the investigator to test all the way to the upper bound; for example, for the 6/18 model with an *n* added of 5, one would be able to test with 6, 11, and 16 subjects per group. One would stop at that point because the addition of 5 subjects would exceed the upper bound of 18.

The stopping criteria are selected from a table (for example, Table 2 of Fitts[7] for a *t* test or ANOVA) for the desired α once the lower and upper bounds of the sample size model are known and the *n* added has been chosen. These 2 criteria are probability values (similar to α), and they are compared with the obtained *P* value to determine whether to stop the experiment. The lower criterion determines when to stop the experiment because the effect is significant, and this criterion will be less than α. When α is 0.05 for a *t* test, for example, this lower criterion will range from 0.013 to 0.034, depending on the sample size model. The upper criterion determines when to stop the experiment because it is unlikely to become significant with the addition of further subjects. This upper criterion can range from 0.150 to 0.460 for an α of 0.05. Reference 7 explains how these criteria were derived by computer simulations and why they are so variable.

Steps for using the variable-criteria SSR. Selection of the sample size model, α, *n* added, and stopping criteria are all a part of the experimental design that is completed before any subjects are tested in the experiment. To begin testing subjects, the investigator randomly samples subjects into each group of the experimental design, with the sample size equal to the lower bound of the sample size model. For example, if the sample size model is 7/28 (lower/upper bound), the first test will be conducted with 7 subjects per group. After the data are collected for these subjects, a null hypothesis test such as a *t* test is conducted, and a *P* value is obtained.

The obtained *P* value is compared with the lower and upper stopping criteria, and there are 3 possible outcomes. In the first outcome, *P* is in the rejection region. If *P* is less than or equal to the lower criterion, the null hypothesis is rejected at the selected level of α, and the experiment is stopped with a significant result. In the second outcome, *P* is in the upper stopping region. If *P* is greater than the upper criterion, the experiment is stopped with a nonsignificant result, and the null hypothesis is retained. In the third outcome, *P* is in the uncertain region. If *P* is greater than the lower criterion but less than or equal to the upper criterion, the decision is uncertain. One can add *n* added subjects to each group by random sampling if the addition of these subjects would not exceed the upper bound of the sample size model. Then, data are collected on these new subjects and added to those collected from the first batch of subjects. A new *P* is calculated based on the augmented sample size and is compared once again with the lower and upper criteria. The cycle continues until the experiment is stopped by one of the stopping rules. Whenever *P* is in the uncertain region and the addition of *n* added subjects would exceed the upper bound of the sample size model, the experiment is stopped without a significant result (the null hypothesis is retained). Otherwise, *n* added subjects can be added for another test with augmented sample size.

Any time a decision is reached about the null hypothesis (significant or not), there is always the possibility that one has made an error of inference. If the null hypothesis has been rejected, one may be correct or one may have made a type I error. If we fail to reject the null hypothesis, one may be correct or one may have made a type II error. The only exception to this circumstance is when the exact population values are known, such as during a computer simulation, and in these circumstances we can count type I and type II errors over a large number of tests. With the variable-criteria SSR, the rates of both type I and type II errors are stable for the entire experiment. The rate of type I errors will be close to the nominal alpha. The rate of type II errors, β, will be equal to 1 minus the observed level of power in the experiment. Thus, if the experiment has 90% power, the rate of type II errors will be 10%.

Available sample size models currently range in size from 3/9 to 10/40 for the lower and upper bounds. The model is best selected by using the power curves[7,8] if reliable prior information is available. The lower bound must be selected carefully because the investigator must abide by the stopping rules after the first test of the experiment. If a small lower bound (such as 3) is selected, one must be willing to stop the experiment at that point if the resulting *P* value is either less than or equal to the lower criterion or greater than the upper criterion. Therefore, one must be willing to submit a sample size of 3 to reviewers for publication if the experiment is significant, and one must be willing to stop the

experiment with a nonsignificant result if the *P* value exceeds the upper criterion. The upper criterion may be as small as 0.150. The greatest risk for a type II error during an experiment is when the sample size is small with low power. If one would be interested even in small effect sizes, it could be best to begin with a larger sample size model to increase power and avoid type II errors. Regarding the upper bound, one should try to choose a sample-size model with an upper bound that is high enough that one would not wish to exceed that number of subjects anyway.

The rate of type I errors is inflated by the intention or the willingness of the investigator to add subjects instead of by the actual addition of subjects to the test.[11] If the investigator sets out to use the variable-criteria SSR with a lower criterion of 0.02, for example, and if the obtained *P* value on the first test is less than 0.02, the investigator should indicate in the publication that the *P* value for the overall SSR was $P \leq 0.05$ not $P \leq 0.02$ (see examples of reporting *P* values that follow).

When to Use the Variable-Criteria SSR

A stopping rule is useful and appropriate for some but not all experiments in the biomedical and behavioral sciences. First of all, the variable-criteria SSR should be used only when a null hypothesis significance test is appropriate. The null hypothesis test is the dominant paradigm for statistical analysis in the biomedical sciences, but many statisticians think it should not be (see the section *Should We Be Doing It This Way?* in reference 7). The procedure has many weaknesses, is poorly understood by many who use it, and is often abused to draw conclusions that are unwarranted. It remains a useful tool for any study where the goal is simply to make a decision as to whether a difference between groups or conditions exists, and if so, in which direction.[10,11] The SSR extensions of the null hypothesis technique allow this decision to be made with the fewest number of subjects on average.[3,7,11,21]

A null hypothesis significance test does not help an investigator to determine whether a difference between 2 groups or conditions is an important difference. Even very tiny effects can be detected with a high level of significance if sufficient sample size is used.[15] The label 'significant' does not mean 'important'—and this is one reason that an investigator should use no more sample size than is sufficient to detect an effect that would be considered important. Arbitrary use of very high levels of power in an experiment wastes animals in pursuit of tiny or meaningless effects. The investigator should communicate the size of a meaningful effect to the IACUC and power the experiment accordingly. The variable-criteria SSR is just an extension of the null hypothesis test, so it suffers from the same defect. See the section *Estimation of Effect Size after Variable-Criteria SSR* for additional discussion.

The variable-criteria SSR assumes that one will be able to conduct an experiment with a few subjects at a time and that the measurements from those subjects will be known before additional subjects are added to an experiment. For that reason, the SSR technique may not be particularly useful in experiments with slow-growing tumors, long-term dietary manipulations, or other experiments in which the results from the first few subjects may not be known for many weeks or months. Even if the subjects can be tested a few at a time, some experiments require that all of the data be processed together, such as some assays in which preserved samples must be compared with a single standard in a single procedure. This situation does not work well with the SSR unless the sample can be split so that part of it can be analyzed preliminarily in inexpensive fashion (for example, an assay of a blood sample) and then again at the end of the experiment if the result is significant.

The variable-criteria SSR was designed for relatively small experiments in the biological sciences in which the investigator expects large effect sizes. Another form of SSR, called CLAST,[3,21] is available for larger studies in areas for which the anticipated effect size may be much smaller (for example, psychology). CLAST is far more efficient than is the fixed stopping rule,[3,7] but CLAST does not hold the type I error rate constant, as does the variable-criteria SSR; with CLAST, the observed alpha can become deflated and slightly reduce the power of the test.[7]

When sample sizes are unequal, the *t* test or ANOVA with the fixed stopping rule can be used without affecting the rate of type I errors as long as the other assumptions of the ANOVA, such as homogeneous variances, have been satisfied. This application is also true of the variable-criteria SSR across a range of sample size models.[8] As with ANOVA with the fixed stopping rule, type I errors with the variable-criteria SSR increase when the variances are different. This inflation of alpha is worse when sample sizes and variances both are unequal. When variances are heterogeneous, the investigator should use a test that constructs an error term based on separate within-group variance estimates (such as the Welch *t* test[20]) instead of on pooled within-group variance estimates. These tests are commonly available in full-featured statistics packages, and they work quite well with the variable-criteria SSR.[8]

The variable-criteria SSR can be used with nonparametric tests such as the Mann–Whitney–Wilcoxon, Wilcoxon signed-rank, and Kruskal–Wallis tests, although one must use different tables of criteria (contact the author). Nonparametric tests do not eliminate the problem with unequal variances even though nonparametric tests are considered to be 'distribution free.' One should not abandon parametric tests in favor of nonparametric tests if the only problem is heterogeneous variances, because separate-variance parametric tests can be effective in this situation.[8] In addition, parametric tests continue to work well under some conditions of nonnormal distribution,[2,7] if there are no other violations of the assumptions of ANOVA.

As long as the variances are similar, the variable-criteria SSR can be used when sample size has been reduced in one group because data have been lost for some reason (for example, technical failure or death). If the variances are markedly different, one should use a separate-variance test or make every effort to equalize the sample sizes.[8,20] Fitts simulated conditions in which additional sample size was added in sequential tests in order to compensate for lost sample size, and this accommodation had a negligible effect on the rate of Type I errors or power with the variable-criteria SSR.[8] If sample size is lost in one group on one test, it can be added back to that group in the next test without adversely affecting the process.

Example: *t* Test Using Variable-Criteria SSR

Suppose an investigator is interested in whether a small molecule has antipyrogenic activity (that is, will reduce fever) with respect to a particular pathogenic bacterium in vivo. Although the in vitro tests have yielded promising results, fever is a complicated process that can be studied in detail only in live animals. Preliminary tests with the small molecule alone have indicated that it is safe to use in large doses, so the investigator plans to

use a single large dose in the first efficacy experiment. The experimental group will receive an injection of the drug, and the control group will receive the vehicle alone. All animals will then be challenged by infection with the bacterium, and their core temperature will be monitored continuously by telemetry. Use of the variable-criteria SSR will help to ensure that only the minimal number of mice will be exposed to the bacterium.

Continuing with the example scenario, previous studies with this dose of the bacterium have indicated that the peak pyrogenic effect (mean ± 1 SD) is 2.4 ± 0.9 °C above normal body temperature. This potential drug will be most interesting, from a research standpoint, if it reduces the fever by at least 50% of the peak effect. However at this stage, whether a difference exists at this single dose is the focus, not how big the difference is. If a difference does exist, the next task is to generate a dose-response curve that is designed to detail the size of effect at each dose. Therefore, assuming that the vehicle-treated animals will experience a peak fever of 2.4 ± 0.9 °C, the investigator wants to ensure a high probability of detecting an absolute effect that is as large as 1.2 °C in the treated group. Assuming that the treated group will also have a standard deviation of 0.9, our standardized effect size *d* is 1.2/0.9 = 1.33 standard deviations. The investigator elects to use conventional values of 0.05 for α and 0.80 for power.

The next decision is whether to use a 1- or 2-tailed test (see section *One-Tailed* t *Tests using Variable-Criteria SSR*), and this example experiment could go either way. Because the drug already is known to have potential antipyrogenic activity from in vitro tests, a 1-tailed test could be used, with a result considered significant only if the drug reduces fever. In this case, the null hypothesis would be that the drug does not reduce fever, and the alternative hypothesis is that the drug does reduce fever. This choice would increase efficiency for detecting an antipyrogenic effect, because 1-tailed tests are more powerful than 2-tailed tests. Consequently, the investigator may detect a significant effect with a smaller sample size. The investigator then would conduct the variable-criteria SSR by using the *P* value from the 1-tailed test and consider a result significant only if the fever was reduced. On the other hand, if the investigator would be interested in an outcome in which the fever was unexpectedly increased instead of decreased, a 2-tailed test should be planned. The mechanism of the unexpected pyrogenic effect of a drug that was theoretically antipyrogenic in an vitro model would be revealed only through use of a 2-tailed test, and in the present example that is what the investigator decides to do. The null hypothesis is therefore that the drug has no effect on the pyrogenic action of the bacterium, and the alternative hypothesis is that the drug either increases or decreases fever.

The original article on the variable-criteria SSR[7] includes a figure for determining a sample size model from a standardized effect size and a desired amount of power in an independent-groups *t* test (Figure 6 in reference 7). The figure contains 2 sets of curves for an α of 0.05, one for smaller upper bounds and one for larger upper bounds. In the example scenario, an inspection of both sets for an effect size of approximately 1.33 on the abscissa and a value of 0.8 for power on the ordinate reveals that either a 6/18 or a 7/21 sample size model would provide approximately the desired power. From a practical standpoint, because the investigator in the example scenario has purchased 6 telemeter systems for continuous measurement of core body temperature, the experiment can be conducted with 3 subjects in each of the 2 groups at one time. For that reason, the investigator decides to use the 6/18 sample size model instead of the 7/21 model.

An initial test is conducted with 3 subjects per group, but the data are not analyzed at that point because the method calls for a lower bound of 6, not 3. In addition, reviewers of manuscripts or grants likely would object that having 3 subjects per group was insufficient to establish the efficacy of a new antipyrogenic drug. Therefore, the investigator analyzes the first set of data once the sample size is 6 per group. After this point, however, the investigator can test after each addition of 3 subjects per group. Therefore, the *n* added is established as 3 subjects per group at each stage of the experiment after the lower bound. Because the upper bound is 18, the investigator should add no additional subjects if the group size ever reaches 18 subjects.

With an α of 0.05, a sample size model of 6/18, and an *n* added of 3, Table 2 in reference 7 indicates that the stopping criteria are 0.0200 and 0.300. Therefore, the experiment will stop with a significant result if $P \leq 0.0200$ or a nonsignificant result if $P > 0.300$. Otherwise, 3 subjects per group will be added and data retested as long as adding 3 subjects per group would not exceed 18 subjects per group.

I simulated this experiment several times as conceived here by using normally distributed random numbers drawn from 2 populations with means and standard deviations of exactly 2.4 ± 0.9 °C and 1.2 ± 0.9 °C, respectively. The simulated experiments were usually powerful, and many rejected the null hypothesis with $n = 6$ per group. I selected for presentation here a less fortunate example in which the means were closer than expected, such that the experiment would require several stages to achieve significance. In this case, after testing with 6 subjects per group, the means and standard deviations were 1.99 ± 0.90 and 1.2 ± 0.78, $t(10) = 1.617$, $P = 0.1369$. This *P* value falls into the uncertain region because it is greater than 0.0200 and less than or equal to 0.300. Because adding *n* added subjects will not exceed the upper bound of 18, one can add 3 subjects per group and repeat the test. With a total sample size of 9 per group, the means and standard deviations were 2.07 ± 0.81 and 1.31 ± 0.75, $t(16) = 2.06$, $P = 0.0558$. These values again fall into the uncertain region, although this result appears more promising than the result of the first test. Even a *P* value of 0.049 would not be significant at this point, because here the criterion for significance is 0.0200, not 0.05. With 12 subjects per group, the statistics were 2.06 ± 0.78 and 1.18 ± 0.85, $t(22) = 2.65$, $P = 0.0146$. Because this *P* is less than 0.0200, one can reject the null hypothesis that the effect of the drug is 0 in the population. By inspecting the direction of the difference, one can conclude that the drug reduces fever induced by the bacterium at $P \leq 0.05$—not $P \leq 0.02$. The criterion of 0.02 was used to ensure that, after inflation of type I errors from repeated sampling, these errors were at most 5%.

In the methods section of the hypothetical manuscript describing this experiment, the investigator would report that the experimental design used the variable-criteria SSR (cite reference 7) with a 2-tailed α of 0.05, a sample size model of 6/18, 3 subjects added per group at each stage of the experiment, and stopping criteria of 0.0200 and 0.300. The investigator should indicate here that significant results will be reported with both a P_{SSR}, meaning the final 'working *P* value' derived from the SSR procedure, and an experimentwise *P* value for the selected level of α (see following).

In the results section of the manuscript describing the example experiment, the investigator would give the means, standard deviations, and ultimate sample sizes and should report that $t(22) = 2.651$. Because the P value of 0.0146 is less than the lower criterion of 0.0200 for the 0.05 level with the variable-criteria SSR, the result should be reported as significant at the 0.05 level. The result should not be reported as '$t(22) = 2.651$, $P = 0.0146$' without identifying the P value as resulting from the variable-criteria SSR procedure because someone who does not read the method section may not realize from the presentation that the experiment had several chances to succeed. The terminology '$t(22) = 2.651$, $P_{SSR} = 0.0146$, $P \leq 0.05$' will alert both the immediate reader and future meta-analyst to refer to the methods section to learn what 'P_{SSR}' means. Additional research likely will identify the meaning of the P_{SSR} value with respect to how replicable the experiment is.[12]

On average, an SSR using the 6/18 sample size model will yield a significant result with a smaller sample size than will the fixed stopping rule.[7] However, that did not occur in this example. Given a 2-tailed independent-groups t test with α of 0.05, power of 0.80, and effect size of 1.33, a regular power analysis suggests a sample size of 10 per group, which is less than our ultimate 12 per group. As mentioned previously, I selected this example from several simulations specifically because it was slow to generate a significant result. If one had conducted the experiment using the fixed stopping rule, one would have conducted the test on all 10 subjects per group before the P value was calculated. If the P had been greater than 0.05, the experiment would have been stopped without a conclusive result of any kind. Instead, when the sample size of 9 per group with the SSR was not significant, one could continue testing with 12, 15, or even 18 per group before stopping. These additional chances to achieve significance with smaller-than-expected mean differences are strengths of the SSR. Now, with evidence that an effect exists in the example scenario, the investigator can proceed to design a dose–response experiment with larger numbers of mice to determine the effect at each dose.

Estimation of Effect Size after Variable-Criteria SSR

In the example just presented, the observed effect size with these sample data was 1.07 instead of 1.33. Because this was a simulation, one can know that it is an underestimate of the actual effect size, 1.33. Several simulations were necessary to find an effect this low. However, this result raises the question of whether the effect size should be estimated at all after one has used an SSR procedure.

If the true treatment effect in the population is unknown, the investigator cannot know whether the sample's observed effect is an underestimate, exactly right, or an overestimate. However, the true effect size can be known in a simulation or thought experiment. Suppose an investigator conducts the previous example experiment using the fixed stopping rule with $\alpha = 0.05$ and $\beta = 0.20$ (that is, power = 0.80) and a true effect size of 1.33 standard deviations. The power analysis suggests using 10 subjects per group. With random sampling, the obtained effect size will be larger than 1.33 about half the time, and half of the time it will be smaller, but the power analysis ensures that the investigator will reject the null hypothesis 80% of the time with 10 subjects per group. If instead one uses 6 per group as with the SSR, one could reject the null hypothesis with this small number and stop testing. This would be most likely to happen if the sampled effect size was larger than the actual effect size. Of all possible samples with the fixed stopping rule and a true effect size of 1.33, 80% (the level of power in this case) would require a sample size of at most 10 to reject the null hypothesis, and only 20% (type II errors) would require a larger sample size. Therefore, the SSR is more likely to stop early with a larger effect size than to stop late with a smaller effect size. As noted in the preceding example, the author found several simulations that stopped early before one was identified that stopped late. Thus, when the effect size in the population is known, the SSR procedure is more likely to stop with an overestimate than with an underestimate of the true effect size.[1]

All of this emphasizes the fact that the null hypothesis test in general, and the SSR in specific, is not the best procedure if the goal is to estimate the size of an effect. A null hypothesis test is good for deciding whether a difference exists, and if so, in what direction. The P value does not contain any information about effect size. Estimation of the actual effect size would require a confidence interval approach and generally would require more animals, because the width of the confidence interval decreases (becomes more precise) as the sample size increases. If it is important to the research to know the actual size of an effect with a certain degree of precision, the investigator should use a confidence interval approach and use a greater number of animals.[14] If a null hypothesis test is not appropriate, the investigator should explain this situation to the IACUC and justify the use of additional animals to achieve greater precision.

If an estimation of the size of the effect is of lesser concern in the experiment, the use of an SSR will save subjects, on average, compared with the fixed stopping rule. If the null hypothesis is false, the variable-criteria SSR will save subjects by stopping early if the sampled effect is larger than the average effect. This situation occurs frequently. The variable-criteria SSR also tends to stop experiments early when the null hypothesis is true. In the previous example, the upper criterion was 0.300, which denotes a 70% chance of stopping the experiment early after the first iteration with $n = 6$ if the null hypothesis is true. This outcome also saves animals.

Repeated-Measures *t* Test Using Variable-Criteria SSR

When a t test requires a one-sample, matched-samples, or dependent-samples design, the procedure is exactly the same as described for the 2-sample case, except that the power of the test is different and varies with the number of pairs of scores and the correlation between the scores. The size of the effect, d, is the mean of the difference scores divided by the standard deviation of the difference scores,[4] where a difference score is the difference between a subject's 2 scores on the dependent variable. Power curves for the dependent samples t test are presented in Figure 8 of reference 7. The best sample size model can be estimated by inspecting the graphs with the effect size on the abscissa and the desired power on the ordinate and looking for the nearest intersecting curve. This case includes only one 'group,' so an n added of 1 means that one will add a single subject's pair of scores to the single group. The stopping criteria are determined from Table 2 of reference 7 exactly the same way as for the independent-samples test. The P value from a dependent samples t test is compared

with the lower and upper criteria as usual. If the result is in the uncertain region, *n* added subjects are added as long as the addition would not exceed the upper bound.

One-Tailed *t* Tests Using Variable-Criteria SSR

A one-tailed *t* test is conducted when the investigator is willing to make a one-sided null hypothesis. The typical 2-tailed null hypothesis is that there is no difference between 2 groups or conditions in the population. The alternative hypothesis, which may be accepted if the null hypothesis is rejected, is that there is a difference between the groups in the population. This difference could be either in a positive or negative direction. To achieve a 2-tailed test, one half of α is placed into each tail of the sampling distribution of the *t* statistic (for example, 2.5% in each tail for an α of 5%). By contrast, a one-tailed test puts all 5% of α in one tail of the sampling distribution so that the investigator is predicting the direction of difference. The null hypothesis of a one-tailed test might be stated as "the treatment does not decrease the mean from the control group," and the corresponding alternative hypothesis would be stated as "the treatment decreases the mean from the control group." Placing all of α in a single tail of the sampling distribution increases power because a smaller *t* is necessary for significance. However, the *t* must have a sign that is compatible with the null hypothesis to be significant. The investigator must be willing to conclude nothing (that is, retain the null hypothesis) if the result is an extreme *t* value in the opposite direction from the predicted direction. If at first an investigator plans to use a one-tailed test but then, after collecting data, decides to use a 2-tailed test because the results are in the unexpected direction, the type I error rate is actually 0.075, not 0.05 (that is, 0.05 from the predicted direction and 0.025 from the unpredicted direction).

A one-tailed *t* test will generate a *P* value appropriate for comparing with the stopping criteria exactly as described in the preceding sections. This option works fine for either independent- or dependent-samples *t* tests. One-tailed tests are discussed widely in statistics textbooks.

Example ANOVA Using Variable-Criteria SSR

The procedure for using a one-way ANOVA for multiple groups with the variable-criteria SSR is exactly the same as with the *t* test except that power varies with the number of groups, not just the sample size per group, so different power curves may be necessary to identify the sample size model that is best to use. Power curves have been provided for 4 groups[7] and for 6 or 8 groups.[8] The rate of type I errors is stable at the nominal alpha for as many as 20 independent groups.[8] The effect size is based on *f*[4] instead of *d* because the numerator is not a simple difference between 2 means, as was the case in the *t* test. Easy calculation methods are available for *f*.[6,19] Each one-way ANOVA generates a single *P* value, and a significant *P* indicates that a difference is present somewhere among the multiple means in the study. Like the usual ANOVA, the variable-criteria SSR does not tell us where the difference is. However, once the stopping rule has been used to determine that an effect exists somewhere among the groups, one can use the usual posthoc tests to examine the differences among the means in exactly the same way as customary (that is, stopping criteria are not involved in these comparisons).

For example, suppose an investigator wants to test whether a knockout of either or both of 2 genes affect the response to a painful stimulus in mice. The investigator obtains wildtype mice and strains of congenic mice with a knockout of either gene A (KOA), gene B (KOB), or both (KOAB). The investigator will test the latency to withdraw a hindlimb when it is placed on a warm plate. Although no tissue damage is anticipated, limiting the number of animals that are exposed to painful stimuli is desirable. However, little prior information is available to use for a power analysis. If one guesses wrong about the sample size with a fixed stopping rule one could test way too many animals in order to find a very large effect or could test an insufficient number of animals and concluding nothing. In this situation, the variable-criteria SSR allows the investigator to start with a modest sample size and work toward larger sample sizes until significance is achieved. The null hypothesis is that all 4 population means are equal. The alternative hypothesis is that the 4 means are not equal.

In this example, the literature on this standard behavioral test indicates that the wildtype strain has a normal latency to withdraw a hindpaw of 30 s with a standard deviation of 7 s. The investigator expects fairly subtle differences, such as 1 standard deviation, and wants to detect a difference even if only one of the mouse lines showed a change this large. To calculate this effect size of interest,[6] one could assign means and standard deviations of 30 ± 7 to 3 groups and 37 ± 7 to 1 group, that is, a 1 standard deviation difference in only 1 group. The power analysis would be the same if the effect was in the other direction, that is, 3 groups at 30 ± 7 and 1 group at 23 ± 7. Entering these data into the sample size calculator reveals that the effect size *f* is 0.43. (Incidentally, the same program indicates that one would need a total sample size of 16 per group for significance with the fixed stopping rule to achieve a power of at least 80%. This is not necessarily the sample size that will be used with the variable-criteria SSR.) By inspecting Figure 10 of reference 7 for a 4-group ANOVA for this approximate effect size one learns that sample size models of 9/27 or larger will provide about 80% power at the 0.05 level of significance. In this example, testing more than 27 mice would not be practical, so the 9/27 model is fine. Because the hot-plate test is rapid, the investigator can test for significance after adding only 1 subject to each group (an *n* added of 1). Table 2 of reference 7 shows that the stopping criteria for the 9/27 model at the 0.05 level of significance with *n* added of 1 are 0.013 and 0.450. The investigator would report the value of α, the sample size model, and the *n* added in the methods section of the resulting manuscript.

I simulated this experiment with populations in which 3 groups had means and standard deviations of 30 ± 7 and 1 group had 37 ± 7. The first test with 9 mice per group yields $F(3, 32) = 2.322$, $P = 0.094$. Adding one subject per group to that dataset for 10 per group gives $F(3, 36) = 3.072$, $P = 0.040$. Adding one more subject per group to that dataset for 11 per group leads to $F(3, 40) = 4.595$, $P = 0.007$. Therefore, with 11 per group, one finds a working p_{SSR} value less than our lower criterion of .013 for the 9/27 model and rejects the null hypothesis that all 4 population means are equal at the 0.05 level of significance for the entire experiment. The investigator would report these findings in the results section of the manuscript as $F(3, 40) = 4.595$, $P_{SSR} = 0.007$, $P \leq 0.05$. The variable-criteria SSR in this instance detected significance among the means with only 44 total subjects, whereas the fixed stopping rule would have required 64 total subjects (16 per group, see previous paragraph). This good outcome is representative of the average reductions in sample size (approximately 30%) with

larger sample size models for an effect of known size in many simulations using the variable-criteria SSR.[7]

The obtained means and standard deviations for the 4 groups are: wildtype, 33.19 ± 6.41 s; KOA, 29.01 ± 5.07; KOB, 28.33 ± 7.64 s; and KOAB, 37.13 ± 5.91 s. Posthoc comparisons of all pairs of means at the 0.05 level with the Tukey Honestly Significant Difference test demonstrates that both of the single-knockout groups, KOA and KOB, are significantly different from the double-knockout group, KOAB.

This random simulation illustrates how a real-world dataset obtained by random sampling can achieve a significant result that is not always easily interpretable. The variable-criteria SSR helps to achieve a significant result with a small sample size in an ANOVA. It does not guarantee that the result will be easily interpretable any more than the fixed stopping rule does. Perhaps the investigator would have liked at least one of the knockout groups to have been significantly different from the WT group. One can conclude only that the double knockout group has a longer latency than either of the single knockout groups. Whether any of the 3 knockout groups are either increased or decreased relative to the wildtype group remains unknown. This experiment used a value of 80% for power, which left a 20% chance of a type II error. One may have obtained a more interpretable result if one had increased power to 90% and used more subjects. This is exactly the same regret that an investigator could face when using the fixed stopping rule with such results. Another article discusses cautions related to using excessive statistical power.[9]

This example demonstrates that, when used with a one-way ANOVA, the variable-criteria SSR stops the experiment when it first detects a significant effect, not when it detects the effect the investigator wants to detect. This problem can be addressed with ANOVA by using a planned contrast as the focus of the variable-criteria SSR instead of the *P* from the omnibus *F*.

Example Planned Contrast Using Variable-Criteria SSR

Any factorial ANOVA can be decomposed into a one-way ANOVA. For example, an ANOVA with 3 levels on one factor and 2 levels on a second factor could be analyzed as a simple one-way ANOVA with 6 groups. Effects can then be explored with planned contrasts. Any one-way ANOVA with *G* groups has at least one set of *G* – 1 independent (orthogonal) planned contrasts that can be made among the means of those groups with 1 df each. Therefore, any between-groups contrast can be used as the focus of the stopping rule.[8] The importance of this practice will become clearer with a specific example.

Suppose an investigator designs an experiment to test whether the deletion of a particular gene from a virus affects early virulence in mice. A well-controlled experiment would include 4 groups: an untreated control group (group C), a group treated with an avirulent virus that still may induce an immune response (group G), a positive-control group treated with the fully virulent virus (group V), and a group given a virus that has a deletion of the gene of interest (group K). The outcome measure will be a measure of the mobilization of immune cells within a short period of time after inoculation prior to progression to severe infection in the virulent group. The investigator already knows that group C will have no active mobilization, group G will have mild mobilization, and group V will have a marked mobilization even early after the infection. The investigator wants to determine whether the gene deletion in group K affects virulence compared with the virus given to group V. The intention is to euthanize the mice before symptoms of a full-blown infection occur in the virulent group V. However, whether the knockout of the viral gene will increase or decrease virulence is unknown, and some mice may experience severe symptoms before they can be euthanized. Although the investigator includes a frequent monitoring plan with clear surrogate endpoints, the experiment is best conducted with the fewest number of animals that will give a statistically significant effect. If the experiment is successful, it will yield insight into the mechanism of virulence of this strain of virus and may lead to an effective vaccine.

As just observed with the omnibus *F*, the variable-criteria SSR will stop the experiment as soon as any difference becomes significant. If that strategy is used in the present experiment, the SSR will stop the experiment as soon as the positive control becomes different from the untreated control. The effect of interest is whether group K and differs from group V. Unless this effect is larger than the difference between groups V and C, a significant difference between groups K and V will not be apparent if the experiment is stopped by the SSR based on the *P* from the omnibus *F*. An alternative is to construct a single-df planned contrast between groups K and V and to use that *P* value to stop the experiment. This approach will allow other effects to become large without affecting the stopping rule.

To plan this test, one calculates the effect size from the size of difference that one would like to be able to detect between groups K and V. In this example, the investigator is hoping for a difference between the means (either positive or negative) of at least half the size of the virulence response of group V. The investigator is familiar with the response of group V, which is at least 3 standard deviations above the value of control group C. If the response of group K is half that large, it would be 1.5 standard deviations. So, the size of the effect in standardized units is $d = 1.5$.

The effect size has been calculated as *d*, not *f*, because the investigator is interested in the difference between 2 means, K and V, as in a *t* test, and not the differences among all 4 means. A *t* test is a special case of the *F* test. The numerator df for an *F* test is the number of groups minus one, so with 2 groups, the numerator has 1 df. The square root of an *F* value with 1 df in the numerator is identical to a *t* value for the same data. Therefore, a single-df *F* test is functionally the same as a *t* test, the *P* values for the *F* and *t* will be identical, and an *F* test with 1 df can be treated the same as an independent-groups *t* test.

Figure 6 in reference 7 for the independent-groups *t* test can be used to determine the sample size model. The 5/19 sample size model is the most appropriate for 80% power and an effect size *d* of 1.5 for an α of 0.05. In this example, the investigator decides to use an *n* added value of 5, maybe because mice often are housed 5 per cage, or perhaps the decision is completely arbitrary. The 5/19 model with an *n* added of 5 per group in the Table 2 of reference 7 gives a lower criterion of 0.026 and an upper criterion of 0.28.

A single-df contrast *F*-ratio can be constructed by dividing a mean square for the intended contrast by the error mean square for the one-way ANOVA. The intended contrast is described by weighting each of the means with numbers such that the sum of the weights for all means equals zero. The sum of squares for the numerator is calculated as the square of the sum of the weighted means divided by the sum of the ratios of squared weights to cell

frequencies. For example, if each of the means in the example virus study is labeled as C, G, V, and K and the respective sample sizes as n_c, n_g, n_v, and n_k and if only the means for V and K are compared, respective weights of 0, 0, 1, and –1 would be used:

$$SS_{1df} = ((0) \times C + (0) \times G + (1) \times V + (-1) \times K)^2 / ((0^2)/n_c + (0^2)/n_g + (1^2)/n_v + (-1^2)/n_k)$$

or, after reducing the formula to delete the zero terms,

$$SS_{1df} = (V - K)^2 / (1/n_v + 1/n_k),$$

which in this case is the squared difference between the 2 means divided by the sum of the reciprocals of the sample sizes. The numerator mean square is the sum of squares divided by the df, which is 1.

The author simulated this experiment using population effect sizes of 0 for the control group C, 0.3 for group G, 3.0 for group V, and 1.5 for group K. As predicted, the omnibus *F* among the 4 groups was significant at the first test with 5 per group ($P < 0.001$) because of the large difference between the untreated control group C and the positive control group V. However, the effect between the V and K groups by the planned contrast with 5 subjects per group was $F(1, 16) = 5.288$, $P = 0.0353$. This result looks promising, but is in the uncertain range between the stopping criteria of 0.026 and 0.28. After adding another 5 subjects per group, the result was $F(1, 36) = 20.326$, $P = 0.000067$. At this point, one can reject the null hypothesis at the 0.05 level of significance, indicating that group K is different from group V. This planned contrast would be reported in the results section of the manuscript as $F(1, 36) = 20.326$, $P_{SSR} = 0.000067$, $P \leq 0.05$. Other comparisons can be made in the usual way with posthoc tests or additional planned contrasts. Because 4 means are tested, 3 (that is, 4 – 1) total orthogonal contrasts can be made, and 2 of these are free after the contrast that was the focus of the stopping rule. Because these other posthoc or planned contrasts were not tested repeatedly to find significance in the experiment, they can be tested at the usual α instead of using a reduced lower criterion.

The example with the knockout mice in a previous section might have used a planned comparison between the WT group and the average of all 3 knockout groups. This decision could have guaranteed significance between the WT group and at least 1 of the other 3 knockout groups if the experiment was stopped with a significant result. The risk of that strategy is that one of the knockout conditions might greatly decrease the mean and another knockout group might greatly increase the mean, so that the average of the groups is zero. This alternative strategy would not detect this interaction.

Planning Simultaneously for Significance and Repeatability

A significance test allows an investigator to assert whether a difference exists, and if it does, in which direction.[10] The investigator also can state the small probability that a type I or type II error has been made. A companion article[9] discusses issues relating to when an effect probably should or should not be replicated before publication to ensure that the effect is repeatable. Experiments with extremely small *P* values are much more likely to be repeatable with a significant effect in the same direction, and replication is less important than for an effect in which the *P* value is close to the α for the experiment. Science is best served if published effects are repeatable, and repetition of results with animal subjects should be done with rationality. A result with a $P < 0.005$ has better than an 80% chance of being significant at the 0.05 level in an exact replication,[12] and such experiments should not waste animals on needless repetitions if replication is the only reason for the repetition (as compared with technical error, for instance). In contrast, a result with a *P* value close to 0.05 has only a 50% chance of being significant at the 0.05 level in an exact replication,[12] and a repetition would probably be wise before publishing the result.

When conducted in stages, an experiment designed to use the variable-criteria SSR can build 'repeatability' and 'significance' with fewer subjects on average than is possible with the fixed stopping rule. If desired, the stages can be conducted independently using different batches of drug, cohorts of animals, experimental days or experimenters.

Consider what happens if an investigator uses an α of 0.005 with the fixed stopping rule in an effort to demonstrate repeatability. This conservative α will require a large *t* value to achieve significance and a large sample size to have acceptable power. Because the fixed stopping rule requires that all subjects be tested before examining the data, all of the large sample size must be used even to find out that the null hypothesis was true. For example, suppose an investigator planned an experiment with an anticipated effect size of 1.4 standard deviation units in an independent-groups *t* test. Using a 2-tailed test, an α of 0.005, and a power of 90%, the sample-size recommendation for the fixed stopping rule is 20 per group. If the null hypothesis is true and the fixed stopping rule is applied, the experiment will not be stopped until all 40 animals in the 2 groups have been tested.

Now consider the same example with an effect size of 1.4 using the variable-criteria SSR. Based on Figure 6 of reference 7 for 90% power and an α of 0.005, the investigator can test using the 8/32 model with an *n* added of between 1 and 8. For example, an *n* added of 8 would allow sequential testing of 8, 16, 24, and 32 subjects. In the original simulations,[7] this model rejected the null hypothesis 92% of the time on average, so the power is at least the same as that for the fixed stopping rule. The stopping criteria from Table 2 of reference 7 for the 8/32 model at the 0.005 level and *n* added of 8 are 0.0018 and 0.20. The investigator would test first with 8 subjects and compare the *P* with 0.0018 to determine whether it is significant at the 0.005 level. The experiment would be stopped at a *P* value exceeding 0.2. If the null hypothesis is true, 80% of experiments will be stopped after this first test with only 8 subjects per group (1.00 – 0.2 = 0.8). If the result is significant, the investigator would report that the experiment was significant at the 0.005 level because *P* was less than the lower criterion of 0.0018 in the variable-criteria SSR, using a lower bound of 8 and an upper bound of 32, and that the probability of a significant exact replication is greater than 80%. The number of subjects per group when the effect reached significance also would be reported. In the original simulations,[7] the average sample size at the rejection of the null hypothesis with the 8/32 model with an *n* added of 8 was 16 per group (data not shown). Therefore, on average, the example study design saves 4 animals per group for an average of 8 total animals for the experiment when compared with the sample size of the fixed stopping rule.

In this example, an α of 0.005 determined whether the effect was replicable, not whether it was significant. Any result that is less than the lower criterion at the 0.05 level should be reported as significant, and any result that is less than the lower criterion at the 0.005 level should be reported as confirmed and likely to be replicated with significance at the 0.05 level.[12] Results with *P* values between the lower criteria for the 0.05 and 0.005 levels can be reported as interesting but unconfirmed. If the experiment is designed to demonstrate repeatability at the 0.005 level, the upper criterion for the 0.005 level should always be used (not the upper criterion for the 0.05 level). The upper criterion has much less effect on the type I error rate than does the lower criterion.[7]

Some investigators may be concerned about the independence of replications regardless of how small the *P* value is after the first repetition. A particular experiment may result in significance because of some error or peculiarity of the testing conditions and perhaps should be performed under completely independent conditions to demonstrate repeatability. The example in the preceding section can include completely independent replications because the experiment is conducted in stages (replications). Instead of conducting the experiment with 20 animals per group all at once, the experiment is conducted with 8 per group in the first test, another 8 per group in the second test, and so on, until the experiment is stopped because one of the stopping conditions is satisfied. Investigators who prefer not to merge the results from multiple replications into a single mean may consider multifactorial ANOVA.

Multifactorial ANOVA with the Variable-Criteria SSR

The preceding experiment can also be organized as a factorial ANOVA instead of as a *t* test.[8] One factor is the treatment variable with 2 levels, and the other factor is the replications with the number of levels needed to produce significance on the treatment factor. This factorial ANOVA will have 2 *F* ratios, representing the main effects of the treatment and replication factors, and a third *F* for the interaction of the treatment and replication factors. To use the variable-criteria SSR in this experiment, the replication and interaction factors are completely ignored, thereby focusing all attention on the treatment factor as data are collected. The previous example has 2 levels of the treatment factor, a control group and a treatment group, and the *P* value for the main effect of this factor will be identical to that of a *t* test between the 2 groups as if the other factor is ignored. One can evaluate this effect after the first test with 8 per group, after the second test with 16 per group, and so on until a decision is reached according to the rules of the variable-criteria SSR. Each test adds a level to the replication factor. In addition to the means and standard deviations, the investigator would report the sample size model used, the value of *n* added, the obtained P_{SSR}, the actual significance level, and the sample size per group at which the effect became significant. Investigators seeking the extra assurance of 'repeatability' can use a significance level of 0.005 instead of 0.05, as was done in the preceding section.

However, the analysis should not end there. Although the investigator's attention was focused on the stopping variable, the data have undergone a full 2 × 3 factorial ANOVA. Once the experiment has been stopped by using the treatment variable, testing of the effects of replications and the interaction of replications and treatments can proceed. Because these 2 other effects in the model were never used in the decision to stop the experiment, they can each be tested at the nominal alpha, such as 0.05, instead of at the lower criterion demanded by the variable-criteria SSR. The *P* value for these tests will have the usual meaning, so they need not be reported as P_{SSR}. If either the replicate factor or the interaction between the replicate and treatment factors is significant, unknown differences between the replications are causing effects that likely will be of concern in future studies. This procedure is more powerful than conducting 2 separate *t* tests and uses far fewer animals to demonstrate that the phenomenon is reproducible.[8] It is also logical, because all attempts to reject the null hypothesis should be reported in the resulting article. This reporting can be done by merging the replicates into a main effect of treatment or by displaying the means separately to illustrate interactions. An investigator should never conduct several identical replications and then present in the paper the replicate that is most favorable to the aims of the study.

The number of levels on the treatment factor is not limited to 2. Any number of groups can be tested by using either the omnibus *F* or a planned contrast as the focus of the stopping rule. This analysis is valid because the different factors in a multifactor ANOVA are completely independent of one another. The variable-criteria SSR has no way of being influenced by the fact that there are other independent factors in the experiment. If the variable-criteria SSR works in a *t* test without a factorial structure, it will work just as well with an *F* test within a factorial structure.

Whether the groups are a part of a main or interaction effect, the number of numerator df should not exceed 19 because the maximal number of groups that have been tested with simulations to assure that the type I error rate is stable is 20.[8] The method may work with additional groups because a large drift of alpha does not occur at 20 groups,[8] but further testing would be required to confirm this assumption.

The foregoing discussion leads to the question of whether the second factor must be a replicate or can be some other factor of interest. For example, if a drug effect is one of the treatments and a strain of mice (transgenic or wildtype) is the second factor, could the investigator use the variable-criteria SSR in this situation as well? The answer is a qualified 'yes'—the decision rule would have to be limited to a single *P* value, as in the treatment × replicate example. The replicates could be analyzed as a third factor or could simply be averaged as with a simple *t* test. The design now includes a 2 × 2 ANOVA with drug compared with vehicle on one factor and transgenic compared with wildtype on the other factor. The investigator could conduct a test with 8 animals per group, then 16 animals per group, and so on, until a *P* value for 1 of the 3 *F* ratios was significant. However, for the test to be valid, one must decide before the experiment which 1 of the 3 *F* ratios to use to make the stopping decision. The design has 3 completely independent effects, and any of these could be the focus of the decision rule. A legitimate approach would be to add sample size until either the drug effect was significant, or until the strain effect was significant, or until the interaction of drug and strain was significant. However, adding sample size until any 1 of the 3 effects was significant would not be legitimate, nor would adding sample size until all of the 3 effects were significant. The observed type I error rate would be inflated by an unknown amount.

Why Am I Using More Animals Instead of Fewer?

Investigators who have been accustomed to doing sequential sampling at the 0.05 level will have more difficulty achieving significance with the variable-criteria SSR because they need a smaller *P* to meet the lower criterion. These investigators have become accustomed to the great power of sequential sampling, but without knowing it, they have incurred a greatly increased level of type I errors. The variable-criteria SSR avoids that by using a few extra animals to control the rate of type I errors.

Other Techniques for Sequential or Interim Analysis

Techniques for sequential or interim analysis date to the late 1940s and have been heavily used in various fields but especially in clinical trials.[1,13] For safety reasons, interim assessment of data may be necessary at several points during a long trial. Some of these methods are available in proprietary statistics packages, and they, like the SSR discussed here, are more efficient in regard to sample size than is the fixed stopping rule when applied to preclinical research problems as well as clinical trials. The available methods vary in complexity and may require advanced mathematics or special training.[3,11] The variable-criteria SSR is the only test for interim analysis that was optimized for small-sample research (that is, 40 or fewer subjects per group), where labor-intensive studies with animal or human subjects often exact considerable monetary expense or ethical cost for adding sample size.

Other SSR used in clinical trials are designed to take a certain predetermined number of assessments of the accumulating data, and the simplest of these SSR require that all groups added to the test must be of the same size (that is, the *n* added must equal the lower bound). Tests that do not make this assumption are more difficult to use. Table 1 lists the adjusted critical value of *P* required by several group SSR tests after 1, 2, 5, 10, or 20 interim analyses with equal sample sizes. For example, with the Pocock method,[17] if a trial was designed to test significance after each fifth of the data was collected (that is, 5 total tests), a test would be considered significant only if the obtained *P* value was less than 0.0158.

The variable-criteria SSR works differently because the investigator begins with a number of subjects at the lower bound and then proceeds adding *n* added subjects at each test. The *n* added can be less than, equal to, or (in some cases) even greater than the lower bound. The number of tests conducted therefore depends on the lower bound, the upper bound, and the *n* added. However, when the lower bound and *n* added are equal, one can select sample size models that provide close comparisons with these other SSR listed in the table (see right side of Table 1). For example, a maximum of exactly 2 tests each can be added with the 4/10, 5/12, 6/12, and 7/14 models with *n* added of 4, 5, 6, and 7, respectively. A maximum of exactly 5 tests is possible if the 3/15 model is used with an *n* added of 3. Available sample size models do not include 10 or 20 tests unless the lower bound and *n* added are unequal. For 10 tests, the lower criterion is listed for the 9/36 model with *n* added of 3. For 20 tests, the lower criterion is listed for the 10/30 model with *n* added of 1 (although the actual maximal number of tests is 21). For more than 2 tests, the variable-criteria SSR would allow the rejection of the null hypothesis with a considerably larger *P* value than the other tests. This benefit is possible because the variable-criteria SSR includes an upper criterion that stops a large percentage of the experiments after only the first test when the null hypothesis is true and this reduces type I errors. More complex SSR from the clinical trials literature can include both an upper and lower boundary for significance and an allowance for testing at unequally spaced intervals, but these techniques are more difficult to use than is the variable-criteria SSR.[18]

The reduction in critical *P* for significance is not a simple function of the number of tests in the variable-criteria SSR; it also depends on the values of the lower bound, upper bound, and *n* added. For an extreme example, at the 0.05 level of significance and a constant *n* added of 1, the lower and upper criteria for the 3/9 bounds are 0.015 and 0.43 for a maximum of 7 tests; the criteria for the 10/40 bounds are quite similar, 0.015 and 0.250, for 31 tests (see Table 2 of reference 7). Because the upper criterion has only a small, 'fine-tuning' influence on the overall obtained type I error rate,[7] the number of tests is only a minor determinant of the critical *P* value required for significance in these cases (that is, 7 compared with 31 tests).

Table 1 supports the conclusion that the variable-criteria SSR is comparable to other SSR techniques, and even a little better within its niche, which is small-sample research. With small-sample biomedical research, any increase in efficiency reduces the number of animal or human subjects that may be subjected to invasive procedures. The variable-criteria SSR is not mathematically rigorous because its criteria are based on simulations (albeit very large ones). For this reason, statisticians should try to improve techniques for small experiments with the following goals: (1) more rigorous mathematically than is the variable-criteria SSR; (2) more powerful in small-sample experiments than are the current simple rigorous methods; (3) understandable to the general population of biomedical researchers who do not consult statisticians; (4) can include lower bounds and *n* added that are flexible and not always equal;[8] and (5) available at low cost.[7]

Conclusions and Advice for IACUC

This article does not substitute for an education in basic statistical methods. Decisions about statistics are best made as the experiment is being designed and by persons who have training in statistical methods and experimental design. The first necessary decision is whether a null hypothesis test is appropriate. If it is, the second decision is whether the variable-criteria SSR is appropriate. The IACUC is composed by law of persons in several roles, including the chair, veterinarian, practicing scientist, and community member, but a statistician is not required. Nonscientist members are unlikely to be able to evaluate the suitability of a statistical design presented in an IACUC protocol application, and veterinarians rarely have a strong background in statistics. Recruitment of a scientist with strength in basic statistical methods will greatly aid the overall evaluation of the justification of animal numbers during the review of a protocol.

If a null hypothesis significance test is appropriate for a study, the most efficient way to assure that no more than the necessary number of animals will be used is to use a SSR. The variable-criteria SSR is an improvement over earlier SSR, COAST, and CLAST, which also control type I errors to less than α but do not hold them constant.[3,11,21] SSR that include an upper bound, such as CLAST and the variable-criteria SSR, allow the determination

Table 1. Critical *P* values to maintain $\alpha = 0.05$ for SSR tests

No. of tests	Šidák[13]	Armitage and McPherson[13]	Pocock[17]	Variable-criteria SSR	Sample-size models for variable-criteria SSR[a]
1	0.05	0.05	0.05	0.05	not applicable
2	0.025	0.03	0.0294	0.03	4/10 – 4 5/12 – 5 6/12 – 6 7/14 – 7
5	0.01	0.016	0.0158	0.025	3/15 – 3
10	0.005	0.011	0.0106	0.019	9/36 – 3[b]
20	0.003	0.008	0.0075	0.015	10/30 – 1[b]

The tests by Šidák, Armitage and McPherson, and Pocock all require that the lower bound is equal to the *n* added. Direct comparison is difficult because the variable-criteria SSR also included an upper criterion. Having an upper criterion means that many tests under the null hypothesis will be stopped before they progress to type I errors, thus reducing the overall error rate compared with the other methods and allowing a more generous critical *P* value.

[a]Data given as lower bound / upper bound – *n* added.

[b]Note that the lower bound and *n* added are not equal for 10 and 20 tests.

of the absolute maximal number of subjects that will be required for a study. This number will be large compared with the actual number of animals that are likely to be used in the study because experiments rarely reach the upper bound whether or not the null hypothesis is true. Nevertheless, the IACUC should assign animals to the protocol based on the upper bound as long as the investigator commits to the use of the SSR for a study in the protocol. If the investigator is using the variable-criteria SSR, the method itself will limit the number of animals to the minimum necessary to achieve significance. Because of that feature, the IACUC does not have to micromanage how many animals are assigned to the protocol. The IACUC can be assured that the experiment will be stopped early, on average, before many subjects have been used if an effect does not exist and that the investigator will add subjects to the experiment only as long as the interim data are promising according to an established rule.

This procedure will probably lead to an accumulation of excess animals on some protocols over the approval period of 3 y. When investigators apply for significant changes to add new experiments to the protocol, they must justify the number of animals in the significant change. The investigator can be asked at this point if these new animals must actually be added to the protocol. In some cases that will not be necessary. If the IACUC requires more control of animal purchasing, animals can be released administratively in stages based on interim reports[16] demonstrating that results are in the 'uncertain' range.

The use of the variable-criteria SSR simplifies the process of predicting a standardized effect size. The lower bound of the model selected for the SSR will be a relatively modest sample size so that larger-than-expected effects will be discovered early with few subjects. The upper bound of the model will allow smaller-than-expected effects to be pursued as long as they represent meaningful effects. For that reason, the upper bound should be close to the sample size required to detect the smallest important effect. With proper consideration to the selection of the lower and upper bounds, the procedure itself regulates the number of animals used within a range of sample sizes that will detect a range of meaningful effects.

References

1. **Bassler D, Briel M, Montori VM, Lane M, Glasziou P, Zhou Q, Heels-Ansdell D, Walter SD, Guyatt GH; STOPIT-2 Study Group, Flynn DN, Elamin MB, Murad MH, Abu Elnour NO, Lampropulos JF, Sood A, Mullan RJ, Erwin PJ, Bankhead CR, Perera R, Ruiz Culebro C, You JJ, Mulla SM, Kaur J, Nerenberg KA, Schünemann H, Cook DJ, Lutz K, Ribic CM, Vale N, Malaga G, Akl EA, Ferreira-Gonzalez I, Alonso-Coello P, Urrutia G, Kunz R, Bucher HC, Nordmann AJ, Raatz H, da Silva SA, Tuche F, Strahm B, Djulbegovic B, Adhikari NK, Mills EJ, Gwadry-Sridhar F, Kirpalani H, Soares HP, Karanicolas PJ, Burns KE, Vandvik PO, Coto-Yglesias F, Chrispim PP, Ramsay T.** 2010. Stopping randomized trials early for benefit and estimation of treatment effects. Systematic review and meta-regression analysis. JAMA **303:**1180–1187.
2. **Boneau CA.** 1960. The effects of violations of assumptions underlying the *t* test. Psychol Bull **57:**49–64.
3. **Botella J, Ximenez C, Revuelta J, Suero M.** 2006. Optimization of sample size in controlled experiments: The CLAST rule. Behav Res Methods **38:**65–76.
4. **Cohen J.** 1988. Statistical power analysis for the behavioral sciences, 2nd ed. Hillsdale (NJ): Erlbaum.
5. **DanielSoper.com.** [Internet]. Statistics Calculators. Version 2.0. [Cited 13 July 2010]. Available at: http://www.danielsoper.com/statcalc/default.aspx.
6. **Faul F, Erdfelder E, Lang A-G, Buchner A.** 2007. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods **39:**175–191.
7. **Fitts DA.** 2010. Improved stopping rules for the design of efficient small-sample experiments in biomedical and biobehavioral research. Behav Res Methods **42:**3–22.
8. **Fitts DA.** 2010. The variable-criteria sequential stopping rule: generality to unequal sample sizes, unequal variances, or to large ANOVA. Behav Res Methods **42:**918–929.
9. **Fitts DA.** 2011. Ethics and animal numbers: informal analyses, uncertain sample sizes, inefficient replications, and type I errors. J Am Assoc Lab Anim Sci [In press].
10. **Frick RW.** 1996. The appropriate use of null hypothesis testing. Psychol Methods **1:**379–390.

11. **Frick RW.** 1998. A better stopping rule for conventional statistical tests. Behav Res Methods **30:**690–697.
12. **Greenwald AG, Gonzalez R, Harris RJ, Guthrie D.** 1996. Effect sizes and *P* values: What should be reported and what should be replicated? Psychophysiology **33:**175–183.
13. **Ludbrook J.** 2003. Interim analyses of data as they accumulate in laboratory experimentation. BMC Med Res Methodol **3:**15.
14. **Maxwell SE, Kelley K, Rausch JR.** 2008. Sample size planning for statistical power and accuracy in parameter estimation. Annu Rev Psychol **59:**537–563.
15. **Meehl PE.** 1967. Theory-testing in psychology and physics: a methodological paradox. Philos Sci **34:**103–115.
16. **Office of Laboratory Animal Welfare, Applied Research Ethics National Association.** 2002. Institutional animal care and use committee guidebook, 2nd ed. Bethesda (MD): Office of Laboratory Animal Welfare
17. **Pocock SJ.** 1977. Group sequential methods in the design and analysis of clinical trials. Biometrika **64:**191–199.
18. **Rudser KD, Emerson SS**. 2008. Implementing type I and type II error spending for two-sided group sequential designs. Contemp Clin Trials **29**:351–358.
19. **Thalheimer W, Cook S.** [Internet]. 2002. How to calculate effect sizes from published research articles: a simplified methodology. [Cited 21 February 2010]. Available at: http://work-learning.com/effect_sizes.htm.
20. **Welch BL.** 1947. The generalization of 'Student's' problem when several different population variances are involved. Biometrika **34:**28–35.
21. **Ximenez C, Revuelta J.** 2007. Extending the CLAST sequential rule to one-way ANOVA under group sampling. Behav Res Methods **39:**86–100.