Behav Res (2018) 50:1988–2003
DOI 10.3758/s13428-017-0968-5

# Variable criteria sequential stopping rule: Validity and power with repeated measures ANOVA, multiple correlation, MANOVA and relation to Chi-square distribution

Douglas A. Fitts[1]

Published online: 21 September 2017

**Abstract** The variable criteria sequential stopping rule (vcSSR) is an efficient way to add sample size to planned ANOVA tests while holding the observed rate of Type I errors, $\alpha_o$, constant. The only difference from regular null hypothesis testing is that criteria for stopping the experiment are obtained from a table based on the desired power, rate of Type I errors, and beginning sample size. The vcSSR was developed using between-subjects ANOVAs, but it should work with *p* values from any type of *F* test. In the present study, the $\alpha_o$ remained constant at the nominal level when using the previously published table of criteria with repeated measures designs with various numbers of treatments per subject, Type I error rates, values of ρ, and four different sample size models. New power curves allow researchers to select the optimal sample size model for a repeated measures experiment. The criteria held $\alpha_o$ constant either when used with a multiple correlation that varied the sample size model and the number of predictor variables, or when used with MANOVA with multiple groups and two levels of a within-subject variable at various levels of ρ. Although not recommended for use with $\chi^2$ tests such as the Friedman rank ANOVA test, the vcSSR produces predictable results based on the relation between *F* and $\chi^2$. Together, the data confirm the view that the vcSSR can be used to control Type I errors during sequential sampling with any *t*- or *F*-statistic rather than being restricted to certain ANOVA designs.

**Keywords** Analysis of variance · Chi square · Effect size · Friedman test · MANOVA · Multiple correlation · Null hypothesis significance test · Power analysis · Reduction of subjects · Repeated measures · Sample size planning · Group sequential stopping rules

The goal of this article is to extend the use of the variable criteria sequential stopping rule (vcSSR; Fitts, 2010a, b, 2011a, b) to include a variety of experimental circumstances involving repeated measures from the same subjects. The vcSSR is a tool to help preclinical and behavioral researchers working with small sample sizes (< 40/group) to conserve sample size and to control the Type I error rate when conducting experiments using *t* tests or *F* tests. Sequential stopping rules (SSRs) are better known in clinical trials, where it has long been recognized that the completion of a large trial all the way to the expected ending sample size may be unethical. If the trial is doing harm, the experiment must be stopped early. If the treatment is performing better than expected, the trial should be stopped so the placebo group can receive the benefit of the new treatment. The method is efficient with sample size.

A problem arises when researchers use informal stopping procedures (one cannot call them "rules") instead of formal ones. Naïve use of informal stopping procedures contributes to the proliferation of false positive results in the literature and threatens the validity of statistical conclusions (García-Pérez, 2012).

As a simple example of the problem and its solution, suppose a researcher designs an experiment to discover the effect of a drug on a behavior. Two groups of subjects will be tested, one with the drug and one with a placebo. Instead of using the conventional method of testing all *N* subjects from the power analysis, the researcher naïvely

**Electronic supplementary material** The online version of this article (https://doi.org/10.3758/s13428-017-0968-5) contains supplementary material, which is available to authorized users.

✉ Douglas A. Fitts
dfitts@uw.edu

[1] Department of Psychology and Office of Animal Welfare (Retired), University of Washington, Seattle, WA 98195, USA

decides to use fewer subjects in hopes that the significant effect can still be detected. Perhaps the drug is expensive or the subjects are rare. If the first *t* test with the smaller sample size is not quite significant at the .05 level, the sample size can always be augmented gradually to include the full *N* subjects. This method is extremely efficient with sample size, because it is often successful if there is a true effect of the drug. What the naïve researcher does not realize is that the Type I error rate (e.g., $\alpha$ = .05) is easily inflated to double the expected rate. This inflation can readily be demonstrated in simulation studies. See Fitts (2010a, 2011a) for a detailed numerical example of how this rate becomes inflated.

In many circumstances, there may be ethical reasons to limit sample size as much as possible (Fitts, 2011a). In such cases, it is less obvious that a strict adherence to the formal rules of the null hypothesis test is the best choice. The solution is to use a SSR in conjunction with the hypothesis test to allow the addition of sample size in graded increments without inflating the rate of Type I errors in the experiment. A simple method to accomplish this is to reduce the criterion *p* value, like the Bonferroni correction for multiple contrasts, so that the Type I error rate does not exceed .05 regardless of what happens during the sequential testing (Frick, 1998).

The vcSSR is a method and a set of tables (vcSSR Tables: Table 2 in Fitts, 2010a; Table 1 in this paper, abridged; or full

**Table 1** Lower and upper criteria of the vcSSR for four sample size models and four levels of α used in this study, reprinted in part from 16 models in Fitts (2010a)

| | | Type I error rate α | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | .005 | | .010 | | .050 | | .100 | |
| Lower/upper bound model | *n Added* | Lower | Upper | Lower | Upper | Lower | Upper | Lower | Upper |
| 4/18 | 1 | .00086 | .35 | .002 | .2 | .015 | .26 | .035 | .36 |
| | 2 | .001 | .45 | .0025 | .219 | .015 | .41 | .045 | .3 |
| | 3 | .0013 | .39 | .003 | .31 | .025 | .19 | .045 | .39 |
| | 4 | .00156 | .425 | .004 | .15 | .03 | .16 | .055 | .3 |
| | 5 | .002 | .2 | .004 | .35 | .03 | .2 | .06 | .32 |
| | 6 | .002 | .26 | .004 | .39 | .025 | .35 | .06 | .33 |
| 6/18 | 1 | .00094 | .4 | .002 | .5 | .0165 | .24 | .039 | .33 |
| | 2 | .0012 | .32 | .0027 | .26 | .018 | .3 | .044 | .33 |
| | 3 | .0014 | .33 | .003 | .32 | .02 | .3 | .049 | .32 |
| | 4 | .0016 | .25 | .00375 | .2 | .023 | .27 | .052 | .33 |
| | 5 | .0019 | .5 | .004 | .35 | .025 | .3 | .059 | .32 |
| | 6 | .0019 | .5 | .0042 | .3 | .026 | .3 | .059 | .33 |
| 8/32 | 1 | .00078 | .4 | .00187 | .238 | .015 | .25 | .035 | .35 |
| | 2 | .00094 | .31 | .0021 | .3 | .015 | .34 | .036 | .4 |
| | 3 | .0011 | .3 | .0025 | .27 | .019 | .26 | .04 | .4 |
| | 4 | .0012 | .43 | .0028 | .25 | .019 | .3 | .05 | .29 |
| | 5 | .0014 | .45 | .0033 | .235 | .025 | .2 | .05 | .34 |
| | 6 | .00156 | .219 | .0031 | .355 | .03 | .15 | .045 | .45 |
| | 7 | .0017 | .23 | .0039 | .2 | .025 | .25 | .05 | .41 |
| | 8 | .0018 | .2 | .0039 | .23 | .026 | .245 | .055 | .34 |
| 10/40 | 1 | .00078 | .35 | .0018 | .24 | .015 | .25 | .035 | .35 |
| | 2 | .00094 | .29 | .002 | .3 | .015 | .32 | .04 | .33 |
| | 3 | .001 | .33 | .0024 | .24 | .019 | .23 | .043 | .33 |
| | 4 | .00125 | .26 | .0028 | .21 | .02 | .24 | .045 | .33 |
| | 5 | .0013 | .24 | .0028 | .27 | .02 | .27 | .049 | .31 |
| | 6 | .0014 | .22 | .0029 | .34 | .021 | .28 | .049 | .33 |
| | 7 | .0015 | .23 | .0031 | .37 | .023 | .25 | .049 | .36 |
| | 8 | .00164 | .38 | .0038 | .2 | .024 | .27 | .05 | .4 |
| | 9 | .0018 | .2 | .00375 | .275 | .025 | .25 | .05 | .4 |
| | 10 | .0018 | .2 | .0037 | .28 | .025 | .26 | .05 | .42 |

vcSSR Tables in Supplement 1) that allow biomedical and behavioral researchers to conduct ordinary null-hypothesis testing experiments with small sample sizes in an efficient manner by adding sample size in stages until the experiment is stopped (1) because an obtained *t* or *F* statistic is significant at a desired Type I error rate, α (.005, .01, .05, or .10), (2) because the *p* value is so high it is deemed futile to continue, or (3) because the further addition of sample size would exceed a predetermined maximum value. The method uses, on average, up to 30 % fewer subjects than the "Fixed Sample Rule" (FSR) that predetermines sample size from an *a priori* power analysis and conducts the experiment in a single step (i.e., as null hypothesis tests were originally intended to be used). Numerous examples of how and when to use the vcSSR are presented in Fitts (2011b).

The vcSSR Tables provide criteria based on the desired beginning and ending sample sizes, the number of subjects that will be added at each step (*n added*), and the *a priori* Type I error rate, α. The two criteria for an experiment consist of a lower criterion for significance and an upper criterion for "futility" for each set of conditions. The criteria were determined by computer simulations to hold the observed rate of Type I errors, $\alpha_o$ (observed alpha), at the selected α throughout the experiment. The criteria are "variable" between these sets of conditions because they were selected individually to hold $\alpha_o$ constant at α, not merely to assure that $\alpha_o$ will be less than α. For that reason, the criteria maximize the power of the test for a given set of conditions (see Fitts, 2010a and 2011b for comparisons with other formal SSR procedures). Stopping rules are available for sample sizes beyond those that are available for the vcSSR, such as CLAST (Botella, Ximenez, Revuelta, & Suero, 2006; Braschi, Botella, & Suero, 2014; Ximenez & Revuelta, 2007), but under identical conditions CLAST is less powerful than the vcSSR. Other widely available stopping rules are more appropriate for large studies such as clinical trials.

To continue with the example of the two-sample drug test, the researcher could conduct the test sequentially using the vcSSR. Suppose the researcher wants to use a Type I error rate of α = .05 (two-tailed), an anticipated effect size of 1.2 pooled standard deviation units (i.e., Cohen's d), and a desired power of at least 80 %. The customary power analysis suggests a total sample size of 24, with 12 per group. The formal way to select a sample size model with the vcSSR is to consult Fig. 6 in Fitts, 2010a for an independent groups *t* test with these parameters (alternatively, one can now consult Supplement 1 (see Electronic Supplementary Material (ESM); see instructions in Supplement 2). The figure suggests a sample size model of either 6/18 or 7/21. These are the lower and upper bounds of sample size. For the 6/18 model, testing is begun with six subjects per group. Sample size can be added according to the rules of the vcSSR provided the sample size never exceeds 18 subjects per group.

A quick-and-dirty way to estimate the starting sample size – which often works well – is simply to take half of the total *n* per group suggested by the customary power analysis (see Supplement 2, ESM). In this case, 12/2 = 6, so the researcher would select one of the available models beginning with six per group, i.e., 6/12 or 6/18, and the researcher can skip the vcSSR power curves. This is the method for selecting the initial sample size in CLAST, and, as mentioned previously, the vcSSR is a little more powerful than CLAST under identical conditions (Supplement 2, ESM). Checking the vcSSR tables for model 6/18 with α = .05, and assuming the researcher will add one subject per group at each iteration (*n added* = 1), then the lower and upper criteria for the vcSSR are .0165 and .240 (see Table 1).

The researcher then begins testing with six per group and conducts a *t* test. If *p* ≤ .0165, the experiment is stopped as significant with α = .05. If *p* > .240, the researcher stops the experiment "for futility" and retains the null hypothesis. Otherwise, the *p* value is "uncertain," and *n added* subjects can be added to each group provided the total of 18 per group for this model is not exceeded by adding the full *n added* subjects. The new *p* value from the augmented sample size is then tested according to the same rules. Typically, the experiment stops because the null hypothesis was rejected, because the *p* value was in the "futile" range, or because adding *n added* subjects would exceed the upper bound. In the latter two cases, the null hypothesis is retained. The researcher is always free to stop the experiment at any time without rejecting the null hypothesis, because early stopping for futility can never inflate the Type I error rate (Frick, 1998).

If, instead of adding 1 per group, the researcher uses *n added* = 4 per group per iteration, from Table 1 the new criteria would be .023 and .270. Possible sample sizes would be four, eight, 12, and 16 per group, after which the testing would have to stop because the addition of four per group would exceed 18. The criteria are a little less stringent (i.e., .023 > .0165) because there are fewer possible tests (only four possible tests instead of 13 with *n added* = 1). There are many kinds of constraints to planned experiments, and researchers are not always able to test after each individual subject is added. Therefore, various options for group sequential testing are available with the vcSSR.

The ability of the published tables of the vcSSR to hold $\alpha_o$ stable at the desired level has been validated with a between-subject completely random ANOVA with up to 20 groups, with either a between-subject or within-subject *t* test, with the Welch *t* test, with unequal sample sizes, and with a loss and replacement of subjects (Fitts, 2010a, b). The method has not until now been validated with repeated measures ANOVA except for a within-subject *t* test with a population correlation (rho, ρ) of .5.

In this paper, the vcSSR is used in large simulations to validate the ability of the existing vcSSR criteria to hold $\alpha_o$

stable with repeated measures ANOVA when the null hypothesis is true regardless of the number of treatments or the correlations between the scores. Power tables are provided for effect sizes (Cohen's *f*; Cohen, 1988) of .1 to .75, thus covering relatively small and large effects, with different numbers of measurements per subject and different values of ρ. No new sets of criteria are necessary for repeated measures ANOVA.

Limited numbers of simulations were conducted to confirm that the same criteria of the vcSSR held the $\alpha_o$ stable when a *F* test was generated from ANOVAs to test whether the percentage of explained variance, Multiple $R^2$, was different from zero, or when a *F* test was generated from MANOVA tests with multiple independent groups and two levels on a within-subjects variable. The relation between *F* and $\chi^2$ is discussed, and data are presented from tests where the *p* value from a Friedman $\chi^2$ repeated measures test was used to control stopping with the vcSSR. These data demonstrate why the vcSSR should not be used with *p* values from tests that do not use the *t* or *F* distribution.

## Method

### Simulations

Simulations were conducted to determine the behavior of various repeated measures models with the original criteria from Table 2 of Fitts (2010a). Each simulation to estimate the observed power was conducted 10,000 times and each simulation to estimate the observed $\alpha_o$ was conducted 40,000 times (4 x 10,000).

Custom programs were designed using the C programming language using 64-bit double precision arithmetic (Fitts, 2010a). Data were sampled using normal deviates generated by function gasdev() modified to use a pseudorandom number generator based on "Ran2()" (Press, Teukolsky, Vetterling, & Flannery, 1992). The function was seeded on the first call using the system clock so that each sequence would be different. Because of the seeding and the large period of the generator it is improbable that any long sequence of numbers was correlated or repeated in these simulations. Normal deviates were then transformed linearly using the desired means and standard deviations to create the generated samples.

Of the entire set of 16 combinations of lower/upper bounds (sample size models) available in Fitts (2010a), I selected four models for testing to cover a wide range of sample sizes, 4/18, 6/18, 8/32, and 10/40, where the first number is the starting sample size and the second number is the maximum sample size that will not be exceeded in the experiment. Table 1 includes the variable criteria for the four models used in most studies in this article (abridged from Fitts, 2010a, Table 2). The full criteria for all models are included in Supplement 1 (ESM).

The ranges of *n added* per group per iteration were 1-6 for models 4/18 and 6/18, 1-8 for the 8/32 model, and 1-10 for the 10/40 model (see Table 1).

The standardized effect sizes (*f*, Cohen, 1988) used in the simulations included 0 (null), 0.10, 0.15, 0.20, 0.25, 0.30, 0.35, 0.40, 0.45, 0.50, 0.55, 0.60, 0.65, 0.70, and 0.75 for multigroup ANOVA. The *f* is calculated by dividing the standard deviation of the treatment means by the pooled standard deviation.

The four levels of α were .005, .01, .05, .10.

New to these simulations were the inclusion of multiple correlated samples, so several values of ρ were tested (.0, .1, .3, .5, .7, and .9).

Independent simulations of 10,000 experiments were conducted for each of the four models at each of the 15 effect sizes, each of the six values of ρ, and each of the available levels of *n added* for that model. Each simulated experiment was conducted per the rules of the vcSSR. That is, the experiment was stopped if the *p* value from the repeated measures ANOVA on the treatments was in the significance range or in the futility range, or if the maximum sample size was reached; otherwise sample size was augmented according to the selected level of *n added* and the ANOVA was repeated. The program recorded the *p* value and sample size at stopping, so it was possible to calculate the empirical proportion of rejections (EPR) of the null hypothesis and the sample size at which the null hypothesis was rejected. When the null hypothesis concerning differences among treatment means was true in the population (effect size 0.0), the EPR was an estimate of $\alpha_o$ for the repeated measures ANOVA. When the null hypothesis was false, the EPR was an estimate of the power of the test at that selected effect size and level of correlation among the scores. Testing a range of effect sizes and correlations allowed the construction of power curves for the repeated measures ANOVA.

### Generating correlated data in multiple samples

The goal was to generate correlated data among multiple sets of treatments. To satisfy the assumption of sphericity for the ANOVA with repeated measures, the covariances among the scores should be similar. To accomplish this, each subject in an experiment was randomly assigned a "true *z*-score" that was never used in the experiments. Each new measurement for a subject was generated by creating a regressed score from the true *z*-score by multiplying the true *z*-score by the square root of ρ and then adding a random error generated by multiplying a random normal *z*-score by the square root of 1 minus ρ. That is, in C pseudocode for the *i*th subject:

```
truescore[i] = random_z()*sqrt(ρ);
```

and later for the *j*th new correlated variable:

```
newscore[i][j] = truescore[i] + random_z()*sqrt(1–ρ)
```

An identical method was used by Corey, Dunlap, and Burke (1998). Treatment effects of varying *z*-score sizes were then added to each score of a given treatment without affecting the correlations among the scores.

The method should generate an arbitrary number of treatments, k, where the average correlation in the bivariate correlation matrix for inter-treatment scores is equal to ρ with only a tiny negative bias, which is attributed to the procedure of averaging the correlations in a matrix, not to the method of generation of correlated scores (Corey et al., 1998). The multiple *R* generated from such simulations should behave as expected from the known characteristics of the unadjusted multiple *R*, i.e., the multiple *R* should vary per the size of ρ but also show bias per the number of variables in the analysis and sample size (McNemar, 1969). The number of unique off-diagonal correlations included in the average was k(k-1)/2, and varied from six with four levels to 45 with 10 levels of the treatment variable. The calculation of the Multiple *R* assumed arbitrarily that the initial level of the treatment variable was the dependent variable and the rest were independent variables. Multiple *R* was calculated using a modification of the Doolittle solution (McNemar, 1969). Adjusted $R^2$ was calculated by the formula:

$$R^2{}_{\mathrm{adj}} = 1{-}\left(1{-}R^2\right)^*\left(\left(N{-}1\right)\Big/\left(N{-}k\right)\right)$$

where $R^2$ is the unadjusted $R^2$ and *N* is the sample size.

### Simulation using .05 instead of the vcSSR

To illustrate how badly $\alpha_o$ becomes inflated by sequential sampling with repeated measures when researchers use .05 as the criterion for significance instead of the variable criteria provided in Table 2 of Fitts (2010a), simulations were conducted at the extremes of sample size and numbers of treatments (four or ten treatments in either the 4/10 or the 10/40 model) using .05 as the lower criterion and .36 (Fitts, 2010a, 1998) as the upper criterion ("BAD"). These were compared with simulations using the proper criteria of the vcSSR.

### Simulation using vcSSR with multiple correlation

Closely akin to the repeated measures ANOVA is the use of the *F*-test to determine significance of the coefficient of determination of a multiple correlation, $R^2$. Because the vcSSR works with the *F* distribution itself instead of one single method of generating an *F* statistic, the vcSSR method should be useful in controlling sample size with multiple correlation. To test this hypothesis, I generated new simulations of multiple correlations, using the same method as for generating data for repeated measures ANOVA, except that the *p* values used with the stopping rules were those obtained from a *F*-test of the null hypothesis that $R^2 = 0$.

Multiple correlation is sensitive to the number of predictor variables relative to the number of subjects, so I used only the larger 6/18, 8/32 and 10/40 vcSSR models with either three or five predictor variables. This equates to four or six levels of the treatment variable in repeated measures ANOVA, with the first generated measurement taken arbitrarily as the criterion and the remaining measurements as independent variables. These two models were tested 10,000 times at each combination of six levels of ρ (.0, .1, .3, .5, .7, .9) and all available levels of *n added* (see Table 1). For each model with three or five predictors, data were generated for the number of subjects at the lower bound, and a *p* value for the significance of Multiple $R^2$ was generated. The rules of the vcSSR were used on the *p* value either to stop the experiment or to add subjects per the value of *n added.* The EPR at each level of ρ was used to determine either the observed $\alpha_o$ or the power of the test to reject the null hypothesis. The average number of subjects required to reject the null hypothesis at each level of ρ and *n added* was determined.

### Simulation using vcSSR with MANOVA

Simulations of MANOVA employed the same four vcSSR sample size models as for the repeated measures ANOVA. The experimental design included one within-subjects variable with two levels and four, six, eight, or 20 independent groups. The Wilks' Lambda was calculated and converted to an approximate *F* statistic using the formula of Rao (1973, p. 556). The *p* values from the *F* tests were then saved and used with the vcSSR stopping rules as sample size was added to the various groups at all of the available levels of *n added* (see Table 1). Simulations included different values of ρ (0.0, 0.3, 0.5, 0.7, 0.9) and several effect sizes for the between-groups component in addition to the null hypothesis of no difference among any of the means. The mean differences for the within-subjects component was always 0. The EPR under the null hypothesis (Type I errors) and power to reject the null hypothesis based on the *F* test of Wilks' Lambda were recorded.

### Friedman $\chi^2$ test

*F*(df1, df2) is a ratio of two independent chi square variables divided by their degrees of freedom, $(\chi^2{}_1/\mathrm{df1})/\,(\chi^2{}_2/\mathrm{df2})$. With infinite degrees of freedom in the denominator, *F* equals just the numerator, $\chi^2{}_1/\mathrm{df1}$. This implies that the vcSSR criteria, which were developed using $p$ values from $F$ tests, should come close to working well with $p$ values from $\chi^2$ tests if the sample size is very large. To test this hypothesis, and to determine whether the vcSSR can be used with a $p$ from a Friedman $\chi^2$ test, I simulated the vcSSR with eight sample size models of varying sizes using the $p$ values from Friedman $\chi^2$ tests to control stopping. The method was identical to simulations with repeated measures ANOVA except that the $p$ value from the nonparametric test was used to control stopping. Each level of *n added* for each model was simulated 40,000 times when the null hypothesis of no differences among the treatment means was true. Type I errors and degrees of freedom at stopping were recorded.

## Results

Supplement 3 (ESM) includes a characterization of the methodology to generate multivariate correlated data in my custom C programs. The data demonstrate that the methods are adequate.

### Validation of vcSSR with repeated measures ANOVA

For the vcSSR criteria to be valid and useful with repeated measures ANOVA, the existing criteria should yield an $\alpha_o$ very close to the nominal $\alpha$ (i.e., .005, .01, .05, or .10) when the null hypothesis is true. When the null hypothesis is false the vcSSR method should yield significance with smaller sample sizes on average assuming an optimal sample size model has been selected.

### Type I errors with repeated measures ANOVA

The observed $\alpha_o$ under conditions of the four tested models (4/18, 6/18, 8/32, 10/40) at four levels of $\alpha$ (.005, .01, .05,.10) and six levels of $\rho$ (.0, .1, .3, .5, .7, .9) are presented in Fig. 1. The data are averaged over the available levels of *n added*, which had no effect on $\alpha_o$, so each mean in the figure is based on 60,000 to 100,000 simulated experiments. The groups of six symbols at each level of treatment represent the different values of $\rho$ increasing from left to right. The observed rates of Type I errors fit well with the nominal values of $\alpha$ in all models.

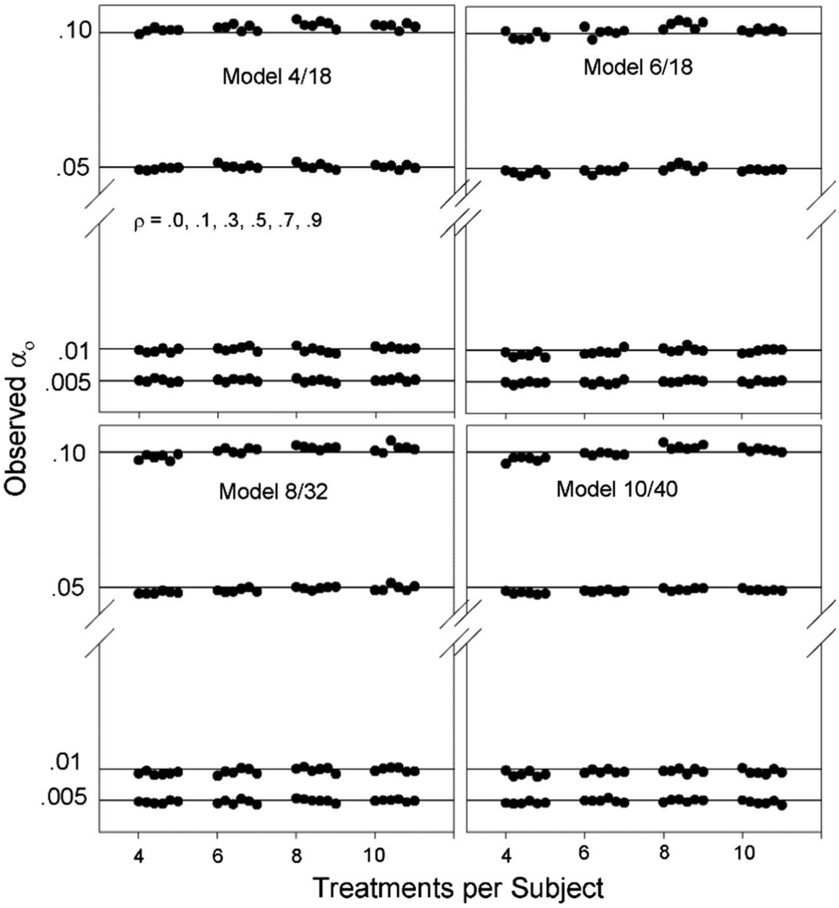

**Fig. 1.** The observed rate of Type I errors ($\alpha_o$) of the four models of the vcSSR used in the simulations as a function of the number of treatments per subject and $\rho$. The existing criteria published in Fitts (2010a) worked well to hold $\alpha_o$ stable, and these were not affected by the number of treatments or by the $\rho$ used in the simulated ANOVAs

### Simulation using .05 instead of the vcSSR

Figure 2 illustrates the "BAD" method of sequential sampling when sample size is added and tested each time at the nominal level of α, in this case .05, instead of using the correct criteria from the vcSSR Tables. In this figure, the data were averaged across all six levels of ρ, which did not affect $\alpha_{o,}$ so the means are based on 60,000 completed experiments where the experiment was stopped because *p* was less than or equal to .05, because *p* was greater than .36, or because the maximum number of subjects had been used (18 or 40 depending on the model). This procedure mimics the practice used by many researchers who are unaware of the inflation of alpha with sequential testing: they stop when they achieve a significant result, when the *p* value is so high that they believe additional subjects will not help, or when the experiment has already consumed so many subjects that the effect is no longer worth pursuing. This higher probability of rejection of a true null hypothesis leads to the increased publication of false-positive results, confusion in the literature, and a loss of effort and resources in unsuccessful replications. By contrast, in Fig. 2, the criteria of the vcSSR hold $\alpha_o$ stable at .05 across sample size models and levels of *n added*. The $\alpha_o$ was unaffected by the number of treatments included in the experiments, so it was the number of tests, not the number of treatments, that inflated $\alpha_o$ with the BAD method.

### Power with repeated measures ANOVA

The power of the vcSSR with repeated measures ANOVA to reject the null hypothesis at different effect sizes and different levels of ρ at the .05 level of significance are given in Fig. 3 for four and six treatments and in Fig. 4 for eight and ten treatments per subject. The curves can be used to select an appropriate sample size model for a repeated measures experiment given an estimate of the overall effect size and the ρ. As anticipated, higher levels of correlation among the variables increased power, and power also varied with the number of levels of the treatment variable, so different curves were necessary for the different numbers of treatments.

### Sample size comparison of vcSSR with fixed stopping rule

Figure 5 illustrates the average savings in sample size when using the vcSSR instead of the FSR with an equivalent amount of power and the same effect size. Sample size calculations for the FSR employed the program G*Power 3.1.9.2 (Faul, Erdfelder, Lang, & Buchner, 2007). The 10/40 model with ten treatments is used in the illustration, so the model obligates the use of at least ten subjects. The average number of subjects is always less with the vcSSR than with the FSR except in those simulations with both large effect sizes and very high correlations between the scores.

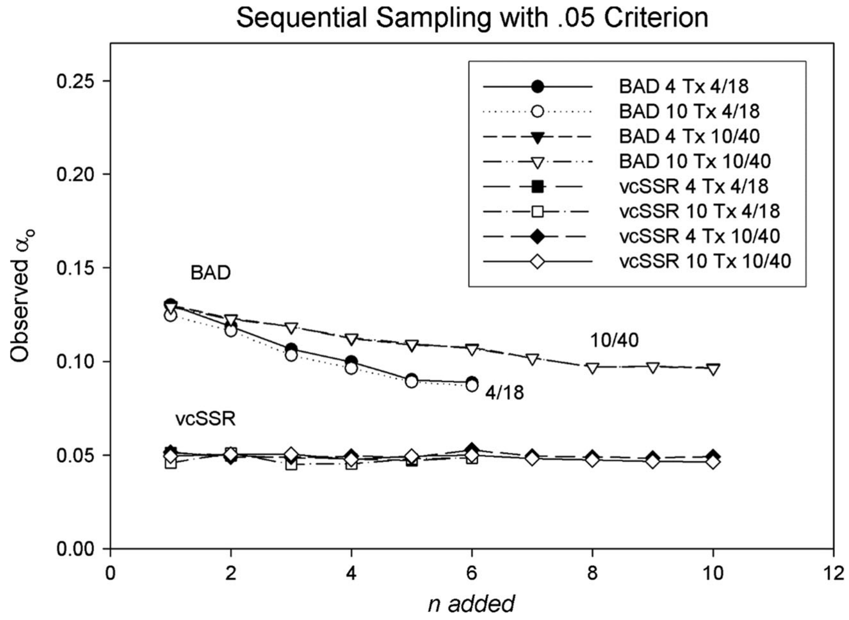

**Fig. 2.** Effect of sequential sampling on Type I errors using lower/upper criteria of .05/.36 ("BAD") in contrast to the values obtained for the vcSSR in Fitts (2010a). Every opportunity to add sample size increases the Type I error rate, and smaller values of *n added* allow more additions to the sample size than larger values before reaching the maximum value. The inflation of $\alpha_o$ was also affected by the sample size model but not by the number of treatments (Tx) in the model. The vcSSR always held $\alpha_o$ stable at .05

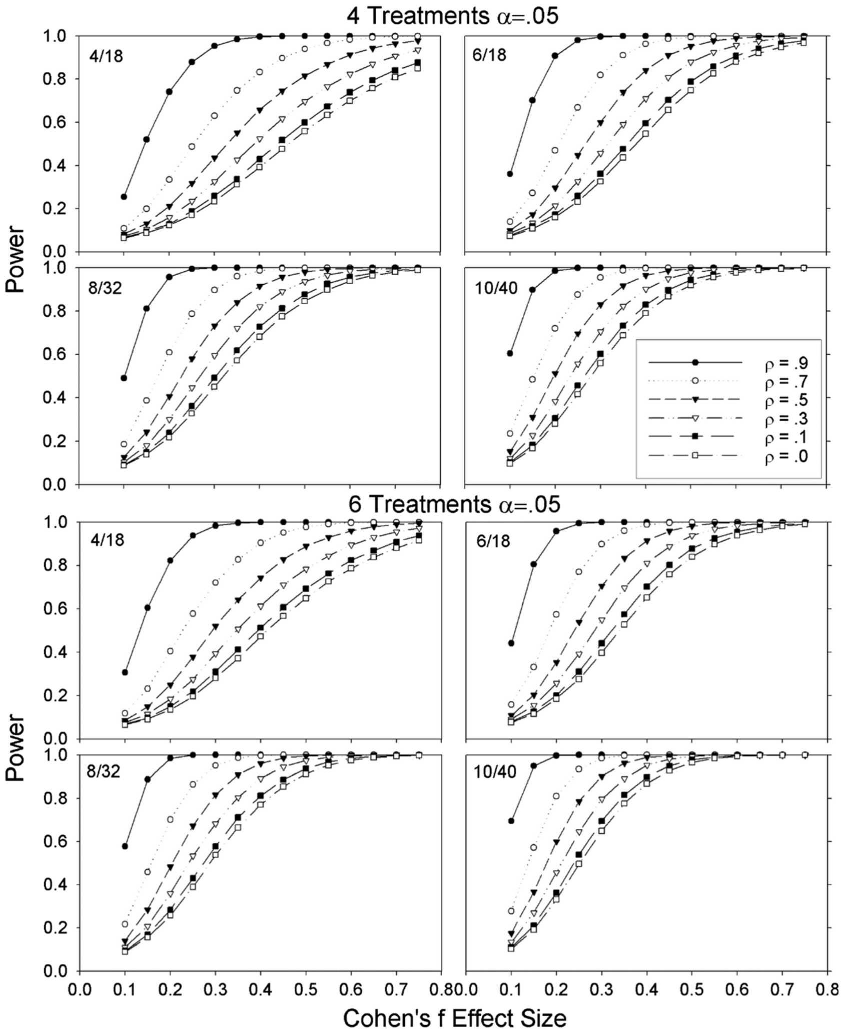

**Fig. 3.** Power of the vcSSR for repeated measures ANOVA at the .05 level with four or six treatments in the four models used in the experiment. Power varies as a function of the sample size model, the ρ, and the size of the effect

## Comparison of vcSSR power between-subjects and within-subjects

Figure 6 illustrates the sample size at the rejection of the false null hypothesis as a function of the population effect size *f* and the mean observed power. Included are both the repeated measures simulations reported here with ρ = .0 and these same four models previously reported in between-subjects simulations with the vcSSR in Fitts (2010a, b). The relative efficiency of the repeated measures simulations even at their worst level of power (i.e., ρ = .0) is evident. As with a power analysis using the FSR, the power of the repeated measures ANOVA with ρ = .0 and *N* subjects per experiment is almost identical to the power of the between-subjects ANOVA with *N* subjects per group when the number of between-subjects groups equals the number of within-subjects treatments (see drop lines to the effect size-power axis in the figures). Between-subject power curves are already available for 16 vcSSR sample size models (Fitts, 2010a, b) and they can be used to calculate the minimum power of repeated measures designs (i.e., assuming ρ = .0). Correlations above 0 will provide more power.

## Simulation using vcSSR with multiple correlation

Figure 7 illustrates power and sample size of the vcSSR with multiple correlation at the .05 level with either three or five predictors and the 6/18, 8/32, or 10/40 models. The EPR under the null hypothesis, an estimate of $\alpha_o$, is given for a ρ value of .0, and power is estimated for other values of ρ. The data were averaged across levels of *n added*, so they represent the means of 60,000 to 100,000 experiments. Most notably for the present study, the $\alpha_o$ was close to the nominal α, .05. The maximum (i.e., worst) values of the five observed $\alpha_o$ at the four levels of α were .00549, .01073, .05229, and .10784.

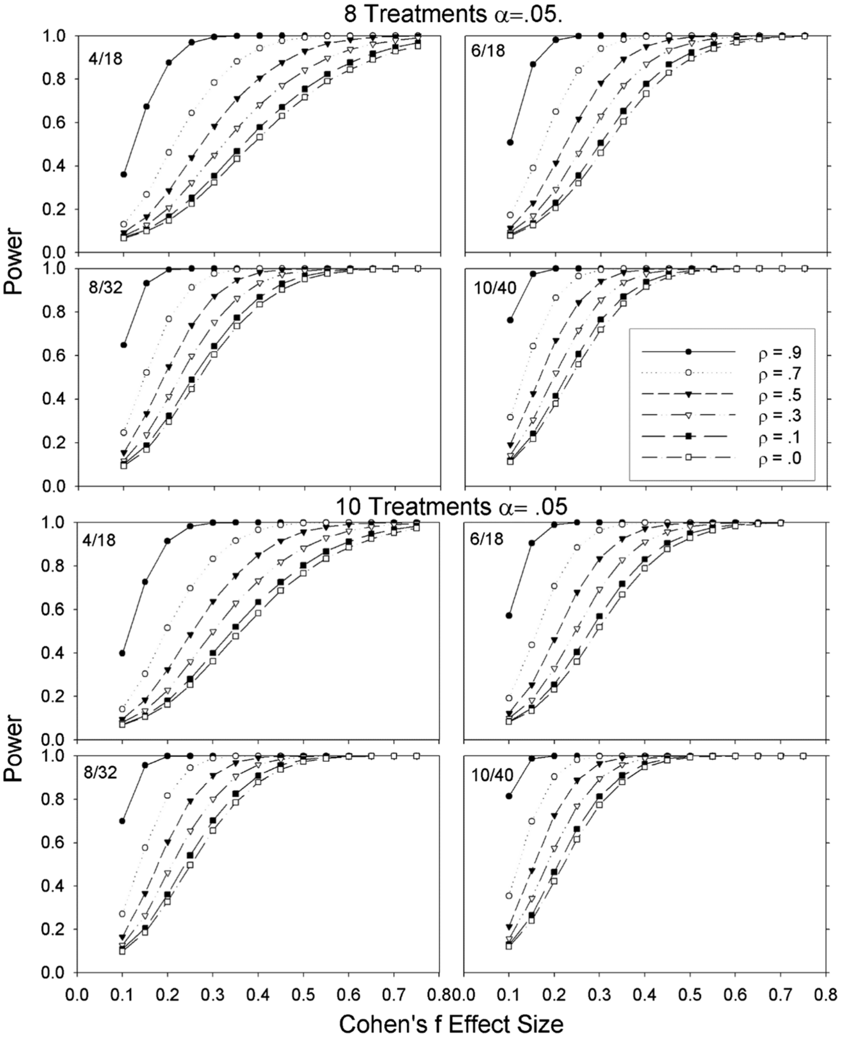

**Fig. 4.** Power of the vcSSR for repeated measures ANOVA at the .05 level with eight or ten treatments in the four models used in the experiment

## Simulation using vcSSR with MANOVA

As demonstrated in Fig. 8, the previously published values of the vcSSR worked well to control Type I errors during simulated experiments using MANOVA with sequential sampling. MANOVA simulations were conducted 10,000 times each using four vcSSR models with five levels of ρ on the within-subjects variable, four group sizes on the between-groups variable, and from six to ten levels of *n added* when the null hypothesis of no differences among any of the means was true. The results did not vary systematically with the level of *n added*, and I have averaged the values across six to ten levels of *n added* in Fig. 8 so that each point represents from 60,000 to 100,000 simulated experiments with the null hypothesis true. Clusters of five points represent the respective levels of ρ in the simulations. All four tested models of the vcSSR controlled Type I errors close to the nominal level of α regardless of the number of groups or the correlation of the within-subjects levels.

Figure 9 illustrates the EPRs for the MANOVA analyses, including a repetition of the Type I error rates at effect size 0 and power for effect sizes .35, .55, and .75 for the between-subject variables. The ρ was varied for the within-subject variable, and the results from different correlations are displayed in the figure in increasing order from left to right for each sample size model at each effect size. The values are offset slightly to avoid overlapping symbols. The relation between the intercorrelations of dependent variables and power is complex in MANOVA, and the negative relationship was expected (Cole, Maxwell, Arvey, & Salas, 1994). The data demonstrate the usefulness of the different models at different effect sizes and group sizes.

## Friedman $\chi^2$ test

Figure 10 illustrates the theoretical relationship between $F$(df1, df2) and $\chi^2$(df1) in the top curves. The $\chi^2(3)=7.82$ and $\chi^2(9)=16.92$ were selected because they are the critical values for α=.05 with 3 and 9 degrees of freedom (df), and .05

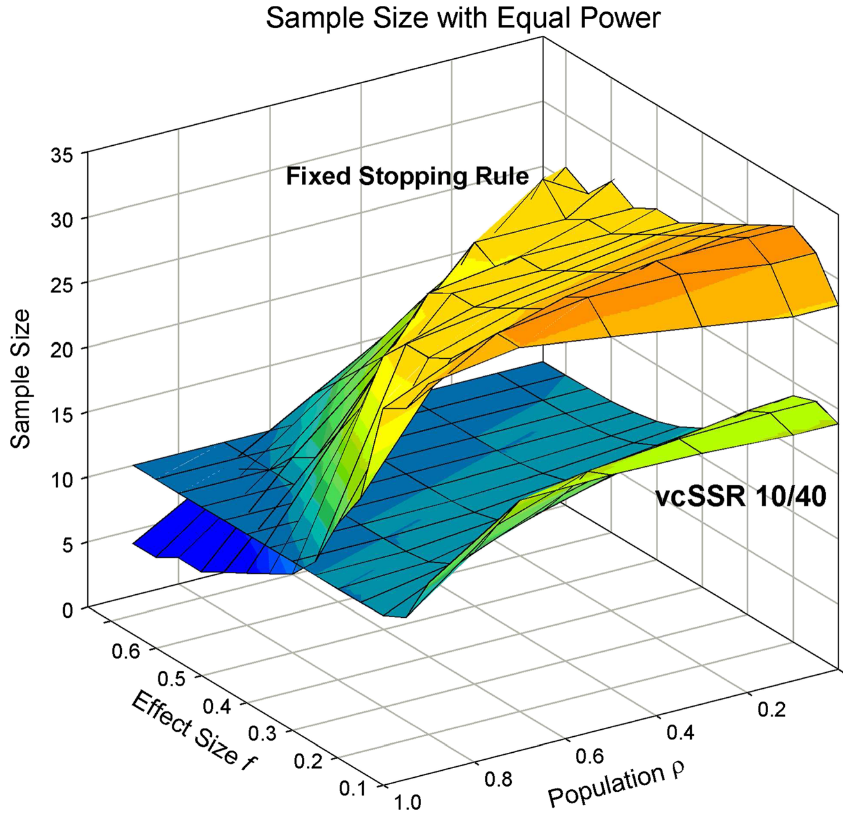

**Fig. 5.** Sample size savings when using the vcSSR in the 10/40 model with ten treatments compared with the Fixed Stopping Rule using an *a priori* power analysis to generate the same power as the vcSSR. In this model, an effect size of 0.3 produces about 80 % or greater power at all levels of correlation (see Fig. 5), and a sample size model with a smaller lower bound would be used for all larger effect sizes

is the lower limit of the probabilities of $F(3, \text{df2})=2.60$ or $F(9, \text{df2})=1.88$ as df2 approaches ∞. The bottom curves illustrate the $\alpha_o$ of Friedman simulations using vcSSR criteria with either 3 or 9 df averaged over all levels of *n added* for each of eight sample size models plotted as a function of what would be the average df2 in a repeated measures ANOVA of the same data at the time the Type I error was made. The data are consistent with .05 as an upper limit of $\alpha_o$ as the df2 increase within the range of sample sizes currently available for the vcSSR. In no case for these existing vcSSR models, however, was the obtained $\alpha_o$ acceptably close to a constant $\alpha$ = .05 so that the vcSSR could reliably be used to control stopping in a Friedman test at that $\alpha$. Use of Friedman in this way would increase Type II errors.

## Effect of *n added* on EPR and sample size

Often in several publications on the vcSSR I included all available levels of *n added* in the development of the power curves and sample size, but I then averaged over all available levels of *n added* when presenting the data in graphs and tables. The rationale for the averaging was to save journal space by not describing a variable that provides little new information. While it is true that I have detected no important systematic effect of *n added* on the rate of Type I errors or power, there is a small systematic effect of *n added* on the ultimate sample size at the time of rejection of the null hypothesis. Figure 10 demonstrates potential effects of *n added* on the empirical proportion of rejections (EPR) and the sample size at the rejection of the null hypothesis for the four sample size models used throughout all original studies in publications on the vcSSR (Fitts, 2010a, b, current article). The results are averaged over all statistical models (between-subjects and repeated-measures ANOVA and MANOVA), all levels of correlation, all effect sizes including the null hypothesis, and all numbers of groups or treatments at the .05 level of significance. There was an equal contribution of all variables to each model, and the results illustrate a general tendency for an increase in sample size at the maximum level of *n added* (six, eight, or ten) compared with an *n added* of 1. The variability in EPR had no systematic trend related to level of *n added* and the variability tended to decrease with an increase in the sample size model. Not surprisingly, larger sample size models provided more power but also used more subjects on average.

Supplement 1 (ESM) is a comma delimited text file containing the raw EPR and sample size for all normative data collected for between-groups or repeated-measures ANOVA or MANOVA in three publications (Fitts, 2010a, b, current article). The file is suitable for import into a spreadsheet or

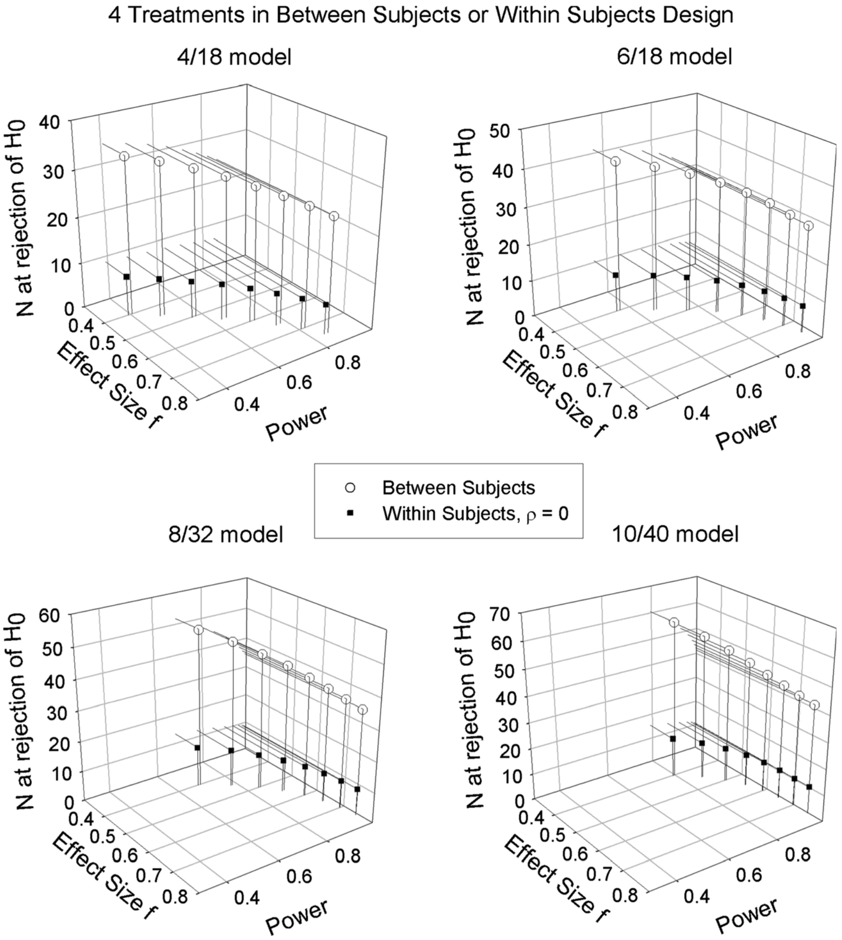

**Fig. 6.** Comparison vcSSR with between-subjects (from Fitts, 2010a, b) and within-subjects (this paper) designs with four groups or treatments. Sample size is illustrated as a function of effect size and observed power when ρ = .0 for repeated measures in the four sample size models tested. The comparative efficiency of repeated measures is evident. Use of a given sample size model with repeated measures and ρ = .0 produced the same power at each effect size as the corresponding between-subjects design (see drop lines)

database. The top row is the header row including the procedures – BetAOV (between-subjects ANOVA or *t* test), RepAOV (repeated measures ANOVA or *t* test) or MANOVA – the number of independent groups, the number of correlated or repeated treatments, the ρ, the α, the level of *n added*, the effect size as Cohen's d or f, the lower and upper bounds of the sample size model, the lower and upper criteria from the vcSSR table, the EPR, and the *n* per group at the rejection of the null hypothesis in simulations. Supplement 2 (ESM) is a document that explains how to use the power curves or the database to select an optimal vcSSR model for an experiment.

## Discussion

In previous papers (Fitts, 2010a, b, 2011b) I suggested that the criteria of the vcSSR might be used with any *t* or *F* test because it works with the *F* distribution rather than any particular method of generating an *F*. The present data confirm that the same variable criteria published in Fitts (2010a) effectively control $\alpha_o$ very close to the nominal level of α in repeated measures ANOVA across a wide range of sample sizes, effect sizes, numbers of treatments, and ρ. In a second challenge, the vcSSR was used to control sample size in a *F* test of a multiple correlation $R^2$ with three or five predictor variables, and here, also, the $\alpha_o$ was held at the nominal level. In a third test, the vcSSR effectively controlled $\alpha_o$ when used with a MANOVA that was analyzed by calculating a Wilks' lambda and then transforming it to an approximate *F* value to get the *p*. Finally, the relationship between *F* and $\chi^2$ produced predictable results when using the vcSSR criteria with a Friedman $\chi^2$ test, a nonparametric analog of the repeated measures ANOVA, instead of a *F* test, but the observed Type I error rates were too low to allow the use of vcSSR to control stopping with a Friedman or other nonparametric or $\chi^2$ test without inflating Type II errors. I recommend using the vcSSR with sequential sampling whenever the decision to stop is based on a *p* value from a *t* test or *F* test and the sample size is within the range provided in the vcSSR Tables (Fitts, 2010a). Use of the vcSSR

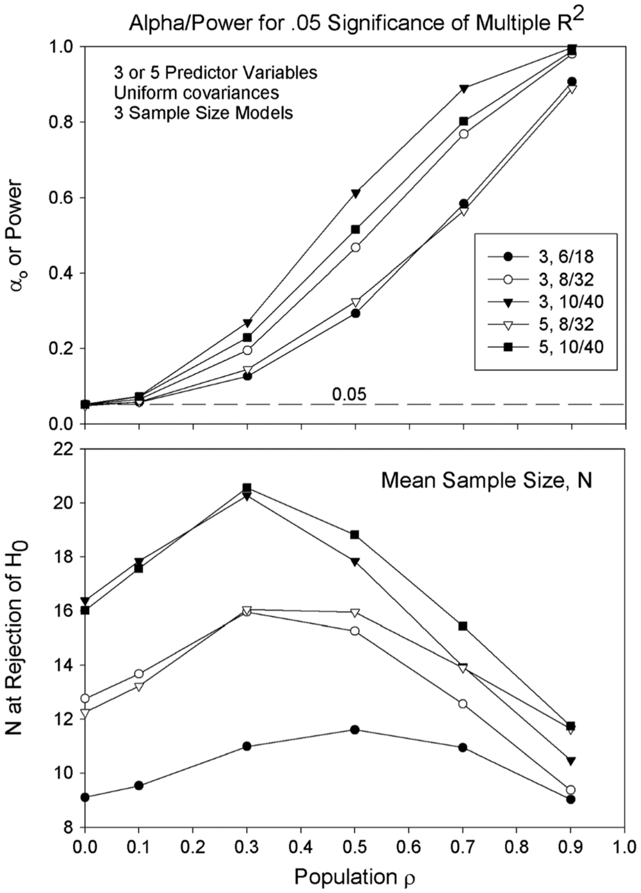

**Fig. 7.** Empirical proportion of rejections (EPR) and sample size of the vcSSR with multiple correlation at the .05 level with either three or five predictors and three sample size models. The EPR estimates $\alpha_o$ under the null hypothesis when ρ = .0, and estimates power for other values of ρ. The data were averaged across levels of *n added.* Most notably, the $\alpha_o$ was very close to the nominal α, .05, in these models of the vcSSR

will control Type I error rates at the nominal level and may consume fewer subjects than the FSR with equal power.

The sample size savings in Fig. 5 represent a maximum, because 10/40 is the largest available sample size model. Note that this model can never use fewer than ten subjects, so sample size models with lower initial sample sizes should be used instead of the 10/40 model when the effect size and correlation are both high. According to Fig. 4 for power with ten treatments, the 10/40 model provides about 80 % power at any level of correlation when the effect size is 0.3, so smaller models would probably be used for any larger effect sizes. At an effect size of 0.3 in the 10/40 model, the vcSSR uses an average of 15 subjects when ρ = .0 and ten subjects when ρ = .9. These contrast with 30 and 15 subjects, respectively, required by the FSR with equal power. Thus, this model of the vcSSR can use from 33 % to 50 % fewer subjects than the FSR at a useful level of power.

The principal message of Fig. 2 using the "BAD" method of sequential sampling is that repeated testing with a criterion of .05 does not merely have a "chance" of increasing Type I errors, it does always increase Type I errors in simulations to a replicable degree. When the null hypothesis is true, the more times sample size is added to a simulated experiment, the greater the increase in observed Type I errors. The inflation of $\alpha_o$ is worst when *n added* is one subject per iteration because this allows the addition of sample size – and opportunities for additional Type I errors – many times before the maximum sample size is reached. Larger values of *n added* (i.e., "group" sequential sampling) provide fewer opportunities to add Type I errors before the maximum sample size is reached. Nevertheless, the observed $\alpha_o$ with the BAD method is always considerably larger than α. The $\alpha_o$ was also affected by the sample size model for the same reason: there are more opportunities to add sample size when the maximum is 40 than when the maximum is 18. By contrast, the criteria of the vcSSR always held $\alpha_o$ stable at .05 across sample size models and all levels of *n added.*

The value of the upper criterion, .36 in the BAD case, has much less influence than the lower criterion, .05, in determining the number of Type I errors (Fitts, 2010a). A variation in the upper criterion was used only to "fine tune" the resulting $\alpha_o$ in my original simulations so that it would be very close to the nominal α. Therefore, the magnitude of the inflation of $\alpha_o$ was not an artifact of my choice of upper bound.

Figure 6 illustrates comparisons between previously-published completely random between-groups designs (data from Fitts, 2010a, b) and the present repeated measures designs with the four vcSSR sample size models tested. The sample size at the rejection of the null hypothesis is plotted as a function of the imposed effect size and the observed power. The magnitude of the savings in sample size from a repeated measures design is evident even with zero correlation. The observed power generated by each model of the vcSSR is the same given either a between-subjects design or a repeated measures design (see the drop lines to the x-y axis). Thus, power curves presented in my previous papers can be used to calculate the minimum power of a repeated measures analysis by assuming the population ρ = .0. Alternatively, the power curves for ρ = .0 in the present paper may be used for completely random designs (the effect sizes between 0.1 and 0.35 are new).

The vcSSR was not originally designed for multiple correlation studies that require large sample sizes, but we see in Fig. 7 that it can apply with the largest sample size models when ρ is large. It did not make sense to test the smaller sample size models where the sample size could equal the number of predictor variables. The principal demonstration of the study was that the vcSSR criteria worked to control $\alpha_o$ with a different method of generating a *F* statistic. One of the models that is potentially useful is the 10/40 model at a population ρ = .7 (Fig. 7, filled squares). The null hypothesis that $R^2$ = 0 was rejected on average with about 15 total subjects with five predictors at

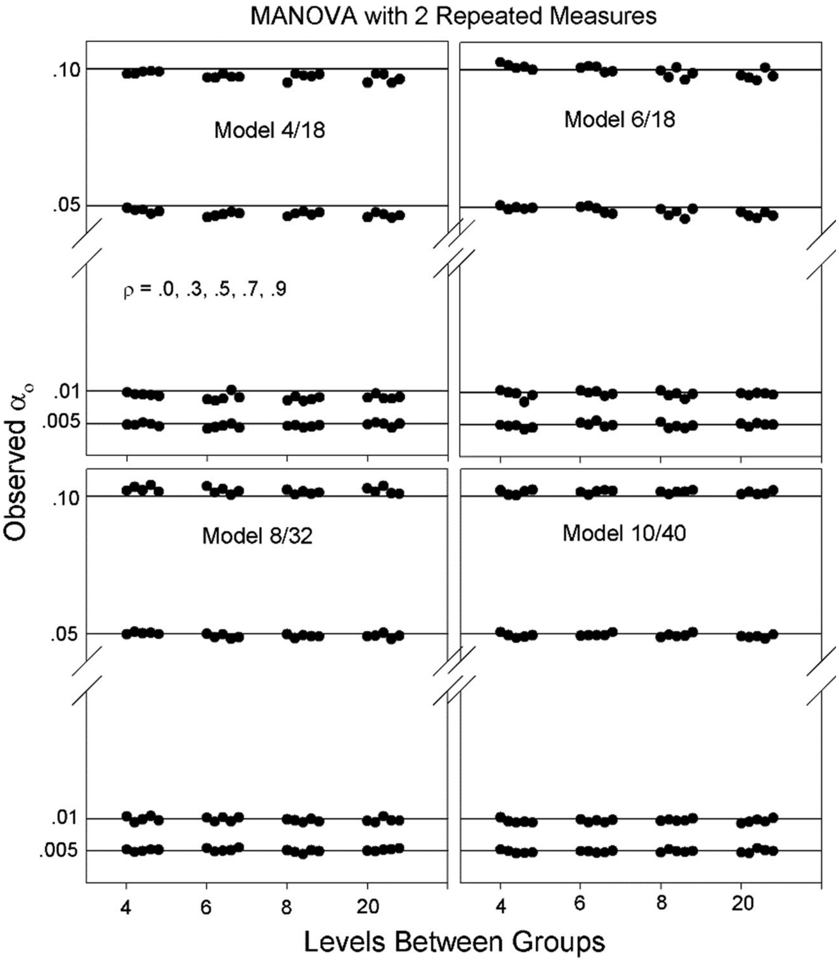

**Fig. 8.** The observed rate of Type I errors ($\alpha_o$) with MANOVA using four models of the vcSSR as a function of the number of groups in the design and the ρ between two levels of a single within-subject variable. Wilks' Lambda was calculated and transformed to an approximate *F* statistic, and the observed rate of Type I errors was calculated from the collected *p* values when the null hypothesis of no difference among the means of the groups or treatments was true. The existing criteria published in Fitts (2010a) worked well to hold $\alpha_o$ stable, and these were not affected by the number of groups or by the ρ used between the dependent measures

~80 % power and with about 14 total subjects with three predictors at ~90 % power. The required sample size of the FSR at equal numbers of predictors and power is 20 in both cases. An approach like the vcSSR could save subjects in multiple correlation studies.

Finding significance with multiple correlation with smaller values of ρ would require new variable criteria to be generated for sample size models with higher bounds. For example, a 20/40 model would probably provide much better power for moderate correlations than a 10/40 model because fewer Type II errors would be made when the sample size was equal to the lower bound.

The vcSSR can be used in experiments where an ordinary repeated measures ANOVA or mixed model is appropriate, even in higher order designs with crossed repeated measures factors, as long as the experiment will be stopped based on a single *p* value from the ANOVA table (discussed in Fitts, 2011b). Other *p* values in the complex analysis can be evaluated at the regular value of α, such as .05, because they are not used in the decision to stop the experiment. For valid results from the vcSSR in ANOVA, the researcher must track the *p* value of only one effect. This can be any main effect, an interaction effect, or even a planned contrast among two or more means (Fitts, 2010a, 2011b). For example, tracking a *p* value from a global *F* in a multi-group experiment may prematurely lead to early stopping when two control groups are different (e.g., a positive vs. a negative control), but a planned comparison can track when the difference of interest becomes significant. The vcSSR may not be appropriate in experiments where sample size is difficult to add, such as longitudinal studies lasting months.

The present simulations were designed to obey all assumptions and requirements of the statistical tests employed, such as normal, homogeneous distributions and equal covariances between the various treatments. The vcSSR can suffer the same deviations from the relevant *F* distribution as the usual ANOVA

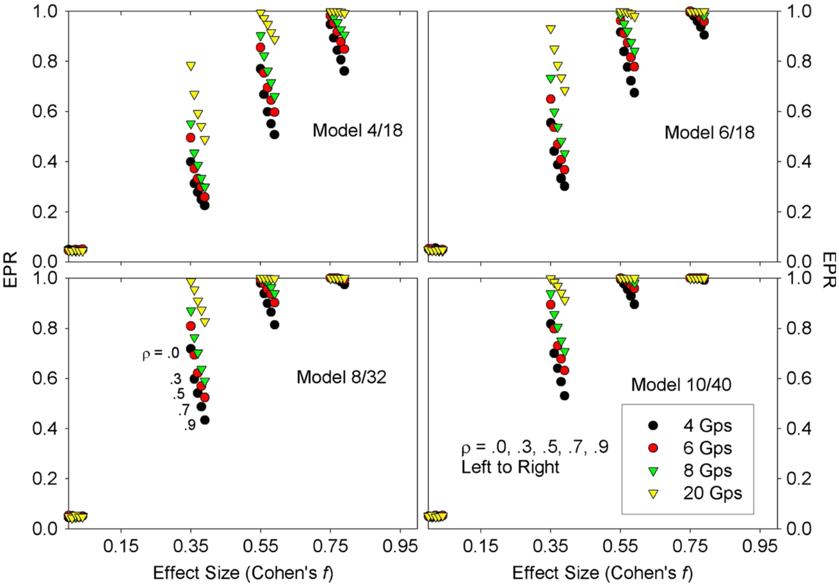

**Fig. 9.** Empirical proportion of rejections (EPR) at $\alpha = .05$ in MANOVA designs with two levels within-subjects and four, six, eight, or 20 levels between-subjects with stopping controlled by vcSSR models 4/18, 6/18, 8/32, or 10/40. Different levels of the population correlation ρ for the within-subjects variables were .0, .3, .5, .7, or .9, and the levels are displayed left-to-right as slightly offset points over each of the population effect sizes, 0, .35, .55, or .75 for the between-subjects variables. The within-subject population effect size was always 0. Type I errors are as given in Fig. 8. All other points represent power. Decreasing power with increasing ρ is typical for this model of MANOVA

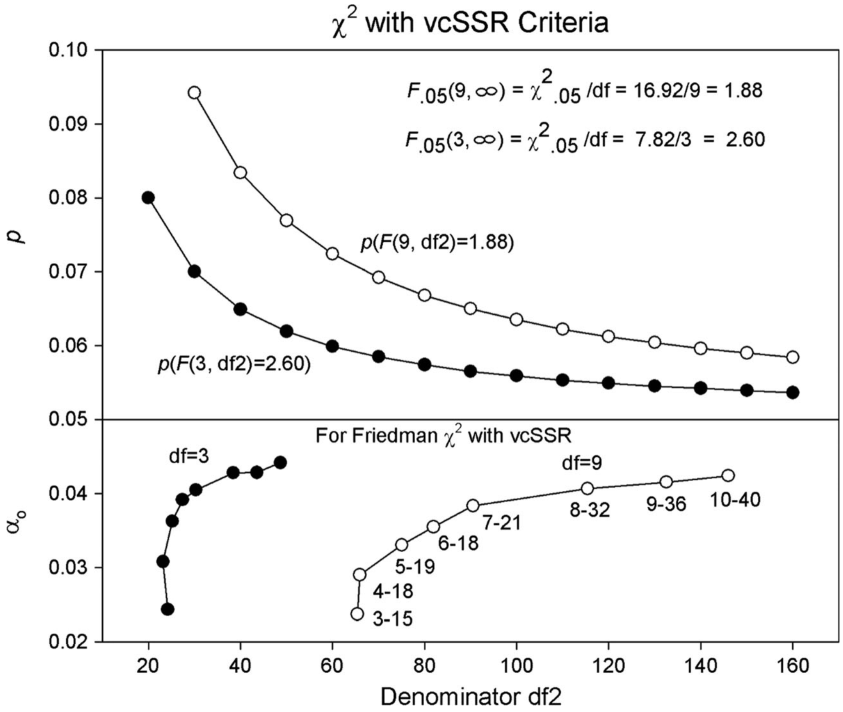

**Fig. 10.** Theoretical relation of F and $\chi^2$ (above .05) and observed rate of Type I errors, $\alpha_o$ (below .05), when $p$ values from Friedman $\chi^2$ tests are used to control sample size in the vcSSR instead of the $p$ values from $F$ tests. The probability of a $\chi^2$(df1) is a limiting value of the probability of $F$(df1, df2) as df2 approaches ∞ (upper curves). The criteria for the .05 level were used in simulations with eight sample size models of the vcSSR with either four or ten treatments. The abscissa is the number of degrees of freedom (df2) that would appear in the denominator of a $F$ test of the same data based on the average total sample size at the time the $H_0$ was rejected when making the Type I error. The α of the vcSSR (.05) appears to be a limit of the Type I error rate for the $\chi^2$ test as df2 increases. No $\alpha_o$ was acceptably close to .05 to allow the reliable use of the vcSSR to control stopping with the Friedman test without inflating Type II errors

when assumptions are violated, such as grossly heterogeneous variances in the between-groups case (Fitts, 2010b). Caution is advised when using the vcSSR with such data. See Everitt (1995), Keselman, Algina, and Kowalchuk (2001), and Oberfeld and Franke (2013) for reviews of various repeated measures or MANOVA designs and for problems with repeated measures designs, such as a deviation from sphericity. If a test for non-sphericity indicates that a correction to the degrees of freedom is required for the repeated measures ANOVA, the *p* value after the correction should be used to stop the experiment.

A repeated measures design is very efficient compared with a between-groups design, but researchers should be aware of when the repetition of the treatment itself will affect the outcome (e.g., habituation, operant learning, drug tolerance, etc.), and take precautions to eliminate these effects when necessary (e.g., washout). For example, certain procedures may be standard for evaluating sensitization to pain or allodynia, but using these tests repeatedly many times on the same subjects may result in habituation that can obscure a treatment effect (Hannaman, Fitts, Doss, Weinstein, & Bryant, 2016). If the effect of repeated treatments is unknown, suitable control groups should be employed.

I used Cohen's *f* as a measure of effect size so the simulations would be directly comparable to my previously collected data with completely random designs (see Fig. 6). Other measures of the size of effect for repeated measures can take into account the magnitude of the correlations instead of just the spread among the means (Bakeman, 2005).

In addition to repeated measures ANOVA, multiple correlation, and MANOVA, I tested the rate of observed Type I errors when using the criteria of the vcSSR to control sample size in simulations of Friedman $\chi^2$ tests with $\alpha = .05$ to test the hypothesis that .05 is a limiting value as sample size increases. The hypothesis was confirmed, but in no case did the vcSSR control the rate of Type I errors at an acceptably close value to .05. I do not recommend using the vcSSR Tables to control sample sizes with tests that do not use the *t* or *F* distributions, such as the Mann-Whitney-Wilcoxon test, the Wilcoxon dependent-samples test, the Kruskal-Wallis test, or the Friedman test. Variable criteria specific to the 16 original sample size models to hold Type I errors constant for these tests are available by writing to the author. The published vcSSR Tables (Fitts, 2010a) will hold Type I errors constant only with statistics that use a *t* or *F* distribution, or with a statistic that can be converted to *F* (e.g., Wilks' lambda).

Some readers may be interested in why the vcSSR behaves the way it does with the Friedman test. With infinite degrees of freedom in the denominator, *F* equals the numerator of a ratio of two independent $\chi^2$ variables, $\chi^2{}_1/\mathrm{df1}$, and the *p* values of *F* and $\chi^2$ are accordingly equal. For example, with 3 df, the value of $\chi^2$ that leaves .05 in the tail of the distribution is 7.82, and the value of $F(3, \infty)$ that leaves .05 in the tail is $7.82/3 = 2.60$. This implies that the vcSSR, if it works with all *F* tests, should work well to control Type I errors in a test using a single $\chi^2$, such as the Friedman test, if the sample sizes are very large. Unfortunately, very large sample size models are not available for the vcSSR, but the data in Fig. 10 are certainly suggestive that .05 is a limiting upper value for the rate of Type I errors across the range of df2 currently available. However, the $\alpha_o$ values were not acceptably close to .05 to use the vcSSR to control stopping in actual experiments. New tables could be constructed with larger sample sizes to control stopping during sequential sampling in experiments with any $\chi^2$ test, or, preferably, a mathematical solution could be found so that the variable criteria could be predicted from a formula for any arbitrary set of lower and upper bounds, $\alpha$, and *n added*.

Why are the lower curves in Fig. 10 shaped the way they are? The Friedman tests in this study would ordinarily be compared with the relevant critical value of $\chi^2$ for the .05 level. However, with the vcSSR the critical value is not always $F = 7.82/3 = 2.60$ for 3 df or $F = 16.92/9 = 1.88$ for 9 df. Instead, the criteria of the vcSSR assume different df2 levels and adjust the criteria for *p* downward, in a more conservative direction, for smaller sample sizes (i.e., fewer df2). This allows fewer rejections with the *p* values from the $\chi^2$ tests as the sample size decreases. Furthermore, at very small sample sizes the $\chi^2$ distribution stops being an accurate model for the actual *p* values of the Friedman test and users would need to consult a table of exact values. I did not use those tables in the simulations.

Data from all available normative simulations (those used to determine Type I error rates and power curves without deviations from the assumptions of ANOVA) of four sample size models in three publications were used to analyze the effect of *n added* on the average EPR and average sample size at the rejection of the null hypothesis across all ANOVA and MANOVA models tested so far (Fitts, 2010a, b, current article). Data were included for all levels of correlations and group, treatment, and effect sizes at the .05 level so that the result is a "main effect" of *n added* across all normative simulations. The results in Fig. 11 confirm my assumption that the EPR was largely unaffected by the level of *n added*. The results also confirm that a minimization of sample size on average is achieved best with small values of *n added*, because the addition of large numbers of subjects to each group at each iteration make it harder to stop at the optimal sample size when the *p* value first becomes less than the lower criterion. There are different constraints to the designs of various experiments, and for logistical or other reasons researchers cannot always test after every individual subject. When a researcher must use a large *n added*, it is still worthwhile to use the vcSSR. Even with large *n added* the sample size will be smaller on average than the FSR (Fitts, 2010a), and the vcSSR will control the rate of Type I errors. It is important to adhere to the selected *n added* throughout the experiment to be certain that the Type I error rate remains at the nominal value. Fortunately, the procedure is robust to such events as missing data or the

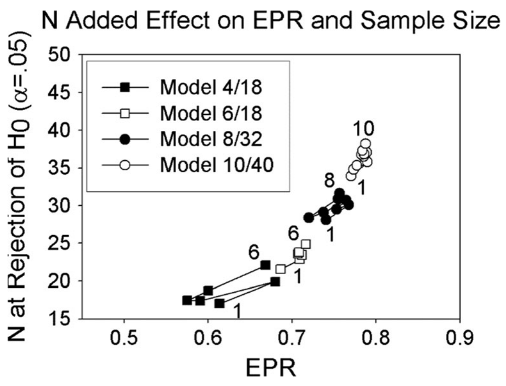

**Fig. 11.** The influence of *n added* on the empirical proportion of rejections (EPR) of the null hypothesis and the sample size, *N*, at the time of the rejection of the null hypothesis in four sample size models that were used in all normative studies of ANOVA or MANOVA in three publications. The results are averaged for the .05 level of significance across all levels of correlations, effect sizes including the null hypothesis, and for all numbers of groups or treatments. Within each sample size model, the sample size is greater at the maximum *n added* (six, eight, or ten) than at an *n added* of 1. Larger sizes of *n added* can make it impossible to stop the experiment at the absolute optimal sample size

replacement of lost data up to 40 % of sample size (Fitts, 2010b). Replacement subjects can be provided in subsequent iterations in groups of *n added* until the upper bound is reached. These events affect power more than Type I errors.

Investigators who need to determine power for α levels other than .05, or who are interested in more detailed information on sample sizes required in various statistical or sample size models, should consult Supplement 1 (ESM), which contains all normative data at the finest level of granularity (about 71,000 records). For example, the power and average sample size at each individual level of *n added* is given in Supplement 1 (i.e., information that is not available in the figures). Supplement 2 (ESM) is a help file that shows how to filter the database to easily find information for any experimental circumstance. Supplement 1 provides the complete criteria for all 16 existing sample size models of the vcSSR without consulting other publications. There is a unique pair of criteria for each of four levels of α at each level of *n added* for each of 16 sample size models (472 total pairs of criteria). Using a database application to filter the data on those variables will easily reveal the complete set of criteria, and it will also aid in selecting the appropriate model based on the desired power of the test.

## References

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, 37, 379–384.

Botella J, Ximenez C, Revuelta J, Suero M. (2006). Optimization of sample size in controlled experiments: The CLAST rule. *Behavior Research Methods* 38:65–76.

Braschi, L., Botella, J., & Suero, M. (2014) Consequences of sequential sampling for meta-analysis. *Behavior Research Methods*, 46, 1167-1183. DOI https://doi.org/10.3758/s13428-013-0433-z

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

Cole, D.A., Maxwell, S.E., Arvey, R., & Salas, E. (1994). How the power of MANOVA can both increase and decrease as a function of the intercorrelations among the dependent variables. *Psychological Bulletin*, 115, 465-474.

Corey, D.M., Dunlap, W.P., & Burke, M.J. (1998). Averaging correlations: Expected values and bias in combined Pearson *r*s and Fisher's *z* transformations. *Journal of General Psychology*, 125, 245-261.

Everitt, B.S. (1995). The analysis of repeated measures: A practical review with examples. *The Statistician*, 44, 113-136.

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. Behavior Research Methods 39, 175-191.

Fitts, D.A. (2010a). Improved stopping rules for the design of efficient small-sample experiments in biomedical and biobehavioral research. *Behavior Research Methods,* 42, 3-22. doi: https://doi.org/10.3758/BRM.42.1.3

Fitts, D.A. (2010b). The variable-criteria sequential stopping rule: Generality to unequal sample sizes, unequal variances, or to large ANOVAs. *Behavior Research Methods,* 42, 918-929. doi: https://doi.org/10.3758/BRM.42.4.918

Fitts, D.A. (2011a). Ethics and animal numbers: Informal analyses, uncertain sample sizes, inefficient replications, and type I errors. *Journal of the American Association of Laboratory Animal Science*, 50, 445-453.

Fitts, D.A. (2011b). Minimizing animal numbers: The variable-criteria sequential stopping rule. *Comparative Medicine*, 61, 206-218.

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

García-Pérez, M.A. (2012). Statistical conclusion validity: Some common threats and simple remedies. *Frontiers in Psychology*, 3, 325. doi: https://doi.org/10.3389/fpsyg.2012.00325

Hannaman, M.R., Fitts, D.A., Doss, R.M., Weinstein, D.E., & Bryant, J.L. (2016) The refined biomimetic NeuroDigm GEL™ Model of neuropathic pain in the mature rat. *F1000Research* 5:2516 doi: 10.12688/f1000research.9544.2

Keselman, H.J., Algina, J., & Kowalchuk, R.K. (2001). The analysis of repeated measures designs: A review. *British Journal of Mathematical and Statistical Psychology*, 54, 1-20.

McNemar, Q. (1969). *Psychological statistics*, (4th ed.). New York: John Wiley & Sons.

Oberfeld, D. & Franke, T. (2013). Evaluating the robustness of repeated measures analyses: The case of small sample sizes and nonnormal data. *Behavior Research Methods*, 45, 792-812. DOI https://doi.org/10.3758/s13428-012-0281-2

Press, W.H., Teukolsky, S.A., Vetterling, W.T., & Flannery, B.P. (1992). *Numerical recipes in C: The art of scientific computing* (2nd ed.). Cambridge, MA: Cambridge University Press

Rao, C. R. (1973). *Linear statistical inference and its applications* (2nd Ed.). New York: Wiley.

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