

## Chapter 607

# Multi-Arm Equivalence Tests for the Ratio of Treatment and Control Proportions in a Cluster-Randomized Design

---

## Introduction

This module computes power and sample size for multi-arm equivalence tests of the ratio of treatment and control proportions when the binary data are gathered from a cluster-randomized design. The formulas are based on results in Donner and Klar (2000) and Machin, Campbell, Tan, and Tan (2018).

A *cluster (group) randomized design* is one in which whole units, or clusters, of subjects are randomized to the groups rather than the individual subjects in those clusters. The conclusions of the study concern individual subjects rather than the clusters. Examples of clusters are families, school classes, neighborhoods, hospitals, and doctor's practices.

Cluster-randomized designs are often adopted when there is a high risk of contamination if cluster members were randomized individually. For example, it may be difficult for doctors to use two treatment methods in their practice. The price of randomizing by clusters is a loss of efficiency—the number of subjects needed to obtain a certain level of precision in a cluster-randomized trial is usually much larger than the number needed when the subjects are randomized individually. Hence, standard methods of sample size estimation cannot be used.

In this multi-arm design, there are  $G$  treatment groups and one control group. A proportion is measured in each group. A total of  $G$  hypothesis tests are anticipated each comparing a treatment group with the common control group using a TOST z-test of the ratio of two proportions.

The Bonferroni adjustment of the type I error rate may be optionally made because several comparisons are being tested using the same data. Making a multiplicity adjustment is usually recommended, but not always. In fact, Saville (1990) advocates not applying it and Machin, Campbell, Tan, and Tan (2018) include omitting it as a possibility.

---

## Example

Suppose that the current treatment for a disease works 70% of the time. Unfortunately, this treatment is expensive and occasionally exhibits serious side-effects. Two promising new treatments have been developed and are now ready to be tested. Two equivalence hypotheses need to be tested in this study: one for each treatment. Hence, three groups are needed to complete this study of the two new treatments.

Clinicians are willing to adopt a new treatment if it is about as effective as the current treatment. They must determine, however, how much more or less effective the new treatment can be and still be adopted. Should it be adopted if 65% respond? 68%? 72%? 74%? There are percentages above and below 70% at which the size of the ratio of the two treatments is no longer considered ignorable. After thoughtful discussion with several clinicians, it is decided that if a response ratio of between 0.8 and 1.25 is achieved, the new treatment will be adopted.

## Multi-Arm Equivalence Tests for the Ratio of Treatment and Control Proportions in a Cluster-Randomized Design

The developers must design an experiment to test the hypothesis that the response rate of the new treatment is between 0.8 and 1.25.

## Multiple Treatments Versus a Single Control

Whether you want to test several doses of a single treatment or several types of treatments, good research practice requires that each treatment be compared with a control. For example, a three-arm design consists of three groups: control, treatment A, and treatment B. Two equivalence tests are run: treatment A versus control and treatment B versus the same control. This design avoids having to obtain a second control group for treatment B.

## Technical Details

Our formulation for cluster-randomized designs comes from Donner and Klar (2000).

Denote a binary (0, 1) observation by  $Y_{ikj}$  where  $i$  is one of  $G+1$  groups,  $k = 1, 2, \dots, K_i$  is a cluster within group  $i$ , and  $j = 1, 2, \dots, M_i$  is an item (often a subject) in cluster  $k$  of group  $i$ . The results that follow assume an equal number of items per cluster per group. When the number of items from cluster to cluster are about the same, the power and sample size values should be fairly accurate. In these cases, the average number of items per cluster can be used.

The statistical hypothesis that is tested concerns the ratio of the treatment group proportion and the control group proportion:  $P_i$  and  $P_C$ . With a simple modification, the large-sample sample size formulas that are listed in the module for testing two proportions can be used here.

The  $G$  equivalence-test hypotheses are

$$H_{0i}: P_i/P_C \leq R0.L \text{ or } P_i/P_C \geq R0.U \quad \text{vs.} \quad H_{1i}: R0.L < P_i/P_C < R0.U \quad \text{for } i = 1, 2, \dots, G$$

where  $R0.L$  and  $R0.U$  are the equivalence limits.

If we define  $R_i = P_i/P_C$ , these are equivalent to

$$H_{0i}: R_i \leq R0.L \text{ or } R_i \geq R0.U \quad \text{vs.} \quad H_{1i}: R0.L < R_i < R0.U \quad \text{for } i = 1, 2, \dots, G$$

When the items are randomly assigned to one of the  $G + 1$  groups, the variance of the sample proportion is

$$\sigma_{S,i}^2 = \frac{P_i(1 - P_i)}{N_i}$$

When the randomization is by clusters of items, the variance of the sample proportion is

$$\begin{aligned} \sigma_{C,i}^2 &= \frac{P_i(1 - P_i)(1 + (M_i - 1)\rho)}{K_i M_i} \\ &= \sigma_{S,i}^2 [1 + (M_i - 1)\rho] \\ &= F_{i,\rho} \sigma_{S,i}^2 \end{aligned}$$

## Multi-Arm Equivalence Tests for the Ratio of Treatment and Control Proportions in a Cluster-Randomized Design

The factor  $[1 + (M_i - 1)\rho]$  is sometimes called the *inflation factor*. The Greek letter  $\rho$  is used to represent the *intracluster correlation coefficient (ICC)*. This correlation may be thought of as the simple correlation between any two subjects within the same cluster. If we stipulate that  $\rho$  is positive, it may also be interpreted as the proportion of total variability that is attributable to differences between clusters. This value is critical to the sample size calculation.

The asymptotic formulas that were used in comparing two proportions (see Chapter 200, "Tests for Two Proportions") may be used with cluster-randomized designs as well, as long as an adjustment is made for the inflation factor. The basic form of the TOST z-tests becomes

$$z_L = \frac{|\hat{D}|}{\hat{\sigma}_{\hat{D}}(R0.L)}$$

where

$$\hat{D} = \hat{p}_i - \hat{p}_c(D0.L)$$

$$\hat{\sigma}_{\hat{D}}(R0.L) = \sqrt{\frac{\tilde{p}_i(1 - \tilde{p}_i)F_{i,\rho}}{N_i} + \frac{\tilde{p}_c(1 - \tilde{p}_c)F_{c,\rho}}{N_c}}$$

The quantities  $\tilde{p}_i$  and  $\tilde{p}_c$  are the maximum likelihood estimates constrained by  $\tilde{p}_i = \tilde{p}_c(R0.L)$ .

## Test Statistics

The Farrington and Manning Likelihood-Score test statistic is used in this routine.

### Farrington and Manning's Likelihood Score Test

Farrington and Manning (1990) proposed a test statistic for testing whether the ratio is equal to a specified value  $R$ . The regular MLE's,  $\hat{p}_i$  and  $\hat{p}_c$ , are used in the numerator of the score statistic while MLE's  $\tilde{p}_i$  and  $\tilde{p}_c$ , constrained so that  $\tilde{p}_i = \tilde{p}_c(R0.L)$ , are used in the denominator. The significance level of the test statistic is based on the asymptotic normality of the score statistic.

The formula for computing the test statistic is

$$z_{FMD} = \frac{\hat{p}_i - \hat{p}_c(R0.L)}{\sqrt{\left(\frac{\tilde{p}_i \tilde{q}_i}{N_i/F_{i,\rho}} + \frac{\tilde{p}_c \tilde{q}_c}{N_c/F_{c,\rho}}\right)}}$$

where the estimates  $\tilde{p}_i$  and  $\tilde{p}_c$  are computed as in the corresponding test of Miettinen and Nurminen (1985).

## Multiplicity Adjustment

Because  $G$  TOST z-tests between treatment groups and the control group are run when analyzing the results of this study, most statisticians recommend that a Bonferroni adjustment be applied. This adjustment is easy to apply: the value of alpha that is used in the test is found by dividing the original alpha by the number of tests. For example, if the original alpha is set at 0.05 and the number of treatment (not including the control) groups is five, the individual tests should be conducted using an alpha of 0.01.

The main criticism of this procedure is that if there are many groups, the value of alpha becomes very small. To mitigate against this complaint, some statisticians recommend separating the treatment groups into those that are of primary interest and those that are of secondary interest. The Bonferroni adjustment is made using the number of primary treatments rather than the total number of treatments.

There are some who advocate ignoring the adjustment entirely in the case of randomized clinical trials. See for example Saville (1990) and the discussion in chapter 14 of Machin, Campbell, Tan, and Tan (2018).

---

## Size of the Control Group

Because the control group is used over and over, some advocate increasing the number of clusters in this group. The standard adjustment is to include  $\sqrt{G}$  clusters in the control group for each cluster in one of the treatment groups. See Machin, Campbell, Tan, and Tan (2018, pages 231-232). Note that often, the treatment groups all have the same sample size.

## Example 1 – Finding the Sample Size

A cluster-randomized, multi-arm trial is being designed to investigate the equivalence of two treatments to the standard drug in patients with a specific type of disease. They plan to use the Farrington and Manning likelihood score test to analyze the data.

Historically, the standard treatment has enjoyed a 70% cure rate. The new treatments both reduce the cost and the seriousness of certain side effects of the standard treatment. Thus, either of the new treatments will be adopted even if they are found to be equivalent to the standard treatment. The researchers will recommend adoption of the either of the new treatments if they exhibit a ratio with the standard treatment that is between 0.8 and 1.25. The investigators would like a sample size large enough to find statistical significance at the 0.05 level and the power is 0.90 in each test. Based on similar studies, they estimate the intracluster correlation to be 0.002.

The researchers will recruit patients from various hospitals. All patients at a particular hospital will receive the same treatment. They anticipate an average of 20 patients per hospital. They want to see the impact on cluster count of having cluster sizes ranging for 10 to 30. Since the control group will be used twice, they set the control group allocation ratio to  $\sqrt{G} = \sqrt{2} = 1.414$ . The two treatment allocation ratios are set to 1.0.

### Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

#### Design Tab

Solve For .....	<b>Sample Size</b>
Power of Each Test .....	<b>0.9</b>
Overall Alpha .....	<b>0.05</b>
Bonferroni Adjustment .....	<b>Standard Bonferroni</b>
Group Allocation .....	<b>Enter Group Allocation Pattern, solve for group numbers of clusters</b>
M (Items Per Cluster).....	<b>10 20 30</b>
R0.U (Upper Equivalence Ratio).....	<b>1.25</b>
R0.L (Lower Equivalence Ratio) .....	<b>1/R0.U</b>
Control Proportion.....	<b>0.7</b>
Control Items Per Cluster.....	<b>M</b>
Control Cluster Allocation .....	<b>1.414</b>
Set A Number of Groups.....	<b>2</b>
Set A Proportion .....	<b>0.7</b>
Set A Items Per Cluster .....	<b>M</b>
Set A Cluster Allocation .....	<b>1</b>
Set B Number of Groups.....	<b>0</b>
Set C Number of Groups .....	<b>0</b>
Set D Number of Groups .....	<b>0</b>
More.....	<b>Unchecked</b>
$\rho$ (Intracluster Correlation) .....	<b>0.002</b>

## Output

Click the Calculate button to perform the calculations and generate the following output.

### Numeric Reports

#### Numeric Results

Solve For: **Sample Size**  
 Group Allocation: Enter Group Allocation Pattern, solve for group numbers of clusters  
 Hypothesis:  $H_0: P_i / P_c \leq R_{0.L}$  or  $P_i / P_c \geq R_{0.U}$  vs.  $H_1: R_{0.L} < P_i / P_c < R_{0.U}$   
 Test Type: Farrington and Manning Likelihood Score Test  
 Number of Groups: 3  
 Bonferroni Adjustment: Standard Bonferroni (Divisor = 2)

Comparison	Power		Number of Clusters Ki	Cluster Allocation	Items Per Cluster Mi	Sample Size Ni	Proportion Pi	Ratio Ri	Equivalence Limits		ICC ρ	Alpha	
	Target	Actual							Lower R0.L	Upper R0.U		Overall	Bonferroni-Adjusted
Control			30	1.414	10	300	0.7		0.8	1.25	0.002		
vs A1	0.9	0.91437	21	1.000	10	210	0.7	1	0.8	1.25	0.002	0.05	0.025
vs A2	0.9	0.91437	21	1.000	10	210	0.7	1	0.8	1.25	0.002	0.05	0.025
Total			72			720							
Control			16	1.414	20	320	0.7		0.8	1.25	0.002		
vs A1	0.9	0.92531	11	1.000	20	220	0.7	1	0.8	1.25	0.002	0.05	0.025
vs A2	0.9	0.92531	11	1.000	20	220	0.7	1	0.8	1.25	0.002	0.05	0.025
Total			38			760							
Control			10	1.414	30	300	0.7		0.8	1.25	0.002		
vs A1	0.9	0.90078	7	1.000	30	210	0.7	1	0.8	1.25	0.002	0.05	0.025
vs A2	0.9	0.90078	7	1.000	30	210	0.7	1	0.8	1.25	0.002	0.05	0.025
Total			24			720							

Comparison: The group that is involved in the comparison between the treatment and control displayed on this report line. The comparison is made using the ratio.

Target Power: The power desired. Power is probability of rejecting a false null hypothesis for this comparison. This power is of the comparison shown on this line only.

Actual Power: The power actually achieved.

Ki: The number of clusters in the ith group. The total number of clusters is reported in the last row of the column.

Allocation: The cluster allocation ratio of the ith group. The value on each row represents the relative number of clusters assigned to the group.

Mi: The average number of items per cluster (or average cluster size) in the ith group.

Ni: The number of items in the ith group. The total sample size is shown as the last row of the column.

Pi: The response proportion in the ith group at which the power is calculated.

Ri: The ratio of the ith group proportion (Pi) and the control group proportion (Pc) at which the power is calculated. The formula is  $R_i = P_i / P_c$ .

R0.L: The lower equivalence ratio is the smallest negative ratio in treatment and control proportions that still results in the conclusion that they are equivalent.

R0.U: The upper equivalence ratio is the largest positive ratio in treatment and control proportions that still results in the conclusion that they are equivalent.

ρ: The intraclass correlation (ICC). The correlation between subjects within a cluster.

Overall Alpha: The probability of rejecting at least one of the comparisons in this experiment when each null hypothesis is true.

Bonferroni Alpha: The adjusted significance level at which each individual comparison is made.

## Multi-Arm Equivalence Tests for the Ratio of Treatment and Control Proportions in a Cluster-Randomized Design

**Summary Statements**

---

A parallel, 3-group cluster-randomized design (with one control group and 2 treatment groups) will be used to test whether the proportion for each treatment group is equivalent to the control group proportion, with equivalence ratio bounds of 0.8 and 1.25 ( $H_0: R \leq 0.8$  or  $R \geq 1.25$  versus  $H_1: 0.8 < R < 1.25$ ,  $R = P_i / P_c$ ). Each of the 2 equivalence comparisons will be made using two one-sided, two-sample, Bonferroni-adjusted (divisor = 2) Farrington and Manning likelihood score tests. The overall (experiment-wise) Type I error rate ( $\alpha$ ) is 0.05. The control group proportion is assumed to be 0.7. The intracluster correlation is assumed to be 0.002. The average cluster size (number of subjects or items per cluster) for the control group is assumed to be 10, and the average cluster size for each of the treatment groups is assumed to be 10 and 10. To detect the treatment proportions 0.7 and 0.7 with at least 90% power for each test, the control group cluster count needed will be 30 and the number of needed clusters for the treatment groups will be 21 and 21 (totaling 72 clusters overall).

---

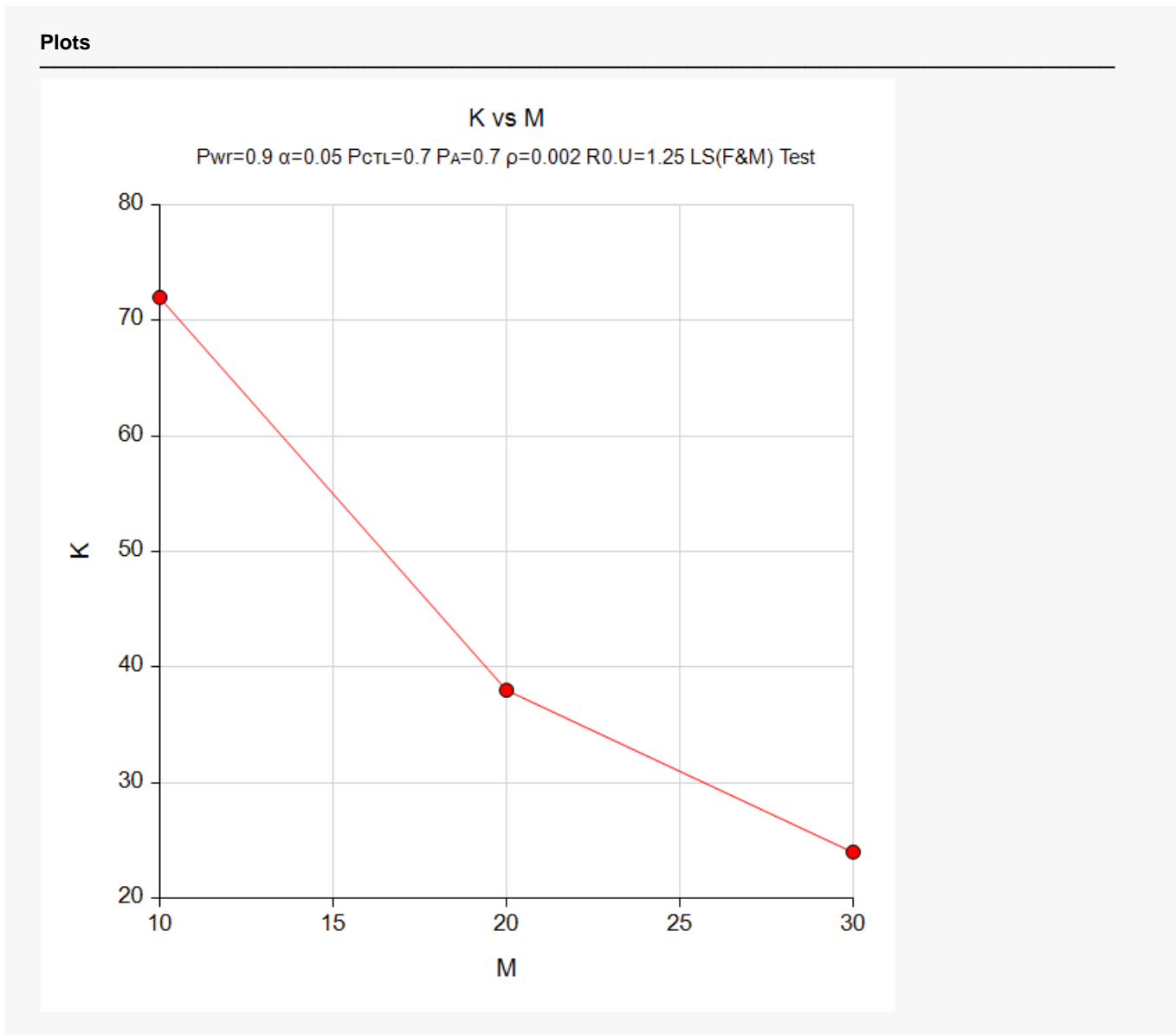
**References**

---

- Blackwelder, W.C. 1998. 'Equivalence Trials.' In Encyclopedia of Biostatistics, John Wiley and Sons. New York. Volume 2, 1367-1372.
- Chow, S.C. and Liu, J.P. 1999. Design and Analysis of Bioavailability and Bioequivalence Studies. Marcel Dekker. New York.
- Chow, S.C., Shao, J., Wang, H., and Lohknygina, Y. 2018. Sample Size Calculations in Clinical Research, 3rd Edition. Chapman & Hall/CRC. Boca Raton, FL. Pages 86-88.
- Donner, A. and Klar, N. 2000. Design and Analysis of Cluster Randomization Trials in Health Research. Arnold. London.
- Farrington, C. P. and Manning, G. 1990. 'Test Statistics and Sample Size Formulae for Comparative Binomial Trials with Null Hypothesis of Non-Zero Risk Difference or Non-Unity Relative Risk.' Statistics in Medicine, Vol. 9, pages 1447-1454.
- Gart, John J. and Nam, Jun-mo. 1988. 'Approximate Interval Estimation of the Ratio in Binomial Parameters: A Review and Corrections for Skewness.' Biometrics, Volume 44, Issue 2, 323-338.
- Julious, S. A. and Campbell, M. J. 2012. 'Tutorial in biostatistics: sample sizes for parallel group clinical trials with binary data.' Statistics in Medicine, 31:2904-2936.
- Lachin, J.M. 2000. Biostatistical Methods. John Wiley & Sons. New York.
- Machin, D., Campbell, M.J., Tan, S.B, and Tan, S.H. 2018. Sample Sizes for Clinical, Laboratory, and Epidemiology Studies, 4th Edition. Wiley Blackwell.
- Miettinen, O.S. and Nurminen, M. 1985. 'Comparative analysis of two rates.' Statistics in Medicine 4: 213-226.
- Tubert-Bitter, P., Manfredi, R., Lellouch, J., Begaud, B. 2000. 'Sample size calculations for risk equivalence testing in pharmacoepidemiology.' Journal of Clinical Epidemiology 53, 1268-1274.
- 

This report shows the numeric results of this sample size study. Notice that the results are shown in blocks of four rows at a time. Each block represents a single design.

## Plots Section



This plot gives a visual presentation to the results in the Numeric Report. We can quickly see the impact on the total cluster count, K, of increasing the cluster size, M.



## Example 2 – Validation using a Previously Validated Procedure

We could not find a validation result in the statistical literature, so we will use a previously validated **PASS** procedure (**Equivalence Tests for the Ratio of Two Proportions in a Cluster-Randomized Design**) to produce the results for the following example.

We will use settings similar to those of Example 1 specifically when  $M = 30$ . In this case, the number of clusters allocated to the control group was 10 and to the treatment groups was 7 in each. The Bonferroni adjustment changes the significance level from 0.05 to 0.025.

The **Equivalence Tests for the Ratio of Two Proportions in a Cluster-Randomized Design** procedure is set up as follows.

### Design Tab

Solve For ..... **Power**  
 Alpha..... **0.025**  
 K1 (Clusters in Group 1) ..... **10**  
 M1 (Average Cluster Size)..... **30**  
 K2 (Clusters in Group 2) ..... **7**  
 M2 (Average Cluster Size)..... **M1**  
 R0.U (Upper Equivalence Ratio)..... **1.25**  
 R0.L (Lower Equivalence Ratio) ..... **1/R0.U**  
 R1 (Actual Ratio) ..... **1.0**  
 P2 (Group 2 Proportion)..... **0.7**  
 ICC (Intracluster Correlation)..... **0.002**

This set of options generates the following report.

### Numeric Results

Solve For: **Power**  
 Groups: 1 = Treatment, 2 = Reference  
 Test Statistic: Likelihood Score Test (Farrington & Manning)  
 Hypotheses:  $H_0: P_1 / P_2 \leq R_{0.L} \text{ or } P_1 / P_2 \geq R_{0.U}$  vs.  $H_1: R_{0.L} < P_1 / P_2 < R_{0.U}$

	Number of Clusters			Average Cluster Size		Total Sample Size N	Proportion			Ratio		Intracluster Correlation ICC	Alpha
	K1	K2	K	M1	M2		Equivalence		Actual	Reference	Equivalence		
<b>Power</b>							P1.L   P1.U	P1.1	P2	R0.L   R0.U	R1		
0.90078	10	7	17	30	30	510	0.56   0.875	0.7	0.7	0.8   1.25	1	0.002	0.025

The power is computed to be 0.90078.

## Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab	
Solve For .....	<b>Power</b>
Overall Alpha .....	<b>0.05</b>
Bonferroni Adjustment .....	<b>Standard Bonferroni</b>
Group Allocation .....	<b>Enter the Numbers of Clusters per Group individually</b>
M (Items Per Cluster).....	<b>30</b>
R0.U (Upper Equivalence Ratio).....	<b>1.25</b>
R0.L (Lower Equivalence Ratio) .....	<b>1/R0.U</b>
Control Proportion.....	<b>0.7</b>
Control Items Per Cluster.....	<b>M</b>
Control Number of Clusters .....	<b>10</b>
Set A Number of Groups.....	<b>2</b>
Set A Proportion .....	<b>0.7</b>
Set A Items Per Cluster .....	<b>M</b>
Set A Number of Clusters .....	<b>7</b>
Set B Number of Groups.....	<b>0</b>
Set C Number of Groups .....	<b>0</b>
Set D Number of Groups .....	<b>0</b>
More.....	<b>Unchecked</b>
$\rho$ (Intracluster Correlation) .....	<b>0.002</b>

## Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results											
Solve For:		<a href="#">Power</a>									
Hypothesis:		H0: $P_i / P_c \leq R0.L$ or $P_i / P_c \geq R0.U$ vs. H1: $R0.L < P_i / P_c < R0.U$									
Test Type:		Farrington and Manning Likelihood Score Test									
Number of Groups:		3									
Bonferroni Adjustment:		Standard Bonferroni (Divisor = 2)									
Comparison	Power	Number of Clusters Ki	Items Per Cluster Mi	Sample Size Ni	Proportion Pi	Ratio Ri	Equivalence Limits		ICC $\rho$	Alpha	
							Lower R0.L	Upper R0.U		Overall	Bonferroni-Adjusted
Control		10	30	300	0.7		0.8	1.25	0.002		
vs A1	0.90078	7	30	210	0.7	1	0.8	1.25	0.002	0.05	0.025
vs A2	0.90078	7	30	210	0.7	1	0.8	1.25	0.002	0.05	0.025
Total		24		720							

As you can see, the power is 0.90078 for both treatment groups which match the power found in the validation run above. The procedure is validated.