Behavior Research Methods, Instruments, & Computers
1998, 30 (4), 690-697

A better stopping rule for conventional statistical tests

ROBERT W. FRICK
State University of New York, Stony Brook, New York

The goal of some research studies is to demonstrate the existence of an effect. Statistical testing, with p less than .05, is one criterion for establishing the existence of this effect. In this situation, the fixed-sample stopping rule, in which the number of subjects is determined in advance, is impractical and inefficient. This article presents a sequential stopping rule that is practical and about 30% more efficient: Once a minimum number of subjects is tested, stop with p less than .01 or greater than .36; otherwise, keep testing. This procedure keeps alpha at .05 and can be adjusted to fit researchers' needs and inclinations.

Suppose that a researcher is testing whether one variable influences another or is comparing two treatments or effect sizes. Suppose also that the researcher is primarily interested in whether there is an effect (or difference) and, if so, the direction of the effect. Finally, suppose that the researcher wants to set alpha at .05, which is to say, have a 5% probability of rejecting the null hypothesis of no effect, if it is true.

Using the *fixed-sample* stopping rule, the researcher would determine the number of subjects to be tested prior to performing the study. However, there is another type of stopping rule, called *sequential stopping rules*, which was first proposed by Wald in 1947. In a sequential stopping rule, the outcome of the statistical test could lead to testing more subjects. Thus, the number of subjects to be tested is not fixed in advance.

The fixed-sample stopping rule achieves an alpha of .05 by having one statistical test and rejecting the null hypothesis when $p < .05$. A sequential stopping rule allows multiple statistical tests. It achieves an alpha of .05 by requiring a lower value of p for any one test, so that the overall alpha is .05.[1]

Inefficiency of the Fixed-Sample Stopping Rule

Power is usually defined as the probability of rejecting the null hypothesis for a given size of effect, and one stopping rule is more efficient than another when it achieves the same power with fewer subjects. Sequential stopping rules are more efficient than the fixed-sample stopping rule. For example, Wald (1947) reported that his sequential stopping rule was 50% more efficient than the fixed-sample stopping rule. The fixed-sample stopping rule is inefficient in three ways.

The first source of inefficiency is not stopping early when statistical significance is nearly assured. Suppose a researcher decided to test 80 subjects but, after testing 40 subjects, found that $p < .001$. Testing 40 more subjects just to fulfill the obligation of the fixed-sample stopping rule is inefficient. However, stopping early in this situation raises alpha, so it is considered inappropriate.

The second source of inefficiency is not stopping early when statistical significance is unlikely. Suppose that, after testing 40 subjects, there was no hint of an effect, so that statistical significance was unlikely to be achieved with the planned 80 subjects. Again, testing 40 more subjects just to fulfill the obligation of the fixed-sample stopping rule is inefficient. Stopping early in this situation lowers alpha, so it is not inappropriate, but it still deviates from the assumptions of the fixed-sample stopping rule.

The third source of inefficiency is the most serious—ending up with a small but statistically nonsignificant value of p, such as .06. Assuming a criterion of $p < .05$, these values of p cannot be used to reject the null hypothesis. Furthermore, small values of p are not appropriate for accepting the null hypothesis either (Frick, 1995). Thus, if the researcher were forced to stop at this point—as would be dictated by the fixed-sample stopping rule—the study would not be good evidence for anything.

The Impracticality of the Fixed-Sample Stopping Rule

Suppose that a researcher is using the fixed-sample stopping rule and finds that $p = .06$. The researcher could decide to improve the methodology and perform a new study. But suppose that the researcher decides the methodology is perfect but that there just were not enough subjects. What should this researcher do?

The researcher is not supposed to test more subjects, because that raises alpha. Some people (e.g., Goodman & Royall, 1988) would say that the researcher cannot perform the same study again, because that essentially doubles the overall alpha—the researcher now has two chances to achieve statistical significance just by chance. But *someone* should presumably do this study again, and

I thank Ira Bernstein, Peter Dixon, Arthur Samuel, and anonymous reviewers for their suggestions. Address correspondence to R. W. Frick, Department of Psychology, SUNY at Stony Brook, Stony Brook, NY 11794-2500 (e-mail: rfrick@sunysb.edu).

there is no reason why that someone cannot be the original researcher. However, performing two identical studies is inefficient; it is more accurate and efficient to combine the studies. In other words, it is more efficient for the researcher to keep testing subjects than to start a new study. Thus, the only practical response to $p = .06$ is to test more subjects. In sum, because of the impracticality of the fixed-sample stopping rule, researchers sometimes find themselves in a situation where the only rational response is not to follow it.

The Problem of Knowing Alpha

A stopping rule is defined, not just by what the researcher did, but also by what the researcher would have done. Suppose that, in one study in which a researcher decided to test 40 subjects, p was less than .05 and that the researcher stopped. The researcher might think that the fixed-sample stopping rule was used. However, suppose that in a second study, 40 subjects were tested, p was less than .10 but greater than .05, and the same researcher decided to test 20 more subjects. Obviously, the fixed-sample stopping rule was not used in the second study. The point is that the fixed-sample stopping rule apparently was not used in the first study either—it seems likely that the researcher would have tested more subjects if p had been less than .10 but greater than .05.

Alpha is determined by the stopping rule, so when the stopping rule is unknown, alpha is also unknown. Alpha is obviously larger than .05 in the second study, because there were two opportunities to achieve $p < .05$. However, the first study also had two opportunities to achieve $p < .05$—if the value of p had been close to .05, the researcher apparently would have tested more subjects and performed another statistical test.

My impression is that researchers do some things that raise alpha, such as stopping early with $p < .05$ or testing more subjects than planned when the value of p is small but greater than .05. However, researchers do other things that *lower* alpha. These include stopping early when p is large, adopting a personal criterion less than .05 for p, and performing a pilot study and not running the study if the result of the pilot study is discouraging. As a consequence, neither researchers nor readers know (1) the stopping rules used by researchers or (2) the cumulative effect on alpha of researchers' deviations from the fixed-sample stopping rule.

Existing Sequential Stopping Rules

A number of different sequential tests have been developed. Unfortunately, many of these choose between two possible effect sizes that must be specified in advance (see, e.g., Arghami & Billard, 1982, 1991; Billard & Vagholkar, 1969; Fiske & Jones, 1954; Park, 1992; Sobel & Wald, 1949; Wald, 1947). For example, in a study testing the effects of hunger on irritability, as measured by some irritability scale, the sequential test might choose between an effect size of 0 and an effect size of 10. (In practice, this is a choice between an effect size of 0 or less and of 10 or more.) In most studies, the choice is not between two values (e.g., Neyman, 1950, p. 324); instead, the choice is between no effect, an effect of unspecified size in one direction, or an effect of unspecified size in the other direction. For example, the researcher might choose between hunger having no effect on irritability, hunger increasing irritability, or hunger decreasing irritability.

An exception to this is the practical application study. The researcher can try to identify a minimal size of effect needed for clinical or practical significance. The sequential test can then be framed as choosing between no effect and minimal clinical significance or larger. Viable sequential stopping rules exist for this situation (see, e.g., DeMets & Lan, 1984; Doll, 1982; Pocock, 1992).

The utility of many sequential stopping rules is also limited by having an upper bound to the number of subjects that might be tested (see, e.g., Arghami & Billard, 1982; Armitage, 1957; Siegmund, 1985). Setting an upper bound has the same primary problem as the fixed-sample stopping rule. Suppose that when a researcher reaches an upper bound of 100 subjects, the value of p is small, but not small enough to allow rejecting the null hypothesis. The researcher will not be happy to stop at this point without being able to reject the null hypothesis. One rationale offered for an upper bound is that no researcher is willing to test an infinite number of subjects (Whitehead & Brunier, 1990). However, most researchers would be willing to test a few more subjects, if p was almost small enough to allow rejecting the null hypothesis. Also, when researchers reached the upper bound, they could be allowed to reject the null hypothesis with higher values of p, such as $p = .10$, without necessarily increasing alpha beyond .05. However, I doubt that the enterprise of psychology would tolerate rejecting the null hypothesis with values of p greater than .05 just because the researcher happened to select an upper bound that was reached.

I could not find any stopping rules that did not either choose between two effect sizes or have an upper bound to the number of subjects that were tested. Thus, there seems to be no sequential stopping rules that would suit the situation faced by the researcher primarily interested in establishing the existence of an effect.

COAST

In the sequential stopping rule I am proposing, the researcher can perform a statistical test at any time. If the outcome of this statistical test is $p < .01$, the researcher stops testing subjects and rejects the null hypothesis; if $p > .36$, the researcher stops testing subjects and does not reject the null hypothesis; and if $.01 < p < .36$, more subjects are tested. I will call this stopping rule COAST, for composite open adaptive sequential test—it has a

composite alternative to the null hypothesis, it is open because it has no upper bound, and it is adaptive to a wide range of situations.

The value of .01 will be called the *low criterion*, and .36 will be called the *high criterion*. The value .36 was determined (from to-be-reported Monte Carlo simulations) to be approximately the value that would balance .01 to produce an alpha of .05. The value of .01 was selected as one that probably would be best suited to the needs of researchers. A smaller value made it too difficult to reject the null hypothesis. A larger value required a smaller high criterion, and it seemed unlikely that researchers would want to stop at p values much smaller than .36.

There is always some minimum number of subjects that the researcher will want to test (Billard & Vagholkar, 1969). To pick an extreme example, most researchers would not want to stop testing with 4 subjects just because the current value of p was less than .01. COAST assumes that the researcher has a lower bound for the number of subjects to test before stopping at $p < .01$. The only requirement for this lower bound is that the researcher should also be willing to stop at $p > .36$ once this lower bound is reached. There is no upper bound, so the researcher is never forced to stop with a marginally nonsignificant value of p.

One goal for COAST was, of course, efficiency. A second goal was to construct a stopping rule that would be easy for researchers to use. The rules for COAST are simple, and the decisions are based on p, which researchers already know how to calculate. I also tried to construct COAST so that it mimicked the decisions a researcher might naturally make.

COAST is built to be used in a particular statistical niche. First, the researcher is asking a focused question about one particular relationship between variables. Thus, there is one critical statistical test for the study, for which the researcher desires statistical significance (if there is an effect). Second, the researcher is primarily interested in the presence and direction of an effect, not its size. Third, the researcher can test more subjects if p is between .01 and .36. Fourth, the researcher wants to set alpha at .05 (for a two-tailed test). Existing sequential stopping rules address other niches, such as clinical trials (see, e.g., DeMets & Lan, 1984; Hwang, 1992; Jennison & Turnbull, 1991; Pocock, 1992; Proschan, Follmann, & Waclawiw, 1992), estimating reliability (see, e.g., Eiting, 1991), testing individual differences (see, e.g., Linn, Rock, & Cleary, 1969, 1972), and psychophysics (see, e.g., Hoffman, 1992).

ALPHA

Monte Carlo Estimations of Alpha

This article will report several Monte Carlo simulations of COAST. In a Monte Carlo simulation, a computer is used to generate scores with a given distribution. The goal was to discover the behavior of COAST by observing its behavior in a very large number of simulated experiments. The simulated experiments match what would happen in real experiments, assuming, of course, that the process of producing subjects' actual scores matched the assumptions of the Monte Carlo simulation.

Unless otherwise noted, these simulations assumed a within-subjects design with two conditions. The difference between conditions for each subject was constructed to be normally distributed. The statistical test was a two-tailed t test, using the null hypothesis that the difference between conditions was 0. The value of t needed for $p = .36$ was computed from a linear average for the t values corresponding to $p = .30$ and $p = .40$.

Table 1
Alpha for Different Lower Bounds of COAST

Lower Bound	Average N	Alpha (%)
10	14.3	4.7
20	27.5	4.8
40	52.4	4.7
80	102.4	5.1

The first issue is determining the value of alpha for COAST. The simulations reported in Table 1 contained 20,000 "experiments," yielding a standard error of the mean of 0.15% for the estimation of alpha. The estimated alpha is reported in Table 1, for different lower bounds. The results of this simulation suggest that alpha is approximately .05. The average number of subjects that would be tested is also given in Table 1.

The mathematics of the situation suggests that alpha will increase slightly as the lower bound increases. Suppose, for a moment, that whenever a new subject was tested, the value of p was independent of the previous value of p. Simple algebra reveals that COAST would have an alpha of approximately 2.7%. In actual studies, the value of p after testing a new subject is strongly correlated with the previous value of p. The correlation between successive values of p increases alpha. When fewer subjects have been tested, there is a smaller correlation between successive values of p, because a new subject has a larger effect on p. Therefore, having a small lower bound tends to decrease alpha.

Fortunately, the change in alpha is small. The difference in alpha from testing 10 subjects to testing 80 subjects is only 0.4%. Therefore, for the purposes of COAST, there is no practical effect of varying when the possible stopping begins.

Other Statistical Tests

This article focuses on Monte Carlo simulations of a within-subjects t test. However, because COAST works with p, its mechanisms can be applied to any statistical test. In a simulation of Pearson's r test of correlation, using 10,000 experiments, alpha was estimated to be approximately 4.6%.

EFFICIENCY

Monte Carlo Simulations of Power

Using the same Monte Carlo procedure, the power of COAST was calculated for effect sizes of .2, .4, and .6 standard deviations. An effect size of .2 standard deviations means that the true difference between conditions was .2 of the standard deviation of differences within conditions. In this simulation and all of the remaining simulations, there were 10,000 "experiments" in the simulation. This yields a standard error of the mean of 0.22% for the estimation of alpha and a standard error of the mean less than 0.5% for the estimate of power. This simulation examined three different lower bounds, at 20, 40, and 60 subjects. The average number of subjects and the power are presented in Table 2.

Table 2 also compares COAST to the fixed-sample stopping rule. For the sake of greater accuracy, the theoretically expected performance of the fixed stopping rule was used. (A few Monte Carlo simulations of the fixed-sample stopping rule produced approximately the same answers as those that were theoretically expected.)

Table 2 presents the number of subjects needed in the fixed-sample stopping rule to achieve the same power as COAST. Efficiency of COAST was calculated as the percentage of savings in number of subjects, on the basis of the expected number of subjects for the fixed-sample stopping rule. As is shown in Table 2, COAST usually required 26%–30% fewer subjects to achieve the same power.

In the comparison of COAST with the fixed-sample stopping rule, the efficiency of the fixed-sample stopping rule was being slightly overestimated. COAST tested slightly fewer subjects for obvious effects and those for which the null hypothesis was correct, which are the optimal situations for testing only a few number of subjects. Correspondingly, it tested slightly more subjects when there was a small effect to be found, which is the optimal situation for testing many subjects. Thus, to a small extent, COAST adapted to the situation without knowing the situation in advance. The fixed-sample stopping rule, of course, does not adapt to the situation. However, being yoked to COAST gave it the benefit of COAST's adaptability.

Table 2
Power of COAST and a Comparison With the Fixed-Sample Stopping Rule

Effect Size	Lower Bound	*N*	Power (%)	*N* of Fixed	Efficiency (%)
.2	20	32.9	28.4	47	+30
.2	40	63.0	45.7	85	+26
.2	60	90.5	59.4	122	+26
.4	20	30.5	70.8	39	+22
.4	40	49.4	91.6	69	+28
.4	60	66.3	97.4	97	+32
.6	20	25.1	93.8	34	+26
.6	40	41.3	99.6	59	+30
Mixed	20	29.0	64.3	40	+28
Mixed	40	51.5	79.0	78	+34

To more fairly compare COAST with the fixed-sample stopping rule, I determined the overall performance of COAST for a mixture of four effect sizes: 0, .2, .4, and .6. (This involved simply combining the results presented in Tables 1 and 2.) The average number of subjects and the power of COAST when there was an effect is also presented in Table 2. This was then compared with the number of subjects needed to accomplish the same power, using the fixed-sample stopping rule. COAST with a lower bound of 20 subjects required 28% fewer subjects; COAST with a lower bound of 40 required 34% fewer subjects. Thus, on the average, COAST required 31% fewer subjects than the fixed-sample stopping rule to have the same probability of finding an effect.

The efficiency of COAST is somewhat smaller than those reported for other sequential stopping rules. For example, Wald's (1947) sequential stopping rule required 50% fewer subjects to achieve the same power as a fixed-sample stopping rule, and Fiske and Jones (1954) suggested savings of 33%–50%. However, they assumed a well-defined alternative to the null hypothesis, whereas COAST does not. It may be that more savings are possible when only one well-defined alternative to the null hypothesis is assumed.

Stopping at $p < .05$

A researcher might worry that the criterion of $p < .01$ in COAST is too difficult. However, the seeming difficulty of achieving $p < .01$ in COAST is illusory. Despite requiring $p < .01$, COAST actually needs, on the average, 30% fewer subjects than the fixed-sample stopping rule needs to have a given probability of achieving $p < .05$.

However, suppose that a researcher is using a stopping procedure identical to COAST, except that the researcher stops as soon as $p < .05$. Obviously, doing so makes it easier to achieve statistical significance. However, this procedure raises alpha too much: In a Monte Carlo simulation with a lower bound of 20, raising the low criterion to .05 produced an alpha of 13.1%, which is inappropriately large.

CUSTOMIZING COAST

The researcher has several choices to make when using COAST. I think of these choices as allowing the researcher to customize COAST to better fit the researcher's circumstances and inclinations.

Choosing the Lower Bound

The choice of a lower bound influences the number of subjects tested. An effect does not have to be statistically significant at the lower bound in order for the null hypothesis to be eventually rejected. However, it does have to produce $p < .36$ by the time the lower bound is reached.

Table 3
Alpha for Spacing the Calculation of *p*

Frequency	Lower Bound	*N*	Alpha (%)
Every subject	20	28.3	4.8*
Every 4 subjects	20	32.1	4.2
Every 10 subjects	20	36.0	3.8
Every subject	40	52.4	4.7*
Every 4 subjects	40	57.8	4.4
Every 10 subjects	40	66.7	4.2

*Taken from Table 1.

Table 4
Power for Unspaced Versus Spaced Stopping

Frequency	Lower Bound	*N*	Power (%)	Efficiency (%)
Effect Size .2				
Every subject	40	63	45.7	+26*
Every 4 subjects	40	69.8	48.1	+23
Every 10 subjects	40	77.2	50.7	+20
Effect Size .4				
Every subject	40	49.4	91.6	+28*
Every 4 subjects	40	51.5	91.8	+26
Every 10 subjects	40	54.4	92.5	+24

*Taken from Table 2.

This is more likely to occur with larger lower bounds. Therefore, larger lower bounds have a larger probability of finding an effect but cause more subjects to be tested.

One factor influencing a researcher's lower bound is experience. Sometimes a researcher will know the approximate number of subjects needed to find a well-established effect or the approximate number of subjects used by other researchers in an area. A second factor is the ease of testing subjects. Suppose subjects are difficult to find or test. The researcher will be more inclined to stop testing with only a few subjects and less inclined to test many subjects.

Spaced Stopping

Alpha and power were calculated on the basis of the assumption that a decision of whether or not to stop could occur after each subject. However, when p is within range of .01 or .36, it might be inconvenient to calculate p after each subject. Furthermore, the demands of counterbalancing might not allow stopping at any point.

Whenever the researcher cannot stop as soon as p is less than .01 or greater than .36, I will say that the opportunities to stop are *spaced*. Spacing leads to what is called a *group sequential stopping rule* (Pocock, 1977). The results of a Monte Carlo simulation for spacing in COAST are presented in Table 3. These simulations assumed that p was calculated after either every 4th or every 10th subject.

Spacing seems to decrease alpha somewhat. As already noted, alpha is influenced by the amount p is likely to change from one opportunity to stop to the next. When opportunities to stop are spaced, slightly more change in p is expected from one opportunity to stop to the next, reducing alpha.

When the spacing is very large, alpha can be considerably lowered. In this case, it might be desirable to compensate for the reduction in alpha (Pocock, 1977). However, there are several reasons for not compensating for changes in alpha in COAST. First, for most studies, the spacing rarely needs to be very large. Second, it is relatively complex to calculate the amount the criteria should be changed to adjust alpha for different sizes of spacing. Third, by maintaining the low criterion of $p < .01$, it is difficult for researchers to violate the assumptions of the stopping rule. They cannot claim to have used spacing in order to raise their low criterion. Instead, their choice is to check after every subject or accept the small "penalty" of a slightly lower alpha.

Spacing increases the number of subjects to be tested. Therefore, it slightly increases power, but it slightly lowers efficiency. These changes in efficiency are shown in Table 4.

As noted, COAST might be suitable for any statistical test, because of its use of p. A Monte Carlo simulation was used to estimate the value of alpha for the interaction in a 2×2 analysis of variance (ANOVA), assuming a between-subjects design and that stopping could occur after every 4 subjects, which is to say, when all four groups had the same number of subjects. Alpha ranged from 4.0% for a lower bound of 20 subjects to 4.6% for a lower bound of 80 subjects.

Not Stopping With $p > .36$

If p was .36 and the trend continued, roughly four times as many subjects as those already tested would be needed to reach $p < .01$. Unless a researcher is willing to test four times as many subjects as currently tested, the researcher would want to stop at this point. But what if the researcher does not want to stop? Continuing despite $p > .36$ (when the lower bound is reached) violates COAST and raises alpha. Table 5 presents alpha for a simulation of COAST, using a larger high criterion.

These simulations demonstrate that having a high criterion of .40 or .50 does not substantially raise alpha. In his review of the robustness of the t test against violations of assumptions, Boneau (1960) called an increase of alpha to 7.8% a "small effect." Thus, the increase in alpha to 6.3%, produced by not stopping until p is greater than .50, probably would be tolerable. Additionally, this increase in alpha might be balanced by the use of spacing.

A high criterion of .75 increases alpha too much and is inappropriate. However, if p was .50 and the present trend continued, it would take, roughly, about nine times as many subjects to achieve $p < .01$. Researchers are very unlikely to want to test this many subjects beyond

Table 5
Alpha for Upper Criterions Greater Than .36

High Criterion	Lower Bound	Alpha (%)	*N*
.40	20	5.4	32.0
.40	40	5.3	61.3
.50	20	6.3	56.8
.50	40	6.3	98.3
.75	20	8.4	319.8
.75	40	9.4	527.7

their lower bound; they are more likely to (1) accept the null hypothesis, (2) try to modify their study to be more powerful and try a fresh start, or (3) investigate some other more promising hypothesis. Therefore, researchers would not be very tempted to deviate from COAST in a way that would substantially increase alpha.

Stopping Early

Suppose a researcher set a lower bound, such as 40 subjects, then "peaked" early, say at 20 or 30 subjects. If the value of p was very large or very small, the researcher might be inclined to stop early. Stopping early with a very small value of p is known as the Haybittle–Peto rule (Haybittle, 1971; Peto et al., 1976).

Stopping early when p is .001 would raise alpha; hence, strictly speaking, it would be inappropriate. However, the effect on alpha is negligible. A value of $p <$ .001 does not occur often when the null hypothesis is true, and when it does, p might still be less than .01 when the lower bound is reached.

To investigate the effects of stopping early, COAST was simulated with a lower bound of 40 subjects. Peaking occurred, beginning with the 20th subject, and the experiment stopped early whenever p was less than .001. The simulation then continued to discover what would have happened had there been no early stop. When the null hypothesis was correct, an early stop at $p < .001$ occurred on only 50 "experiments" out of 10,000. The null hypothesis would have been rejected on 33 of these experiments without the early stop. Thus, the early stopping at $p < .001$ increased alpha only 0.17%.

It is also more efficient to stop when it seems unlikely that the null hypothesis will be rejected (Lan, Simon, & Halperin, 1982). In the above simulation of early stopping, the experiment also stopped early if the number of subjects was greater than 20 and p was greater than .75. This happened quite frequently, on 46% of the experiments. This had little effect on alpha, because on only four of these early stops would COAST have rejected the null hypothesis had there been no early stop. Thus, the decrease in alpha was 0.04%, leading to a cumulative change for early stopping of increasing alpha 0.13%. The results of this simulation are presented in Table 6.

I also simulated early stopping for effect sizes of .2, .4, and .6. These results are included in Table 6. Stopping early reduced the average number of subjects tested and power, but the reduction in power was very small, because the eventual decision of COAST was usually unchanged by the early stop. To compute overall efficiency, I again assumed that there was an even mixture of effect sizes of .0, .2, .4, and .6. The overall performances are presented in Table 7, comparing early stops with what would have happened without early stops. The early stops served to increase efficiency.

Table 6
Effects of Early Stopping at $N = 20$

Effect Size	Savings in N	Change in Power/Alpha %
0	4.2	−.1
.2	2.9	−.4
.4	3.3	−.5
.6	8.6	0

Table 7
Overall Effect of Early Stopping

	N	Alpha (%)	Power (%)	N of Fixed	Efficiency (%)
No early stops	51.7	5.2	79.2	79	+35
Early stops	46.9	5.1	78.9	78	+40

Table 8
Effects of Giving Up at $N = 100$ and $p > .10$

Effect Size	Savings in N	Change in Power/Alpha %
0	2.6	−.2
.2	1.8	−1.8
.4	0.0	0
.6	0.0	0

Table 9
Overall Effect of Giving Up

	N	Alpha (%)	Power (%)	N of Fixed	Efficiency (%)
No giving up	29.4	4.6	64.4	40	+27
Giving up	28.2	4.4	63.8	39	+28

Giving Up

COAST does not specify an upper bound to the number of subjects that can be tested, and, in calculating alpha and power for COAST, I assumed that the researcher would continue testing subjects until a "decision" was reached. In practice, researchers will sometimes "give up," stopping testing even though the value of p is between .01 and .36. Giving up is always allowable, because it reduces alpha. Giving up reduces the probability of finding an effect, but it also reduces the expected number of subjects to be tested.

For a Monte Carlo simulation of giving up, I assumed that there was a lower bound of 20 subjects and that the researcher would want to give up at 100 (or more) subjects if the value of p was greater than .10. As is shown in Tables 8 and 9, giving up did not occur often. It had a small effect on alpha and power, and it increased efficiency.

Secondary Analyses

Researchers often perform secondary analyses after a study is done, addressing additional issues. The researcher is unlikely to test more subjects for a secondary analysis, so it seems appropriate to use the fixed-sample stopping rule for these analyses. (Furthermore, it currently seems acceptable to present marginally significant results for these secondary analyses, as long as they are then interpreted with caution.)

EVALUATION

Meeting the Needs of Researchers

COAST meets the needs of researchers. First, it is efficient. This efficiency would be useful when testing

subjects was difficult, expensive, or time consuming or when, for ethical reasons, it would be desirable to test as few subjects as is necessary.

Second, as Edwards, Lindman, and Savage (1963, p. 239) note, "Many researchers would like to feel free to collect data until they have either conclusively proved their point, conclusively disproved it, or run out of time, money, or patience." COAST allows researchers to do this.

Third, COAST was designed to mimic as closely as possible the decisions a researcher would make on an ad hoc basis (though it does avoid the inefficient choices a researcher might make). The result is a conglomeration of techniques compatible with researchers' needs and inclinations.

Having the criteria of .01 and .36 remain constant is not the most efficient technique (Weiss, 1953). Instead, it is more efficient for the criteria to be more extreme (e.g., .001 and .75) for fewer subjects and less extreme (e.g., .03 and .20) for more subjects. Different methods of adjusting the criteria have been proposed (see, e.g., Lan, DeMets, & Halperin, 1984; O'Brien & Fleming, 1979; Slud & Wei, 1982). However, adjusting the high and low criteria requires some relatively sophisticated calculations. COAST adopts the simple procedure of leaving these criteria constant.

Technical Feasibility

When sequential stopping rules were first proposed, statistical tests had to be computed by hand. Needless to say, it would be have been very tedious to perform the multiple statistical tests allowed by a sequential stopping rule. Nowadays, when statistical tests are performed by computer, this would be relatively easy.

I have assumed that the researcher uses a computer program that reports the exact value of p. If not, the value of p must be estimated from a table. Published tables will not contain the test values corresponding to p equals .36, but the average of .20 and .50 works well (as suggested by Monte Carlo simulations not reported in this article). In any case, COAST is not sensitive to the exact value of the high criterion, so the researcher can use a rough estimate for p at the high criterion.

It is not immediately obvious how a researcher who used COAST would present a confidence interval for the estimated effect size. A conservative approach would be to calculate a 99% confidence interval on the basis of the observed value of p, then report it as a 95% confidence. (This is analogous to achieving $p < .01$ to accomplish an alpha of less than .05.) A researcher who was interested just in reporting a confidence interval and was not concerned with whether that confidence interval included any particular value would not need a stopping rule.

Meeting the Needs of Science

Psychology as a science apparently wants to maintain alpha at .05, and researchers seem honest about trying to keep alpha at the desired .05 level. However, the impracticalities and inefficiencies of the fixed-sample stopping rule make this difficult. Researchers might deviate from the fixed-sample stopping rule in ways that raise alpha and in ways that lower alpha, so that the alpha for the overall stopping rule is unknown.

It is also useful to have stopping rules that are difficult to violate. If a researcher can easily test more subjects, violations are easy with the fixed-sample stopping rule—a researcher can test until $p < .05$, then report the results. As noted, a researcher who sets a lower bound of 20 subjects and then does not stop until p is less then .05 or greater than .36 has a functional alpha of .13.

COAST avoids these problem. It is relatively resistant to violations, because it has the unchanging criterion of $p < .01$. Researchers would have little reason not to follow the rules, and most deviations the researcher might want to make do not substantially change alpha. Furthermore, although some deviations raise alpha slightly, some lower it. Researchers could raise alpha to inappropriate levels by not stopping when p was greater than .50, but as noted, they would have little motivation for doing so.

CONCLUSION

The current state of affairs in psychology with regard to stopping rules is not good. Researchers are told that the only acceptable stopping rule is the fixed-sample stopping rule, even though the fixed-sample stopping rule is impractical, inefficient, and not the only acceptable stopping rule. Researchers should have a choice of stopping rules, so that they can choose the stopping rule that most efficiently fits their needs.

Because the fixed-sample stopping rule is impractical and inefficient, researchers should not be blamed for making adjustments to it. However, it would be better if researchers used well-analyzed stopping rules that were efficient and controlled alpha at about .05.

I proposed a new stopping rule, called COAST, suited to cases in which a psychologist's concern is demonstrating the existence of an effect. In terms of acceptability to the enterprise of psychology, COAST holds alpha at about .05 and is relatively resistant to violations. In terms of acceptability to researchers, COAST is easy to use and is about 30% more efficient than fixed-sample stopping rules; most of the procedures of COAST would fit the researcher's inclinations, and COAST can be customized to even better fit the researcher's needs and inclinations.

REFERENCES

ARGHAMI, N. R., & BILLARD, L. (1982). A modification of a truncated partial sequential procedure. *Biometrika*, **69**, 613-618.

ARGHAMI, N. R., & BILLARD, L. (1991). A partial sequential *t*-test. *Sequential Analysis*, **10**, 181-197.

ARMITAGE, P. (1957). Restricted sequential procedures. *Biometrika*, **44**, 9-26.

BILLARD, L., & VAGHOLKAR, M. K. (1969). A sequential procedure for testing a null hypothesis against a two-sided alternative hypothesis. *Journal of the Royal Statistical Society B*, **31**, 285-294.

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

DEMETS, D. L., & LAN, K. K. G. (1984). An overview of sequential methods and their application in clinical trials. *Communications in Statistics: Theory & Methods*, **13**, 2315-2338.

DOLL, R. (1982). Clinical trials: Retrospect and prospect. *Statistics in Medicine*, **1**, 337-344.

EDWARDS, W., LINDMAN, H., & SAVAGE, L. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, **70**, 193-242.

EITING, M. H. (1991). Sequential reliability tests. *Applied Psychological Measurement*, **15**, 193-205.

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

FRICK, R. W. (1995). Accepting the null hypothesis. *Memory & Cognition*, **23**, 132-138.

FRICK, R. W. (1996). The appropriate use of null hypothesis testing. *Psychological Methods*, **1**, 379-390.

GOODMAN, S. N., & ROYALL, R. (1988). Evidence and scientific research. *American Journal of Public Health*, **78**, 1568-1574.

HAYBITTLE, J. L. (1971). Repeated assessment of results in clinical trials of cancer treatment. *British Journal of Radiology*, **44**, 793-797.

HOFFMAN, H. S. (1992). An application of sequential analysis to observer-based psychophysics. *Infant Behavior & Development*, **15**, 271-277.

HWANG, I. K. (1992). Overview of the development of sequential procedures. In K. E. Bruce (Ed.), *Biopharmaceutical sequential statistical application* (pp. 3-18). New York: Dekker.

JENNISON, C., & TURNBULL, B. W. (1991). Group sequential tests and repeated confidence intervals. In M. Ghosh & P. K. Sen (Eds.), *Handbook of sequential analysis* (pp. 283-311). New York: Dekker.

LAN, K. K. G., DEMETS, D. L., & HALPERIN, M. (1984). More flexible sequential and non-sequential designs in long-term clinical trials. *Communication in Statistics: Theory & Methods*, **13**, 2339-2353.

LAN, K. K. G., SIMON, R., & HALPERIN, M. (1982). Stochastically curtailed tests in long-term clinical trials. *Communication in Statistics: Sequential Analysis*, **1**, 207-219.

LINN, R. L., ROCK, D. A., & CLEARY, T. A. (1969). The development and evaluation of several programmed testing methods. *Educational & Psychological Measurement*, **29**, 129-146.

LINN, R. L., ROCK, D. A., & CLEARY, T. A. (1972). Sequential testing for dichotomous decisions. *Educational & Psychological Measurement*, **32**, 85-95.

NEYMAN, J. (1950). *First course in probability and statistics*. New York: Holt.

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

PARK, C. (1992). An approximation method for the characteristics of the sequential probability ratio test. *Sequential Analysis*, **11**, 55-72.

PETO, R., PIKE, P., ARMITAGE, P., BRESLOW, N. E., COX, D. R., HOWARD, S. V., MANTEL, N., MCPHERSON, K., PETO, J., & SMITH, P. G. (1976). Design and analysis of randomized clinical trials requiring prolonged observation of each patient. *British Journal of Cancer*, **35**, 585-611.

POCOCK, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika*, **64**, 191-199.

POCOCK, S. J. (1992). When to stop a clinical trial. *British Medical Journal*, **305**, 235-240.

PROSCHAN, M. A., FOLLMANN, D. A., & WACLAWIW, M. A. (1992). Effects of assumption violations on Type I error rate in group sequential monitoring. *Biometrics*, **48**, 1131-1143.

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

SLUD, E., & WEI, L. J. (1982). Two sample repeated significance tests based on the modified Wilcoxon statistic. *Journal of the American Statistical Association*, **77**, 862-868.

SOBEL, M., & WALD, A. (1949). A sequential decision procedure for choosing one of three hypotheses concerning the unknown mean of a normal distribution. *Annals of Mathematical Statistics*, **20**, 502-522.

STERLING, T. D., ROSENBAUM, W. L., & WEINKAM, J. J. (1995). Publication decisions revisited: The effect of the outcome of statistical tests on the decision to publish and vice versa. *American Statistician*, **49**, 108-112.

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

WEISS, L. (1953). Testing one simple hypothesis against another. *Annals of Mathematical Statistics*, **24**, 273-281.

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

NOTE

1. I will assume that *p* is calculated in the standard way, as the probability of achieving the observed difference or more, given the null hypothesis and the number of subjects tested. The .05 criterion is widespread (see, e.g., Sterling, Rosenbaum, & Weinkam, 1995), but it has come under attack. Reasons for this attack and the appropriateness of this criterion are discussed in Frick (1996).

(Manuscript received January 15, 1997;
revision accepted for publication July 16, 1997.)