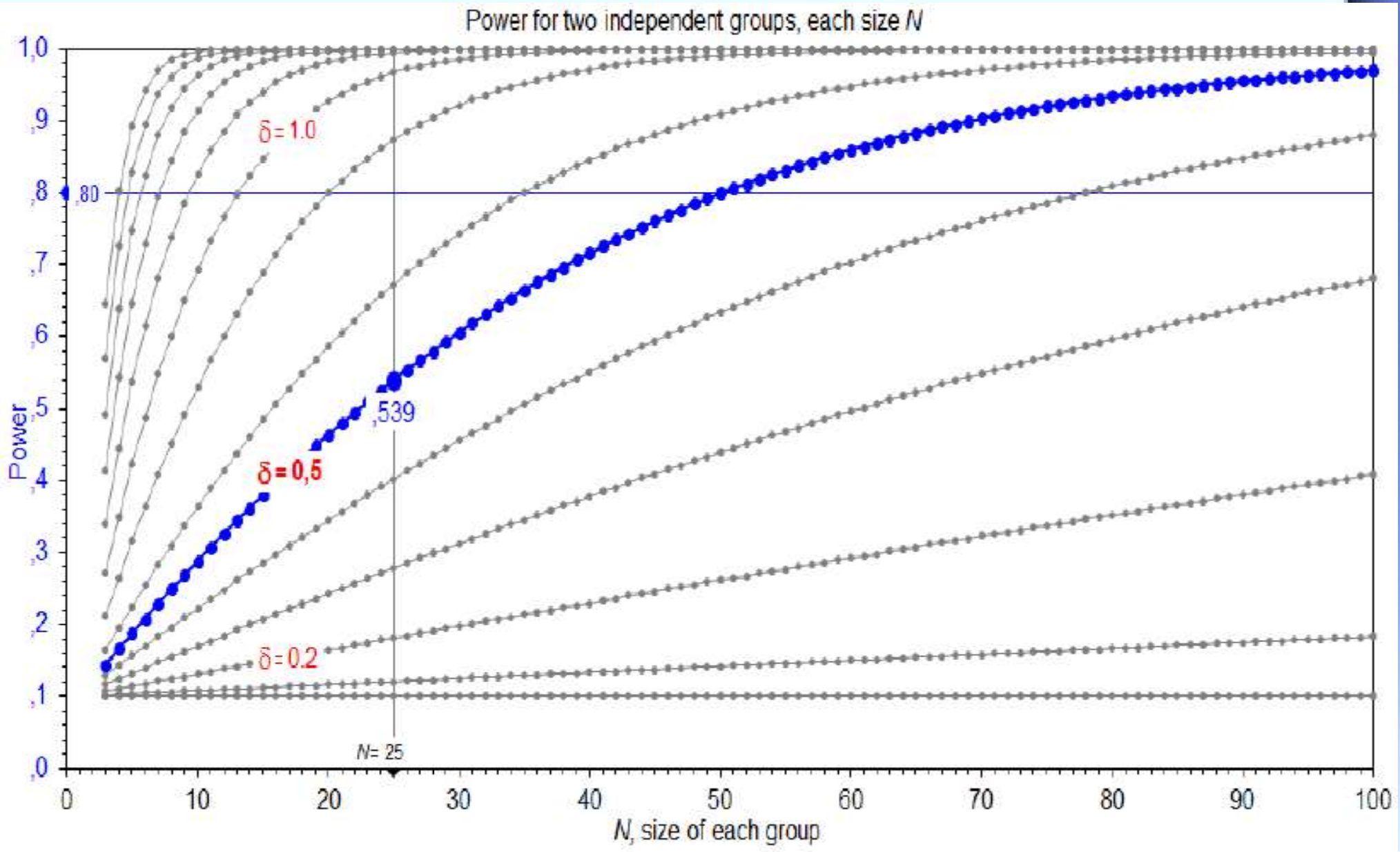
Between vs. Within

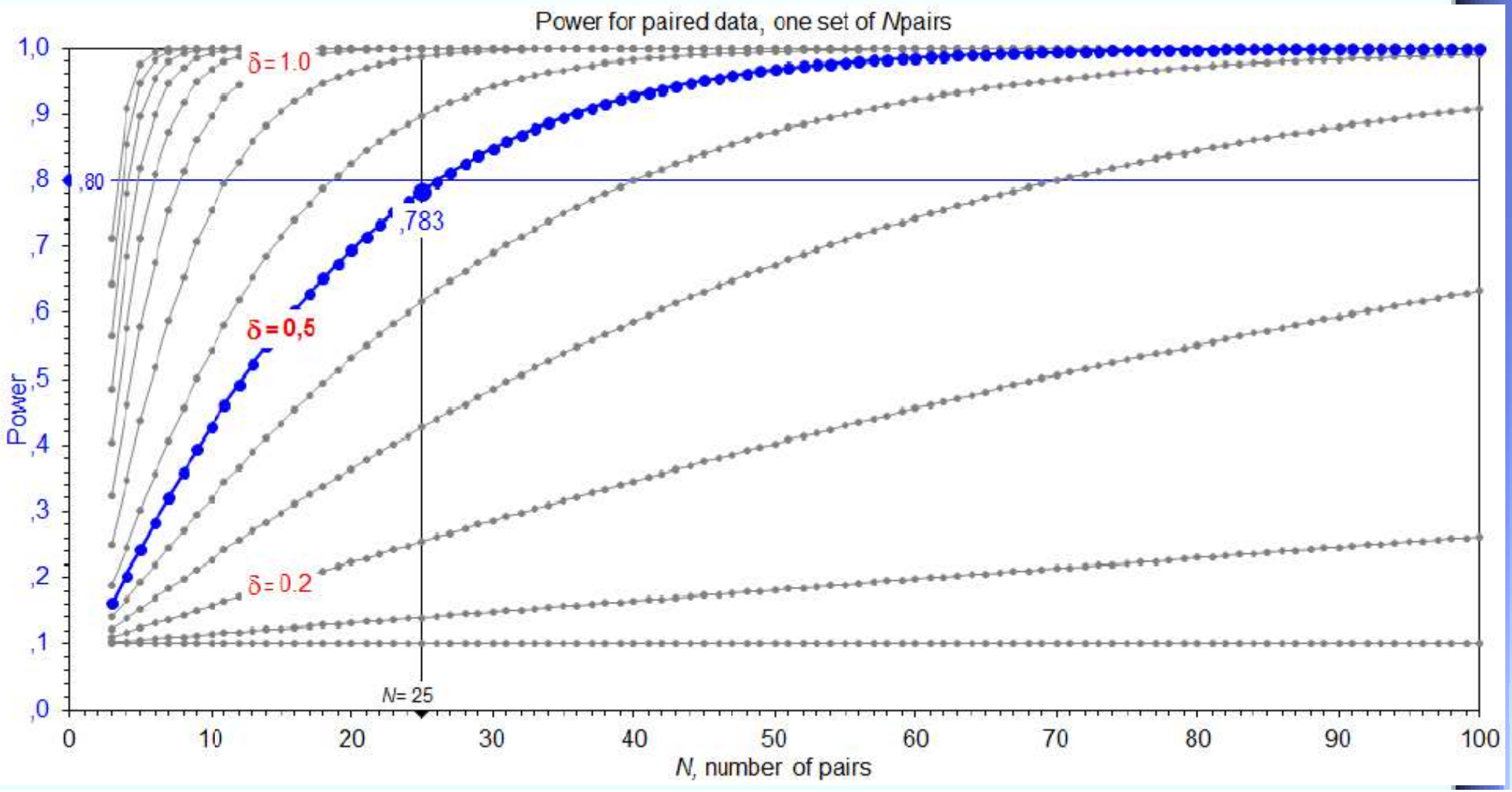
- Everything else being equal, within studies are more powerful than between studies
- Example with a simple two groups/two measures design
- Later on more articulated examples

Power Between Ss



Power Within Ss

- Power for within Ss studies is greater (*ceteris paribus*) but depends also on r (e.g., r = .50) between DVs

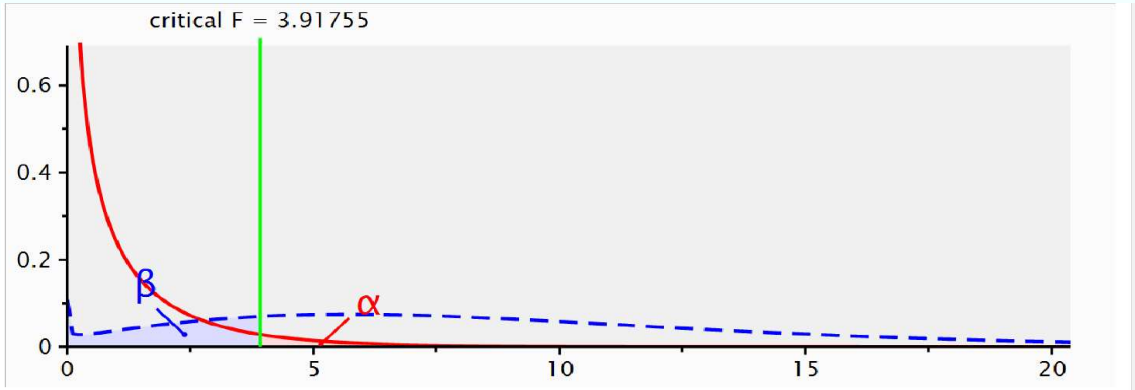


Correspondence between some ES

Effect Size	d	0.20	Effect Size	d	0.50	Effect Size	d	0.80
d	0.2		d	0.5		d	0.8	
r	0.0995		r	0.2425		r	0.3714	
η^2	0.0099		η^2	0.0588		η^2	0.1379	
f	0.1		f	0.25		f	0.4	
Odds Ratio	1.4373		Odds Ratio	2.4766		Odds Ratio	4.2675	

https://www.psychometrica.de/effect_size.html#transform

ANOVA 2 x 2



Analysis

Test family: F tests

Statistical test: ANOVA: Fixed effects, special, main effects and interactions

Type of power analysis: A priori: Compute required sample size – given α, power, and effect size

Inputs

Input Parameters		Output Parameters	
Determine =>		Noncentrality parameter λ	8.0000000
Effect size f	0.25	Critical F	3.9175498
α err prob	0.05	Denominator df	124
Power (1–β err prob)	0.80	Total sample size	128
Numerator df	1	Actual power	0.8013621
Number of groups	4		

Output

Contrast approach (from Means)

More complex designs can be sometimes simplified with a focused contrast approach

The key point is to be explicit about the focal hypothesis and the pattern of expected means and to use contrast weights that reflect the focal hypothesis

CONTRAST WEIGHTS

	A1	A2
B1	1	-1
B2	1	-1

MAIN EFFECT (A) {1, -1, 1, -1}

CONTRAST WEIGHTS

	A1	A2
B1	1	1
B2	-1	-1

MAIN EFFECT (B) {1, 1, -1, -1}

CONTRAST WEIGHTS

	A1	A2
B1	1	-1
B2	-1	1

INTERACTION (AxB) {1, -1, -1, 1}

Contrast approach

Suppose you expect this pattern of means

Table 2: Example of 2 × 2 design.

	A1	A2
B1	10	0
B2	0	0

Note: Pooled standard deviation is equal to 5.

Calculate ***f*** for main effects and interactions, after coding

$$f = \frac{\left|\sum c_i \cdot \mu_i\right|}{\sqrt{k \cdot \sum c_i^2 \cdot \sigma^2}}$$

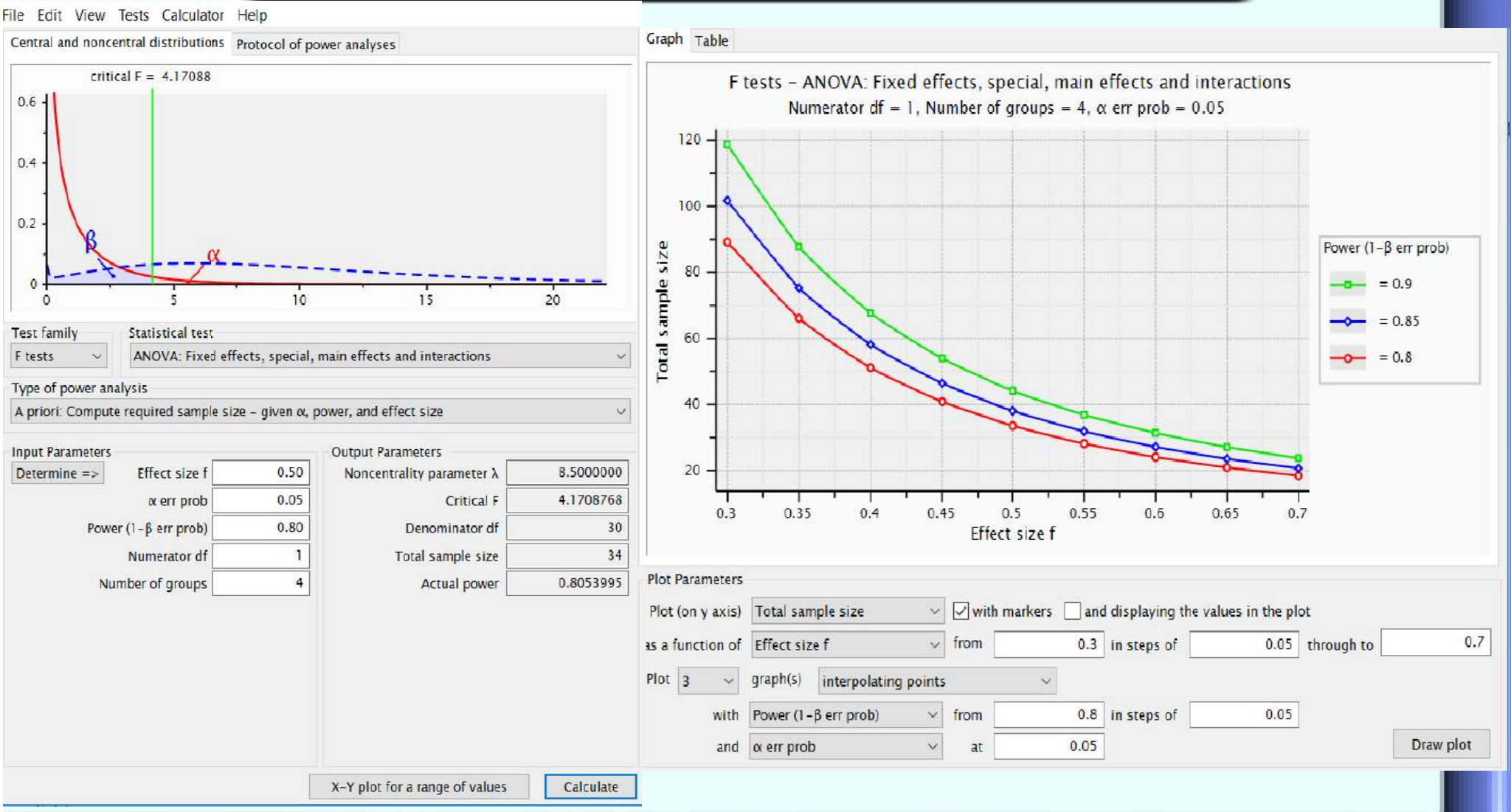
WEIGHTS (c_i)		
	1	1
	-1	-1

$$f = \frac{10}{\sqrt{4 \cdot 4 \cdot 25}} = .50$$

Might be useful to recall that $f = \frac{d}{2}$ or $d = 2f$

Tip: if it helps, you can think with standardized means (e.g., SD=1, transform EMs in standardized expected differences, as for Cohens d)

Plugging it in G*Power



Planning a "moderation" study

A common research scenario

Study 1: Main effect

Study 2: Test of a moderator of the main effect

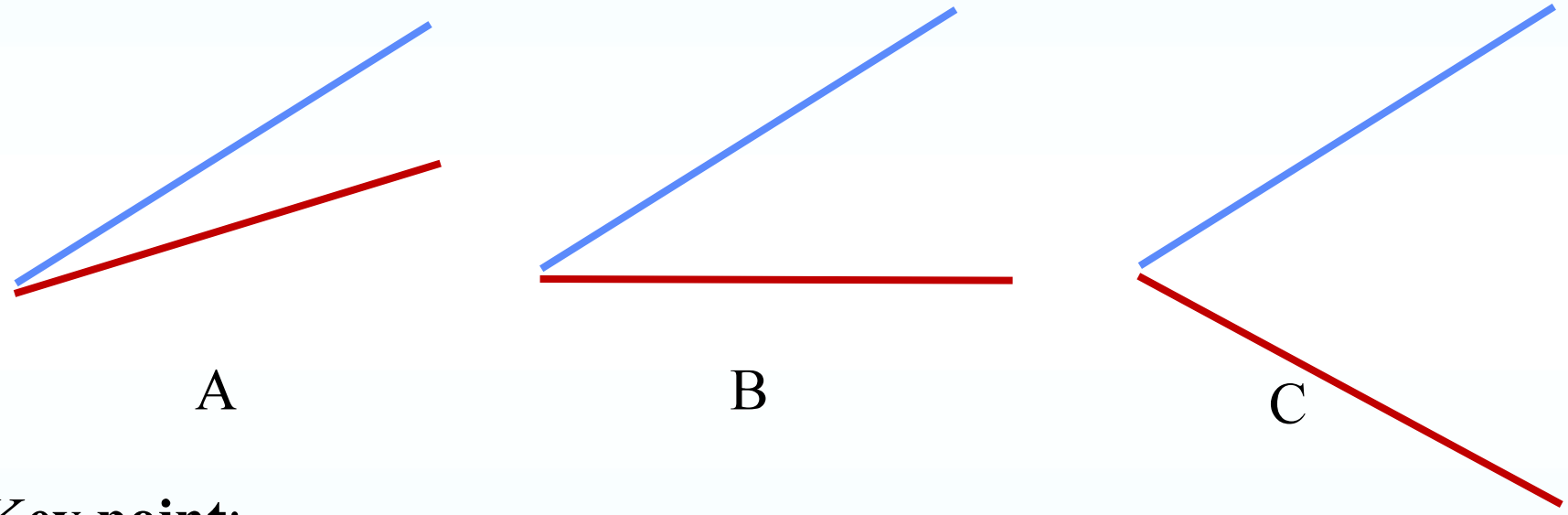
Table 3: Example of 2 × 2 design expected results.

		Case 1		Case 2	
		A1	A2	A1	A2
B	replicated	5	2	5	2
	moderated	0	0	2	5

Note: Pooled standard deviation is equal to 1.

Key point: Power calculations will change depending on design and type of expected moderation

Shape of the moderation effect



Key point:

ES (hence required N) will change depending on the shape

$$N_A >> N_B > N_C$$

Using a contrast approach

Case 1: Expect to replicate main effect and moderator suppresses it

Table 3: Example of 2 × 2 design expected results.

		Case 1	
		A1	A2
B	replicated	5	2
	moderated	0	0

Note: Pooled standard deviation is equal to 1.

Power calculation for main effect in original study (2 conditions, A1 vs. A2)

$$f = \frac{\left|\sum c_i \cdot \mu_i\right|}{\sqrt{k \cdot \sum c_i^2 \cdot \sigma^2}}$$

$$C_A = (1) \cdot 5 + (-1) \cdot 2 = 3$$

$$f = \frac{3}{\sqrt{2 \cdot 2 \cdot 1}} = 1.50$$

Using a contrast approach

Case 1: Expect to replicate main effect and moderator suppresses it

Table 3: Example of 2 × 2 design expected results.

		Case 1	
		A1	A2
B	replicated	5	2
	moderated	0	0

Note: Pooled standard deviation is equal to 1.

Power calculation for interaction effect

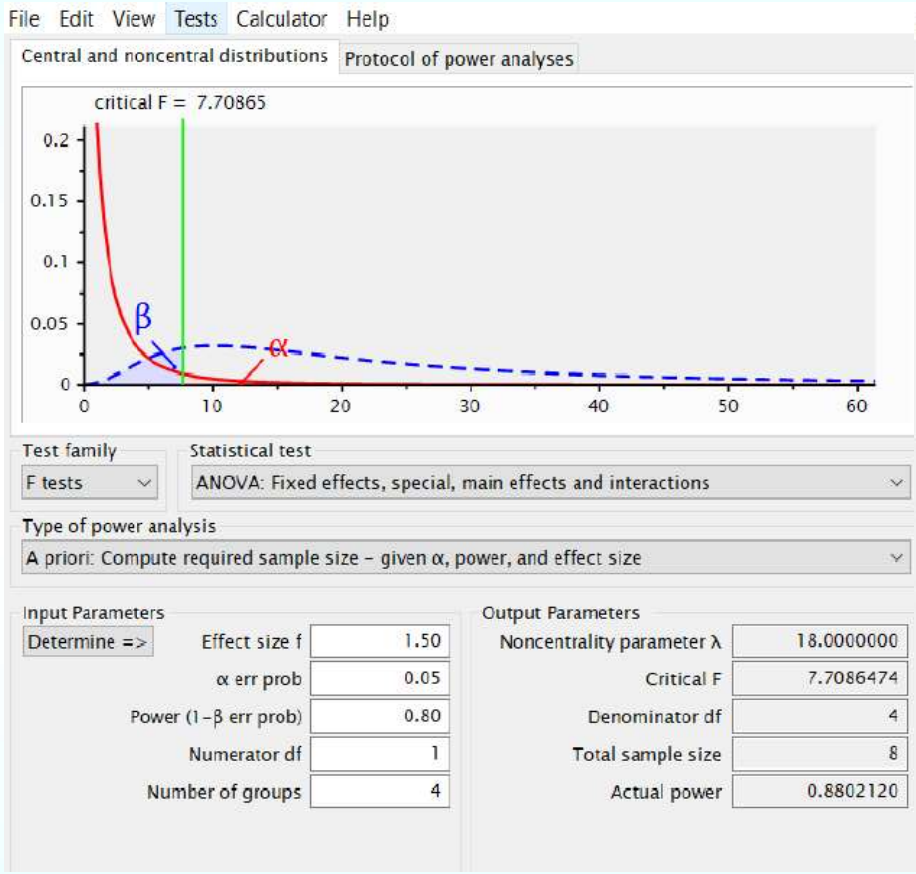
$$f = \frac{\left|\sum c_i \cdot \mu_i\right|}{\sqrt{k \cdot \sum c_i^2 \cdot \sigma^2}}$$

$$C_{AB} = (1) \cdot 5 + (-1) \cdot 2 + (-1) \cdot 0 + (1) \cdot 0 = 3$$

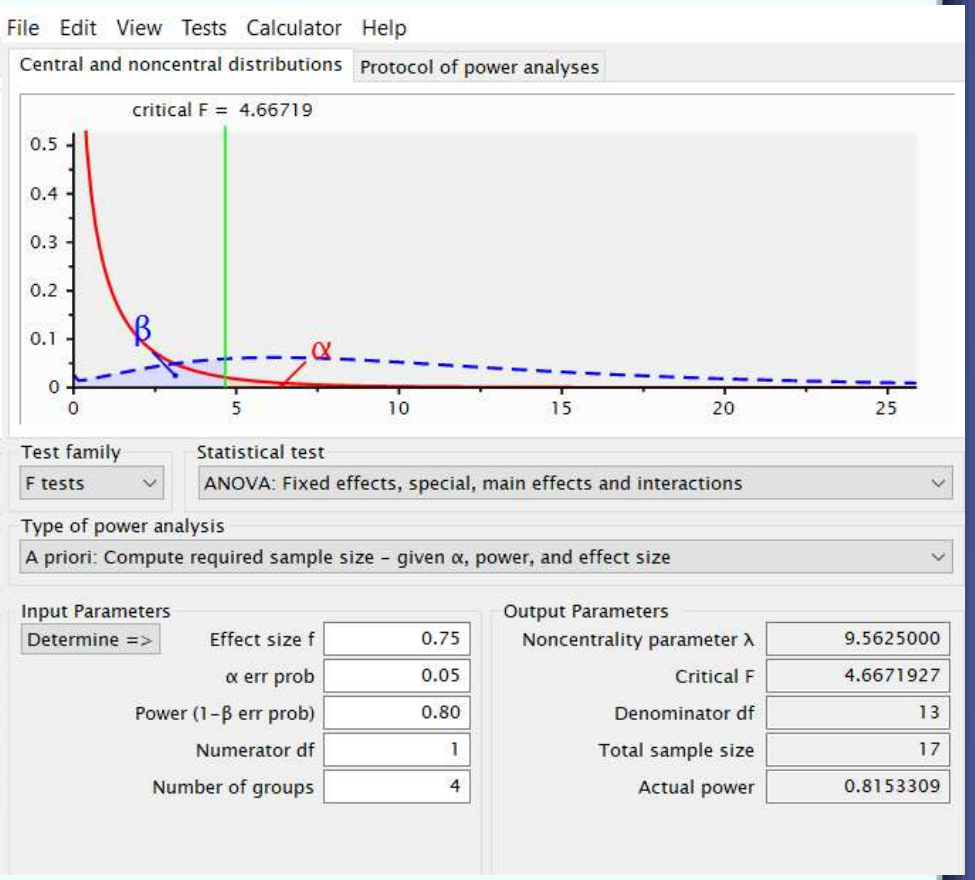
$$f = \frac{3}{\sqrt{4 \cdot 4 \cdot 1}} = 0.75$$

Plugging it in G*Power

Main effect



Moderated (suppression) effect



Needed sample size is about doubled

Using a contrast approach

Case 2: Expect to replicate main effect and moderator reverts it

Table 3: Example of 2 × 2 design expected results.

		Case 2	
		A1	A2
B	replicated	5	2
	moderated	2	5

Note: Pooled standard deviation is equal to 1.

Power calculation for interaction effect

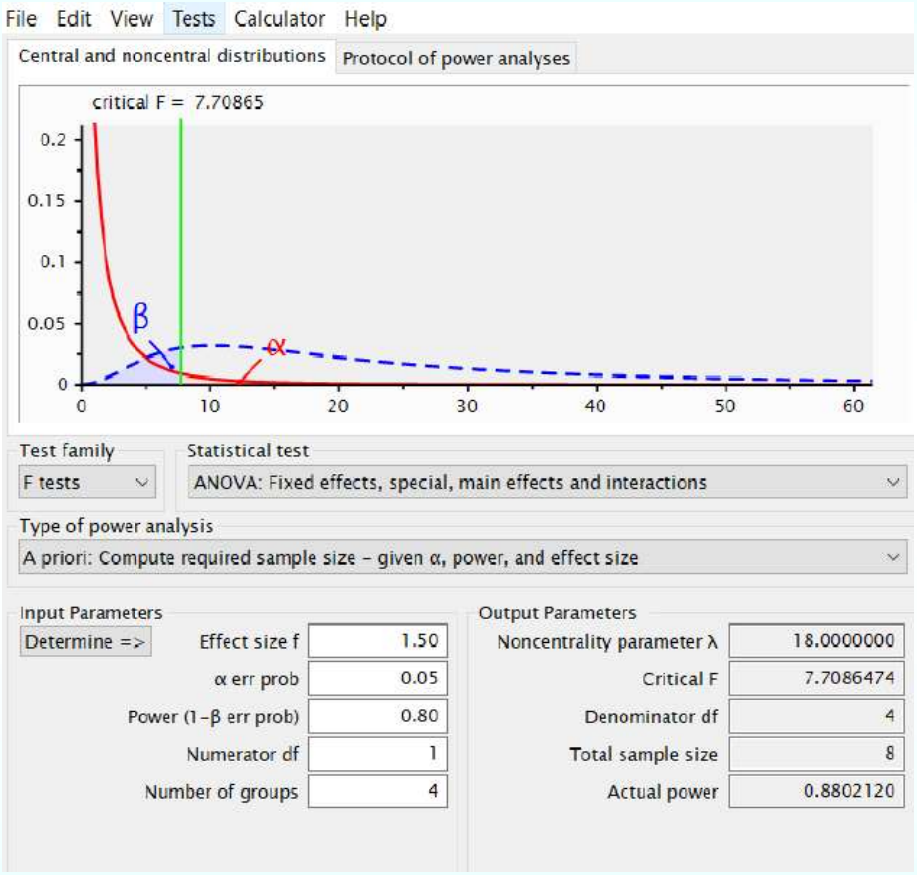
$$f = \frac{\left|\sum c_i \cdot \mu_i\right|}{\sqrt{k \cdot \sum c_i^2 \cdot \sigma^2}}$$

$$C_{AB} = (1) \cdot 5 + (-1) \cdot 2 + (-1) \cdot 2 + (1) \cdot 5 = 6$$

$$f = \frac{6}{\sqrt{4 \cdot 4 \cdot 1}} = 1.50$$

Plugging it in G*Power

Main effect or Moderated (reverted) effect



Needed sample size is the same !

General rule

One can think in terms of **percentage (expected change) of moderation effect**

$$f_n = \frac{p_m}{100} \cdot f_o \cdot \sqrt{\frac{k_o}{k_n \cdot l}}$$

f_n = *expected ES of moderator for planned research*

f_o = *observed ES of original effect*

k_o = *number of cells in original study*

k_n = *number of cells in planned research*

l = *number of levels of moderator*

p_m : *0%= no moderation (replicated effect)*
100%= moderation as suppression
200%= moderation as reverted effect

Applying the rule

$$f_n = \frac{p_m}{100} \cdot f_o \cdot \sqrt{\frac{k_o}{k_n \cdot l}}$$

CASE 1 (suppressed moderation)

$p_m = 100\%$

$f_o = 0.50$

$k_o = 2;\ k_n = 4;\ l = 2$

$$f_n = 1.50 * \frac{1}{2} = 0.75$$

CASE 2 (reverted moderation)

$p_m = 200\%$

$$f_n = 2 * 1.50 * \frac{1}{2} = 1.50$$

CASE 3 (weak suppressed moderation)

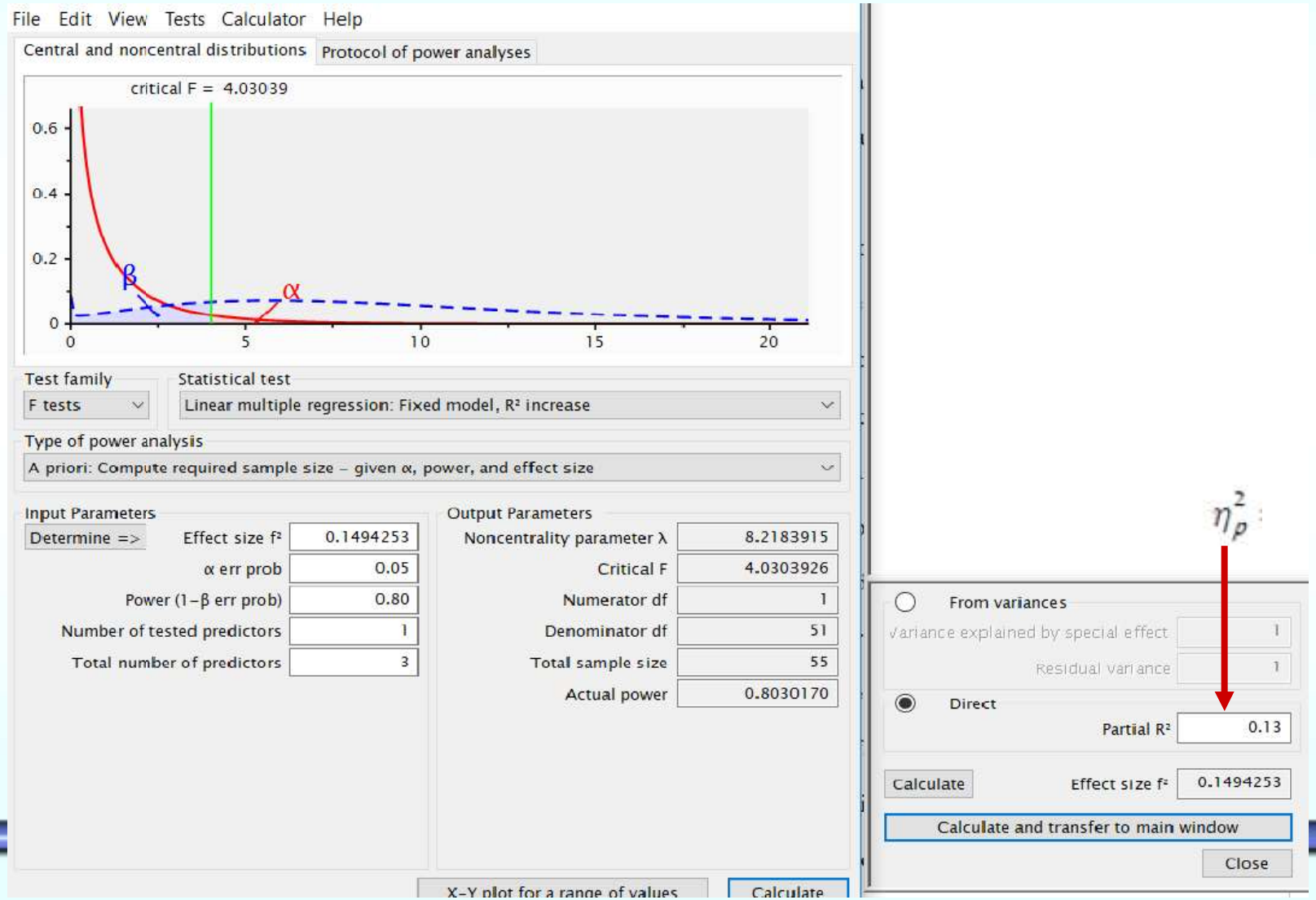
$p_m = 50\%$

$$f_n = 0.50 * 1.50 * \frac{1}{2} = 0.375$$

No need to guess Means and SD !

Regression analysis

Power calculation for a term is straightforward



Moderators in regression analysis

But ES are sometimes not easy to guess

Interaction effects (moderator) can explain not much variance (also for technical reasons) yet be theoretically key

One likely easier way to guess ES is to think in terms of difference between standardized coefficients (β)

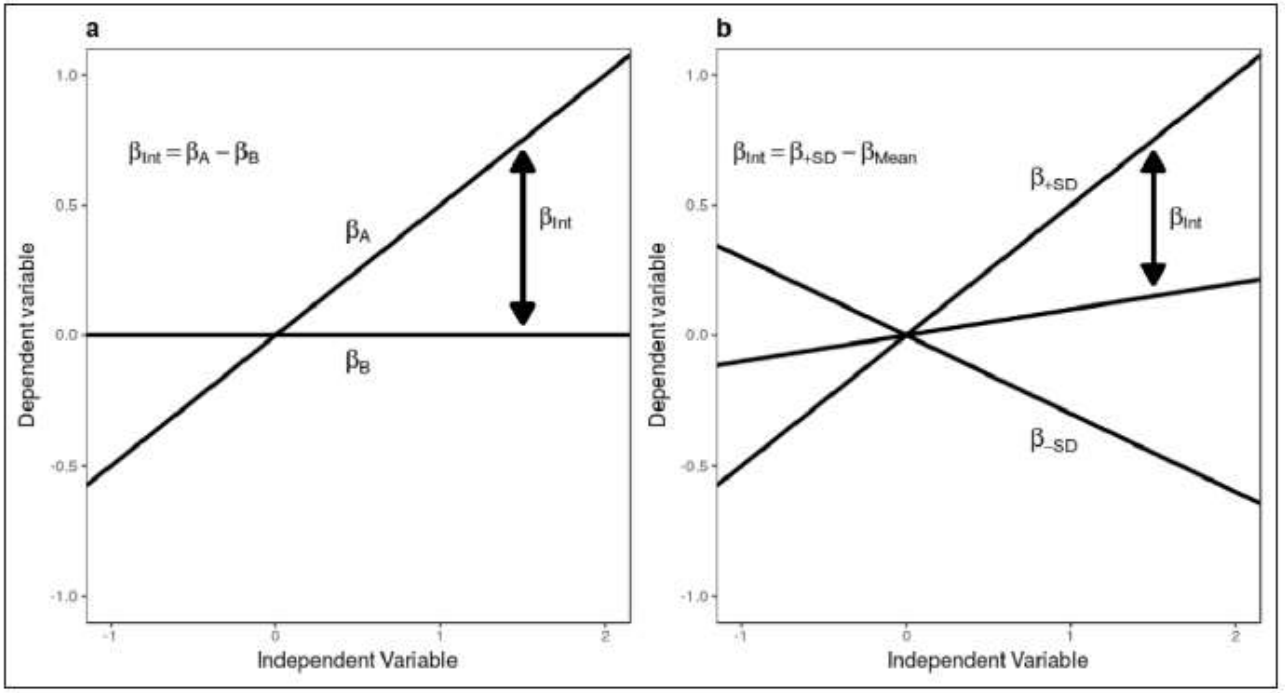


Figure 7: Geometrical interpretation of the interaction beta coefficient, with a dichotomous moderator (a) and a continuous moderator (b).

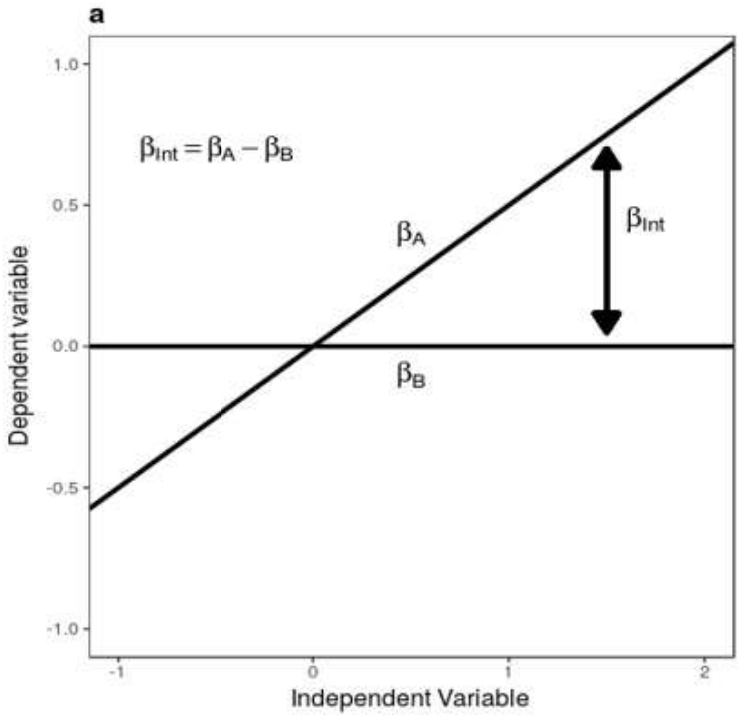
MLR with dichotomous moderator

The moderation effect can be approximated as $$f^2 \approx \frac{\beta_{int}^2}{2 \cdot (2 - r_a^2 - r_b^2)}$$

EFFECT SIZE OF INTERACTION BETWEEN A CONTINUOUS AND A DICHOTOMOUS VARIABLE

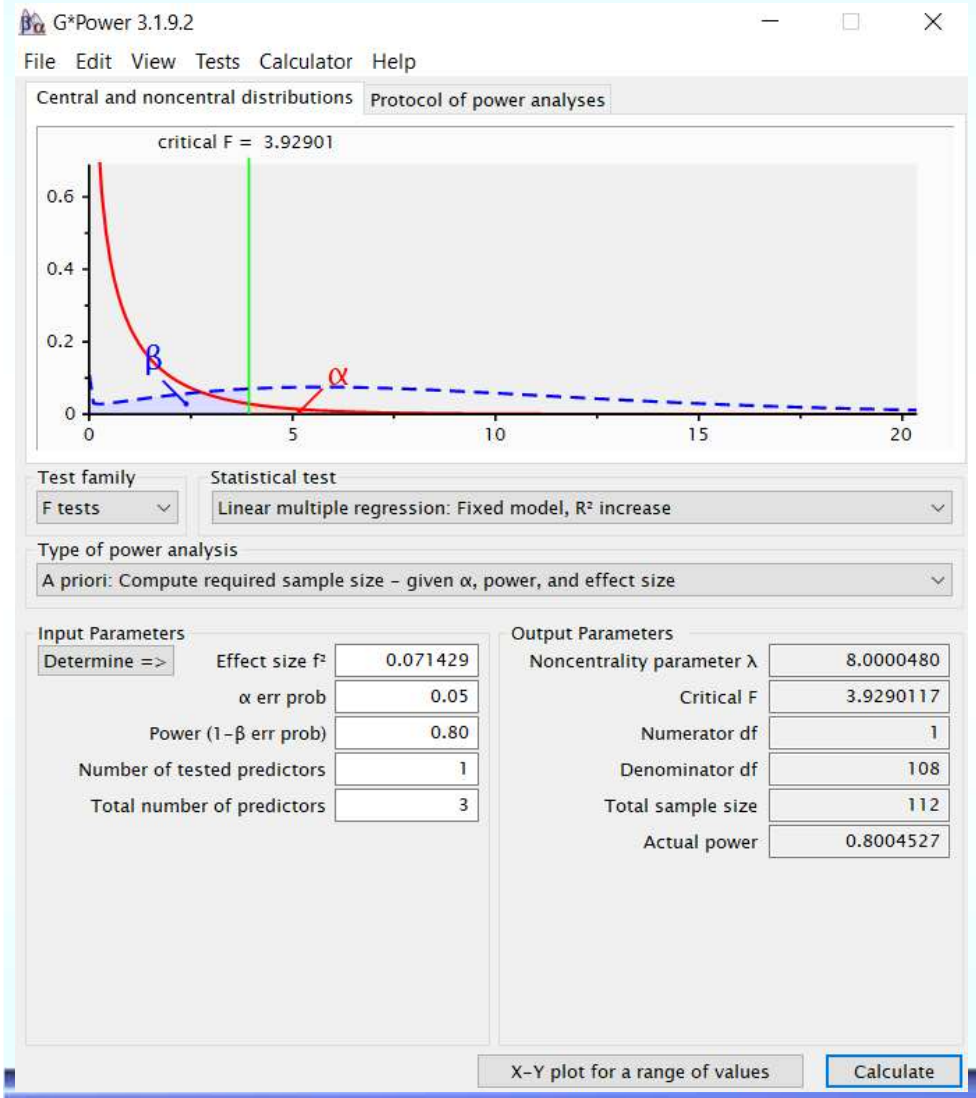
Input the expected correlations between the continuous independent variable X and the dependent variable Y at the two levels A and B of the moderator (referred to as β_a and β_b, or equivalently as r_a and or r_b). In the yellow cells you can see the implied interaction term (β_{int}) and the interaction effect size (f_n)

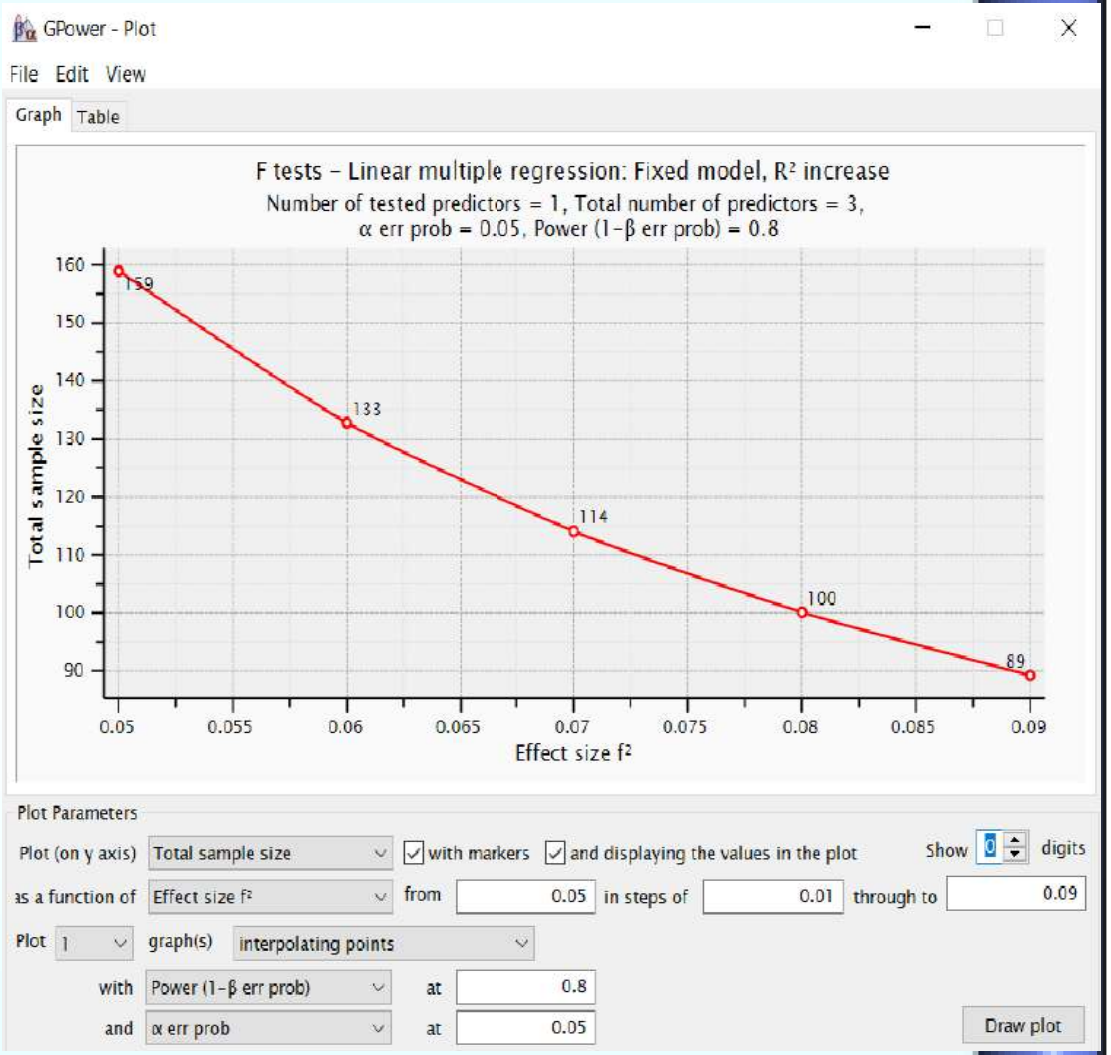
β_a or r_a	0,5
βa or r_b	0
β_{int}	0,5
f_n	0,071429



Routine Excel at https://github.com/mcfanda/primerPowerIRSP

Plugging it in G*Power





MLR with continuous moderator

The moderation effect can be approximated as $f^2 \approx \frac{\beta_{int}^2}{1 - r_{yx}^2 - r_{ym}^2}$

INTERACTION BETWEEN TWO CONTINUOUS VARIABLES

Option 1: Input the values of correlations between the dependent variable Y and the predictor X when the moderator M is equal to its mean (r_{yx}), the correlation between the dependent variable Y and the moderator M (r_{ym}), and the expected percentage of moderation (p_m)

r_{yx}	0,35	
r_{ym}	0,25	
p_m	50	%
β_{int}	0,175	
f_n	0,037577	

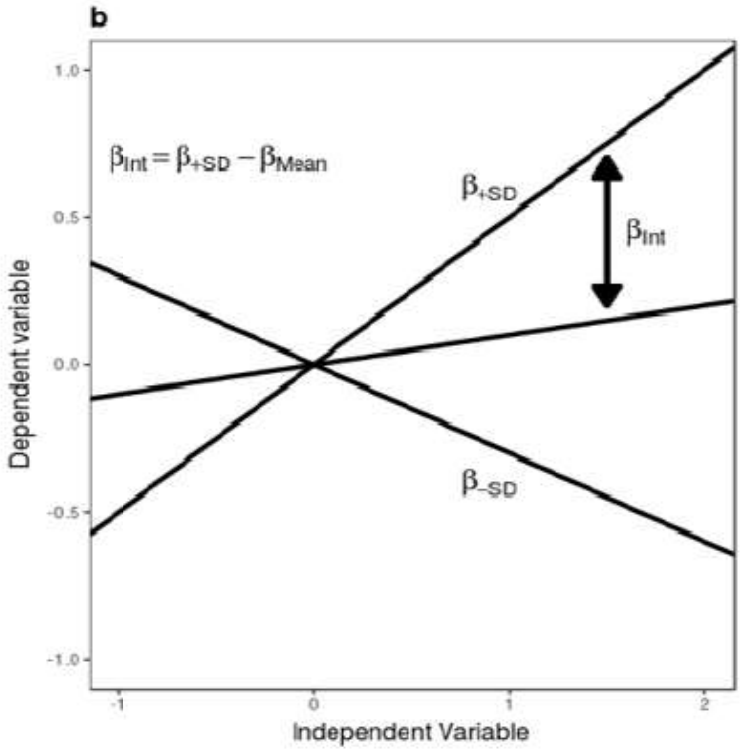
Routine Excel at https://github.com/mcfanda/primerPowerIRSP

MLR with continuous moderator

The moderation effect can be approximated as $f^2 \approx \frac{\beta_{int}^2}{1 - r_{yx}^2 - r_{ym}^2}$

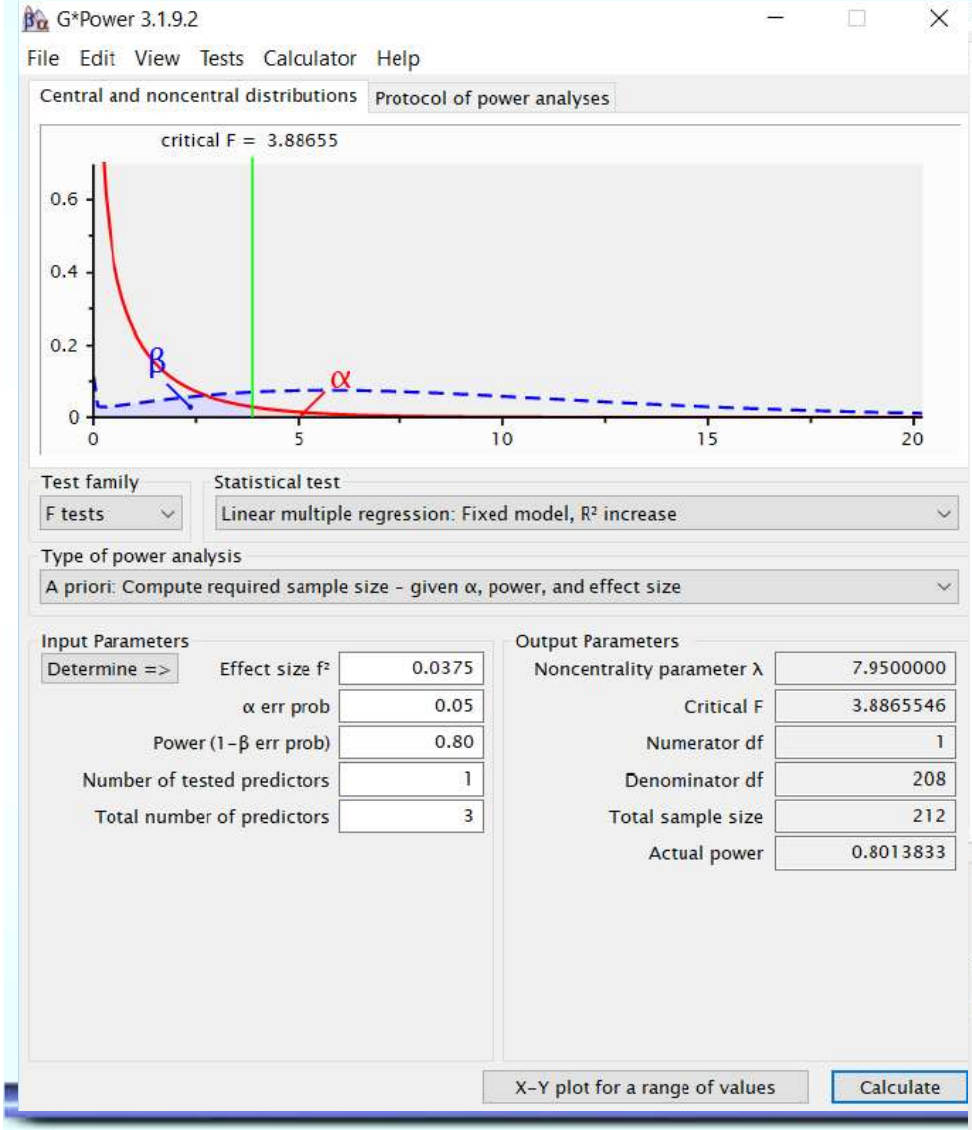
Option 2 Input the expected value of correlations between the dependent variable Y and the predictor X when the moderator M is equal to its mean (referred to as r_{yx} or equivalently as β_{Mean}), the correlation between the dependent variable and the moderator r_{ym}, and the expected correlation between the dependent variable Y and the predictor X when the moderator is 1SD above (or below) its mean, referred to as r_{yx} ±SD or β_{Mean} ±SD

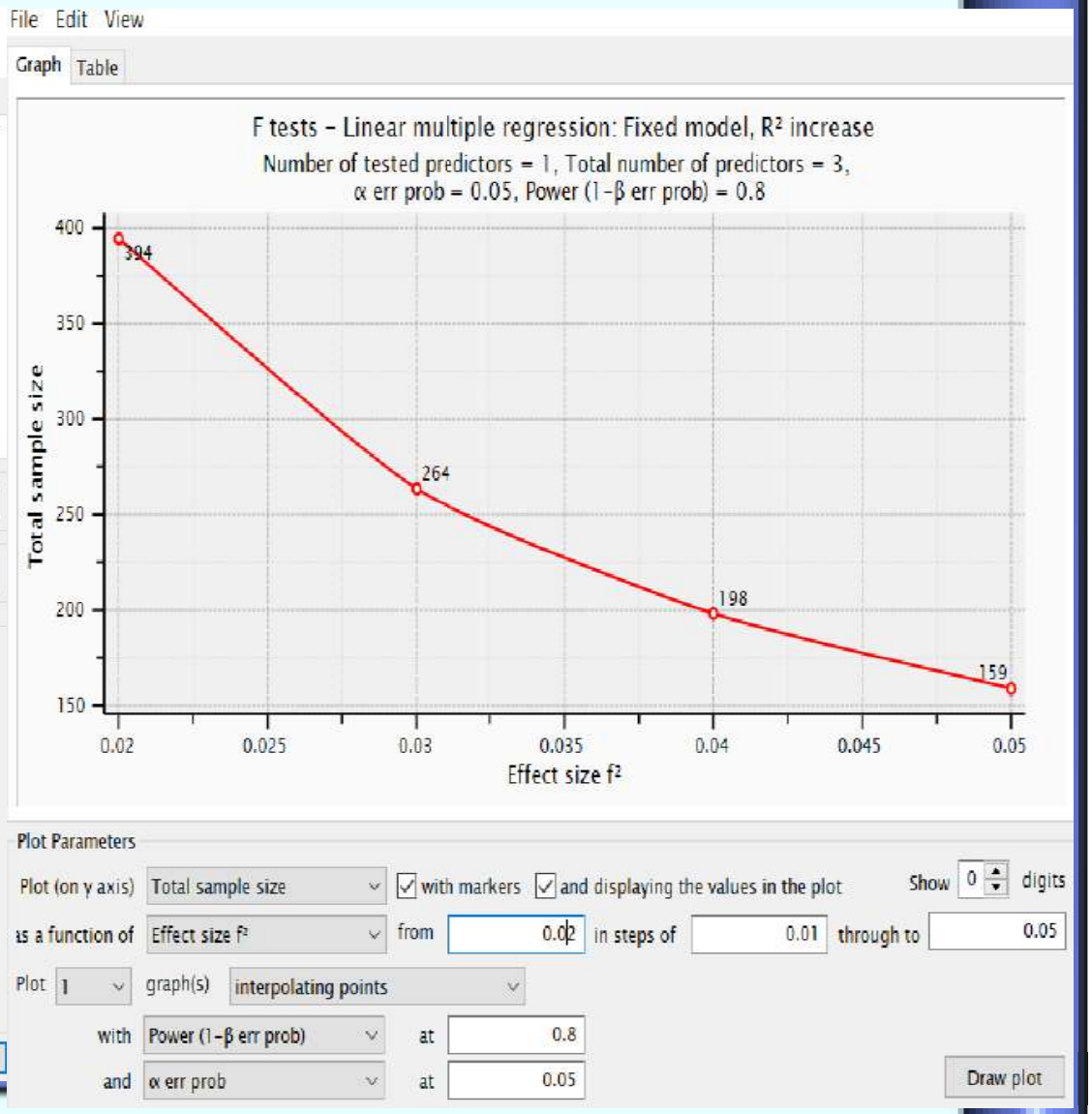
r_{yx} or β_{Mean}	0,35
r_{ym}	0,25
r_{yx} ±SD or β_{Mean} ±SD	0,525
β_{int}	0,175
f_n	0,037577



Routine Excel at https://github.com/mcfanda/primerPowerIRSP

Plugging it in G*Power





Mediation in regression analysis

Not all statistical models have available easy or generally valid analytic solutions for power

For mediation analysis, G*Power is not of help

Simulation approach can be a solution (**see Giulio tomorrow**)

R-packages are available

powerMediation (Qiu, 2017)

Bmem (Zhang, 2014)

Also Shiny app based on R

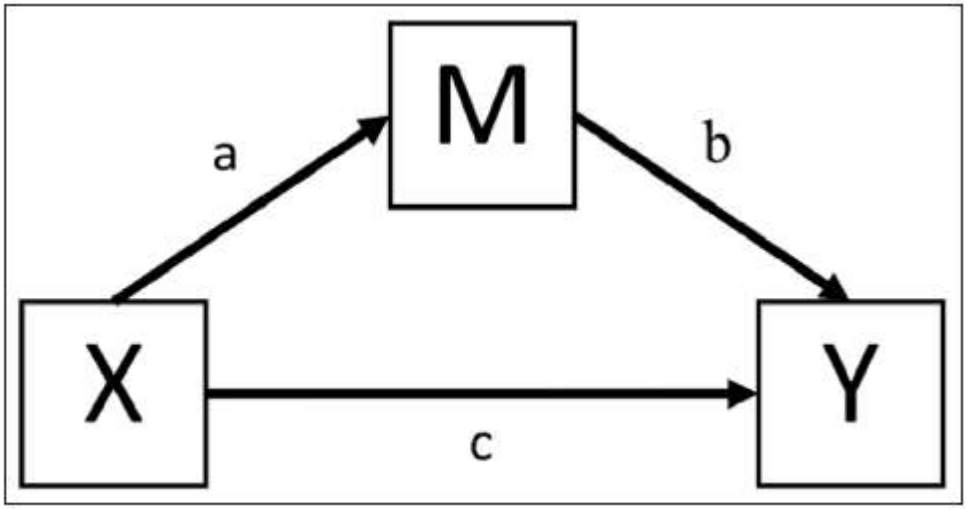
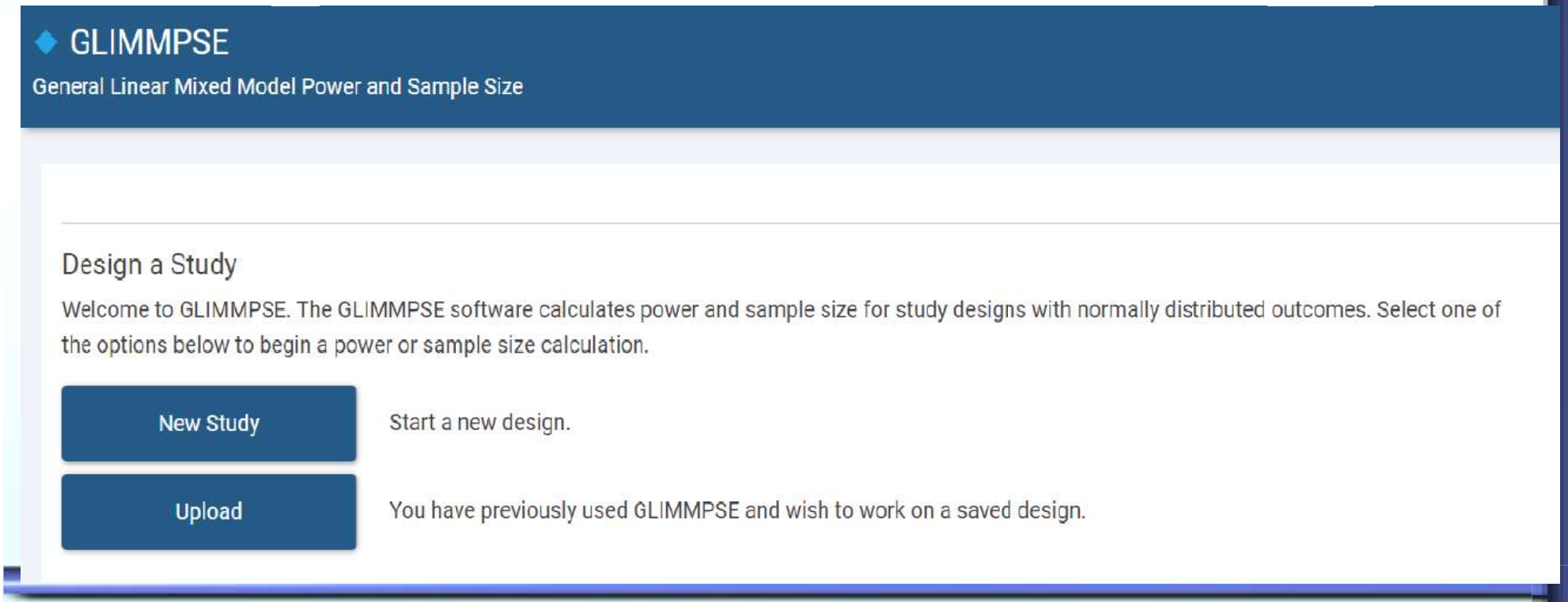


Figure 9: Mediation model.

ANOVA Within and Mixed

Web app: GLIMMPSE (https://glimmpse.samplesizeshop.org) but check also https://samplesizeshop.org/

2 x 2 Mixed ANOVA



Solve for sample size

GLIMMPSE

General Linear Mixed Model Power and Sample Size

2 x 2 Mixed ANOVA: Study title

Please pick a concise title for the study:

2 x 2 Mixed ANOVA

Solve for

Please indicate whether you would like to solve for power or total sample size.
If you have a rough idea of the number of research participants you will be able to recruit, then solve for power.
If you have few restrictions on recruitment then you may wish to solve for sample size.

Power | Sample Size

: : Target power

Progress Help Sa

Please choose one or more power values, for which you wish to calculate minimum sample size.

All target power values must be between 0 and 1, exclusive.

Target Power	remove
0.8	
0.9	

Define test and alpha

es : Statistical tests

Progress Help Sa

Please choose one or more statistical tests. If you are unsure which to pick, we recommend the Hotelling Lawley Trace test due to its equivalence to a mixed model test.

- [x] Hotelling Lawley Trace
- [ ] Pillai-Bartlett Trace
- [ ] Wilks Likelihood Ratio
- [ ] Box Corrected
- [ ] Geisser-Greenhouse Corrected
- [ ] Huynh-Feldt Corrected
- [ ] Uncorrected

: Type I error rates

Progress Help Sa

A Type I error occurs when a scientist declares a difference when none is present in the population. The Type I error rate is the probability of that kind of error, a false positive, and is often referred to as α (alpha). A Type I error rate can range from 0 to 1. Although the most commonly used value is 0.05, we recommend 0.01.

Type I Error Rate	remove
0.05	

Define outcomes and Within factor

: Outcomes

Progress Help Sav

Enter the name of each outcome variable one at a time in the underlined space below. For example, in a study investigating cholesterol-lowering medication, the outcome variables could be HDL, LDL, and total cholesterol.

Note that repeated measurement information will be addressed on the next screen.

Please name the one or more outcomes.

Outcome	remove
Performance	

: Repeated measures

Progress Help Sa

GLIMMPSE allows you to define within-participant factors, specified as repeated measures. An independent sampling unit provides one or more observations such that observations from one unit are statistically independent from any other distinct unit while observations from the same unit may be correlated. Repeated measures are present when a response variable is measured on each independent sampling unit on two or more occasions or under two or more conditions. The values of the repeated measures (that is, the levels of the within-participant factors) distinguish the occasions or conditions.

If the study includes repeated measures, click "Add Repeated Measure" and follow the prompts.

You may specify up to 5 repeated measures. Each repeated measure you add will apply to each outcome you specified on the previous page.

Define Repeated Measure

Define Within factor

What is the name of the dimension you will be measuring?

The text entered in the "Dimension" text box indicates the dimension over which measures were taken (e.g. time, days, locations, etc.).
The choice of "Type" indicates whether the repeated measures are numeric (e.g. time), or categorical (e.g. arm, leg, hand)

Dimension: Repeated measures

time

Cancel | Next: Type

What type of data is time?

Categorical | Numeric

Cancel | Back | Next: No. Measurements

Number of measurements of time?

2

You must have between 2 and 10 repeats (inclusive)

Cancel | Back | Next: Spacing

Spacing

If the repeated measures are numeric, the spacing values must be unique nonnegative integers, in ascending order.

Set values myself | Select values by series

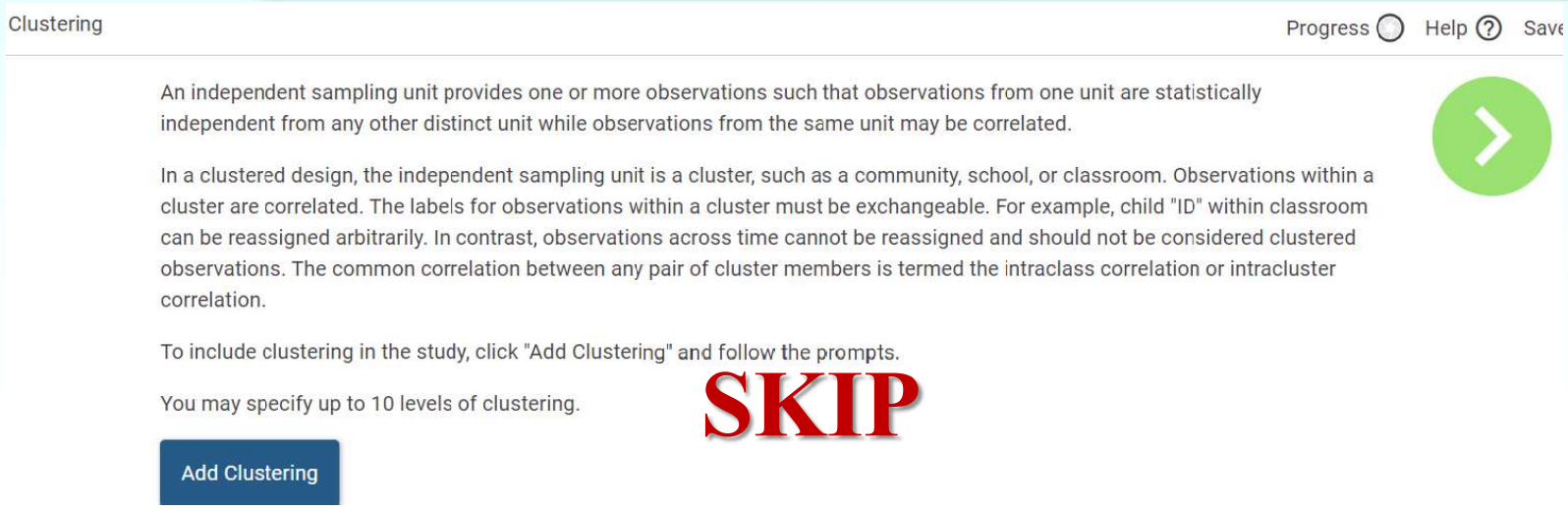
Measurement #1 at 1

Measurement #2 at 2

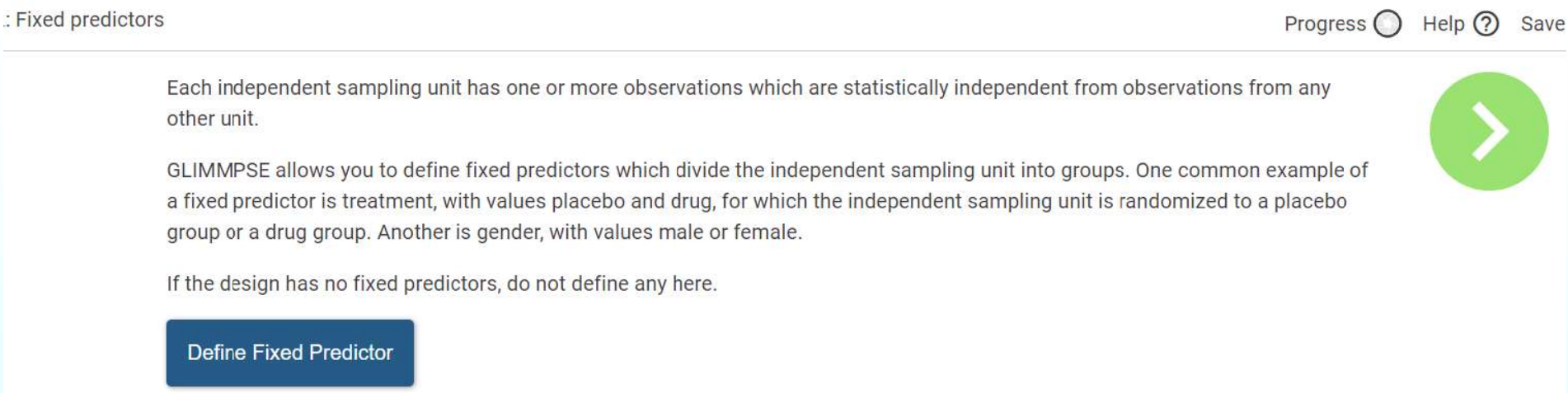
Cancel | Back | Update repeated measure

Repeated Measure Dimension	Type	Measurements	Edit	Remove
time	Numeric	["1", "2"]		

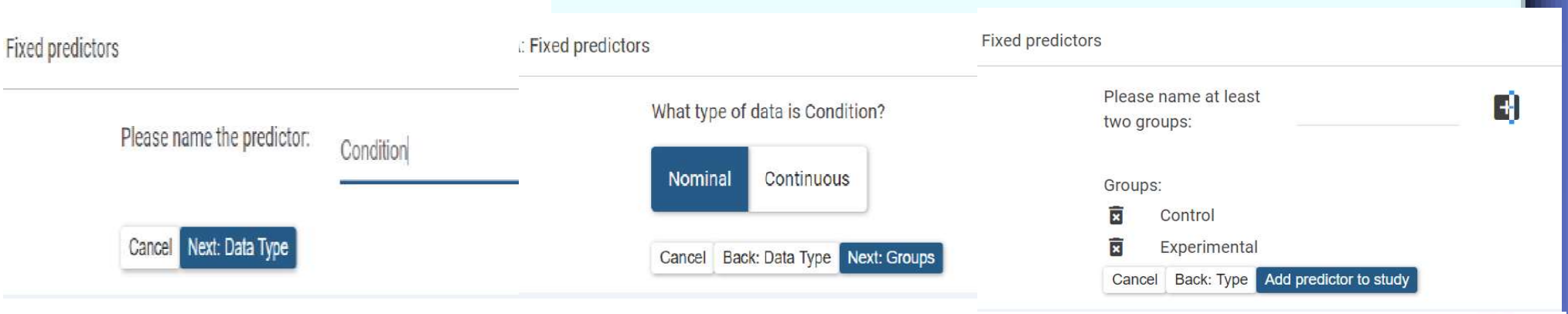
Define Between factor



SKIP



Define Between factor



Each independent sampling unit has one or more observations which are statistically independent from observations from any other unit.

GLIMMPSE allows you to define fixed predictors which divide the independent sampling unit into groups. One common example of a fixed predictor is treatment, with values placebo and drug, for which the independent sampling unit is randomized to a placebo group or a drug group. Another is gender, with values male or female.

If the design has no fixed predictors, do not define any here.

Define Fixed Predictor

Fixed Predictors

Name	Type	Units	Groups	Remove	Edit
Condition	NOMINAL		["Control", "Experimental"]		

Select key hypothesis for power analysis

Hypothesis choice

Progress Help Save

Each power or sample size calculation is based on selecting a specific study hypothesis. The options below show the hypotheses which are available for the current study design. Specify the hypothesis that represents your scientific question.

GLIMMPSE chooses sensible contrast matrices based on cell means coding. Should you wish to define your own contrast matrices, pick the highest order interaction and choose from the advanced options in the hypothesis components.

Select a hypothesis from the list.

	Effects Available for Consideration	Nature of Variation
◉	Condition x time: Interaction	Between x Within
○	time: Main Effect	Within
○	Condition: Main Effect	Between
○	Grand Mean	Between

Test hypothesis

: Hypothesis

Progress

What type of contrast do you wish among the means defined by your groups and repeated measures?

All mean differences zero

A parameter is a characteristic of a population.The parameters of interest are differences between groups at individual repeated measures.

The null hypothesis is that all pairwise differences between groups are the same among all pairs of repeated measures.

Show Advanced Options

Theta 0

Progress Help Sa

A hypothesis compares parameters to a constant, the contrast comparison constant, Θ_0. This is almost always zero. If you choose a value other than zero, be sure that you understand that the hypothesis you define is scientifically meaningful. Also note that the description and interpretation of your hypothesis given when choosing your contrasts will be affected.

0

Group size ratios

Progress

For equal group sizes, input a "1" in the block next to each group. This is the default study design.

For unequal group sizes, specify the ratio of the group sizes. For example, consider a design with an active drug group and a placebo group. If twice as many study participants receive the placebo, a value of "2" would be selected for the placebo group, and a value of "1" would be selected for the active drug group.

Group size ratios

Condition	Control	1
	Experimental	1

Expected means under key hypothesis

Marginal means

Progress (

The table below shows the mean values for outcome **Performance** within each group in the study. Each group is represented by a row in the table, and each repeated measure dimension is represented by a column.

Enter the mean values you expect to observe for outcome **Performance** within each group. The table should contain at least one value that is non-zero. Also, at least two groups should have means which differ by a scientifically meaningful amount.

Expected mean values, per group, for *Performance*

		time	
		1	**2**
Condition	**Control**	5	5
	Experimental	5	6

Set blank values to value

Scale factors (different scenarios) and SD

Scale factor for the marginal means

Progress

In power analysis, it is not possible to know the exact values of means before the experiment is observed. Scale factors allow you to consider alternative values for the means by scaling the values entered on the previous screen.
For example, entering the scale factors 0.5, 1, and 2 would compute power for the mean values divided by 2, the mean values as entered, and the mean values multiplied by 2.

Enter a scale factor:

number > 0 +

Scale Factor	remove
1	

Variability across outcomes

Enter the standard deviation you expect to observe for each outcome.

Outcome	Standard Deviation
Performance	1

Repeated measure standard deviation ratios

Progress

Define the ratios of standard deviations for time. One of your values should be 1 and the others should represent the ratio of that value to that value:

For example, if you believe that the standard deviation doubles at each time, enter the values 1, 2, 4, 8... etc.

time	Standard Deviation Ratio
1	1
2	1

Repeated measures correlations and scale factors

Repeated measure correlation

Progress

For a given research participant, responses vary across outcomes and across repeated measurements. The amount of variability can dramatically impact power and sample size.

Define the **time** correlation matrix, by entering correlations you expect to observe among the chosen spacing values of **time**:

Unstructured | LEAR

time

	1	2
1	1	0,5
2	0,5	1

(each off-diagonal correlation must be between -1 and 1, exclusive)

: Scale factor variance

Progress

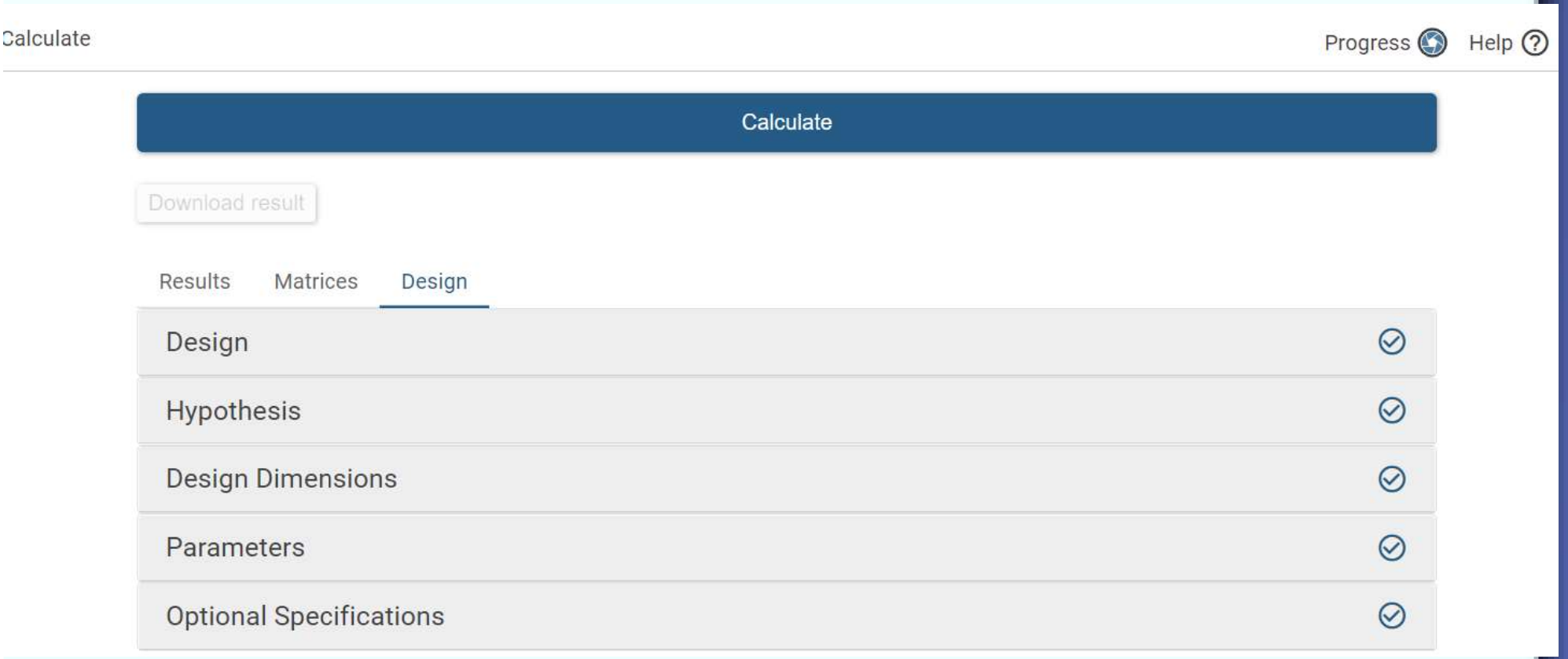
Changes in variability can dramatically affect power and sample size results. It is not possible to know the variability until the experiment is observed. Scale factors allow you to consider alternative values for variability by scaling the calculated covariance matrix. For example, entering the scale factors 0.5, 1, and 2 would compute power for the covariance matrix divided by 2, the covariance matrix as entered, and the covariance matrix multiplied by 2.

You may add up to 10 scale factors.

Choose a number greater than zero

Scale Factor	remove
1	

Finally, the calculation…



…and the results!

Progress Help Save

Calculate

Download result

Results | Matrices | Design

Power	Total Sample Size	Target Power	Means Scale Factor	Variability Scale Factor	Test	Power Method	Type I Error Rate
0.807	34	0.8	1	1	Hotelling Lawley Trace	conditional	0.05
0.912	46	0.9	1	1	Hotelling Lawley Trace	conditional	0.05

Suppose expected correlation is lower

time

	1	2
1	1	0,25
2	0,25	1

(each off-diagonal correlation must

Calculate

Download result

Results | Matrices | Design

Power	Total Sample Size	Target Power	Means Scale Factor	Variability Scale Factor	Test	Power Method	Type I Error Rate
0.807	50	0.8	1	1	Hotelling Lawley Trace	conditional	0.05
0.904	66	0.9	1	1	Hotelling Lawley Trace	conditional	0.05

Suppose no correlation

time

	1	2
1	1	0
2	0	1

(each off diagonal correlation p

Calculate

Download result

Results | Matrices | Design

Power	Total Sample Size	Target Power	Means Scale Factor	Variability Scale Factor	Test	Power Method	Type I Error Rate
0.808	66	0.8	1	1	Hotelling Lawley Trace	conditional	0.05
0.906	88	0.9	1	1	Hotelling Lawley Trace	conditional	0.05

Recap Examples Mixed ANOVA

1) The design was a 2 x 2 Mixed ANOVA
2) We varied the expected correlations

Required N for power at .80

- **r=.00, N= 66**
- **r=.25, N= 50**
- **r=.50, N= 34**

Required N goes down as the correlation between DVs of the Within factor goes up

Suppose instead a 2 x 2 Between Ss

The design you've described, means that every level of **Group** occurs at every level of **Condition**. This concept applies to every combination of fixed predictors.

Define Fixed Predictor

Fixed Predictors

Name	Type	Units	Groups	Remove	Edit
Condition	NOMINAL		["Control", "Experimental"]		
Group	NOMINAL		["No previous experience", "Previous experience"]		

	Effects Available for Consideration	Nature of Variation
◉	Condition x Group: Interaction	Between x Between

Expected mean values, per group, for *Performance*

Condition, Group		
	Control, No previous experience	5
	Control, Previous experience	5
	Experimental, No previous experience	5
	Experimental, Previous experience	6

Calculate

Download result

Results | Matrices | Design

Power	Total Sample Size	Target Power	Means Scale Factor	Variability Scale Factor	Test	Power Method	Type I Error Rate
0.801	128	0.8	1	1	Hotelling Lawley Trace	conditional	0.05
0.903	172	0.9	1	1	Hotelling Lawley Trace	conditional	0.05

Suppose instead a 2 x 2 Within Ss (r=.25)

The design you have described, means that every level of **Variable B** is measured at every level of **Variable A**. This concept applies to every combination of repeated measures.

Define Repeated Measure

Repeated Measure Dimension	Type	Measurements	Edit	Remove
Variable A	Categorical	["1", "2"]		
Variable B	Categorical	["1", "2"]		

Effects Available for Consideration	Nature of Variation
Variable A x Variable B: Interaction	Within x Within

Expected mean values, per group, for *Performance*

Variable A, Variable B

1,1	1,2	2,1	2,2
5	5	5	6

Variable A

	1	2
1	1	0,25
2	0,25	1

Variable B

	1	2
1	1	0,25
2	0,25	1

Calculate

Download result

Results | Matrices | Design

Power	Total Sample Size	Target Power	Means Scale Factor	Variability Scale Factor	Test	Power Method	Type I Error Rate
0.807	20	0.8	1	1	Hotelling Lawley Trace	conditional	0.05
0.904	26	0.9	1	1	Hotelling Lawley Trace	conditional	0.05

2 x 2 Within Ss with r=.50 and r=.0

Variable A

	1	2
1	1	0,5
2	0,5	1

Variable B

	1	2
1	1	0,5
2	0,5	1

Calculate

Download result

Results | Matrices | Design

Power	Total Sample Size	Target Power	Means Scale Factor	Variability Scale Factor	Test	Power Method	Type I Error Rate
0.803	10	0.8	1	1	Hotelling Lawley Trace	conditional	0.05
0.911	13	0.9	1	1	Hotelling Lawley Trace	conditional	0.05

Variable A

	1	2
1	1	0
2	0	1

Variable B

	1	2
1	1	0
2	0	1

Results | Matrices | Design

Power	Total Sample Size	Target Power	Means Scale Factor	Variability Scale Factor	Test	Power Method	Type I Error Rate
0.808	34	0.8	1	1	Hotelling Lawley Trace	conditional	0.05
0.900	44	0.9	1	1	Hotelling Lawley Trace	conditional	0.05

Power Comparison

- Three 2 x 2 ANOVA designs (Mixed, Between, Within)
- In each design the same pattern of expected means

	A1	A2
B1	5	5
B2	5	6

- Always SD=1
- Always powered for interaction effect
- Required N for power at .80

  -Between = 128
  -Mixed (r=.00) = 64
  -Mixed (r=.25) = 50
  -Mixed (r=.50) = 34
  -Within (r=.00) = 34
  -Within (r=.25) = 20
  -Within (r=.50) = 10

- You can draw your own conclusion…

How to increase power?

- **<u>Increase sample size</u> (also multi-lab collaborations)**
- Use blocking or repeated measures (**<u>within Ss design</u>**) BUT sometimes can be inappropriate
- Administer stronger treatments (e.g., experimental manipulation) BUT be wary of possible reduced ecological validity
- Avoid restrictions of range for dependent variables
- Standardize experimental procedures
- Increase reliability of measures
- Use more homogenous subject samples BUT increased risks to generalizability of results
- Meta-analytic mindset

Increasing power without increasing sample size I

$$SE = \sqrt{\frac{S^2}{n}}$$

Increasing Statistical Power Without Increasing Sample Size

Gary H. McClelland
University of Colorado at Boulder
August 2000 • American Psychologist 963

Increasing the Power of Your Study by Increasing the Effect Size

TOM MEYVIS
STIJN M. J. VAN OSSELAER
Journal of Consumer Research. Feb2018, Vol. 44 Issue 5, p1157-1173.

- Standard errors depend on N and SD (smaller SD means smaller SE)
- SD can be reduced with more reliable measures, more precise experimental designs, less Ss variability (e.g., also within Ss designs)
- Plan your design as simple and as clean as possible

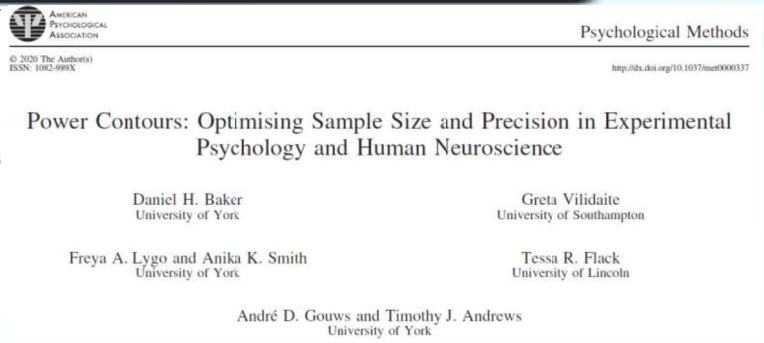
Increasing power without increasing sample size II

Power, Dominance, and Constraint: A Note on the Appeal of Different Design Traditions

Jeffrey N. Rouder[1,2] and Julia M. Haaf[2]
[1]Department of Cognitive Sciences, University of California, Irvine, and [2]Department of Psychological Sciences, University of Missouri

Advances in Methods and Practices in Psychological Science
2018, Vol. 1(1) 19–26

DOI: 10.1177/2515245917745058
www.psychologicalscience.org/AMPPS
SAGE



American Psychological Association
Psychological Methods
© 2020 The Author(s)
ISSN: 1082-989X
http://dx.doi.org/10.1037/met0000337

Power Contours: Optimising Sample Size and Precision in Experimental Psychology and Human Neuroscience

Daniel H. Baker, University of York
Greta Vilidaite, University of Southampton
Freya A. Lygo and Anika K. Smith, University of York
Tessa R. Flack, University of Lincoln
André D. Gouws and Timothy J. Andrews, University of York

- The SD (σ) depends on both N (variance between Ss) and K (trials, variance across trials within Ss)
- This means that N can be (to some extent) traded for K

$$\lambda = \sqrt{\mu^2 \times \frac{IK}{K\sigma_\beta^2 + 2\sigma^2}}.$$

$$\sigma_s = \sqrt{\sigma_b^2 + \frac{\sigma_w^2}{k}}$$

$$\sigma_b = \sqrt{\sigma_s^2 - \frac{\sigma_w^2}{k}},$$

Increasing power without increasing sample size II

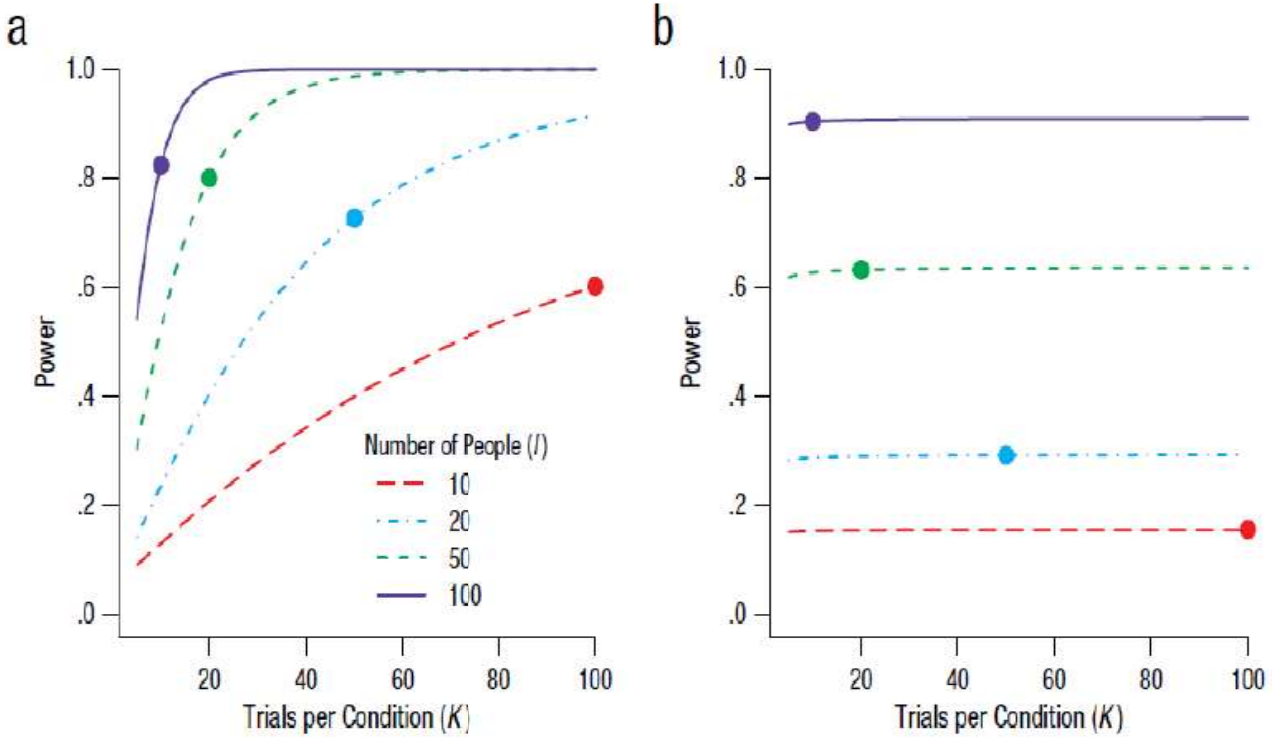
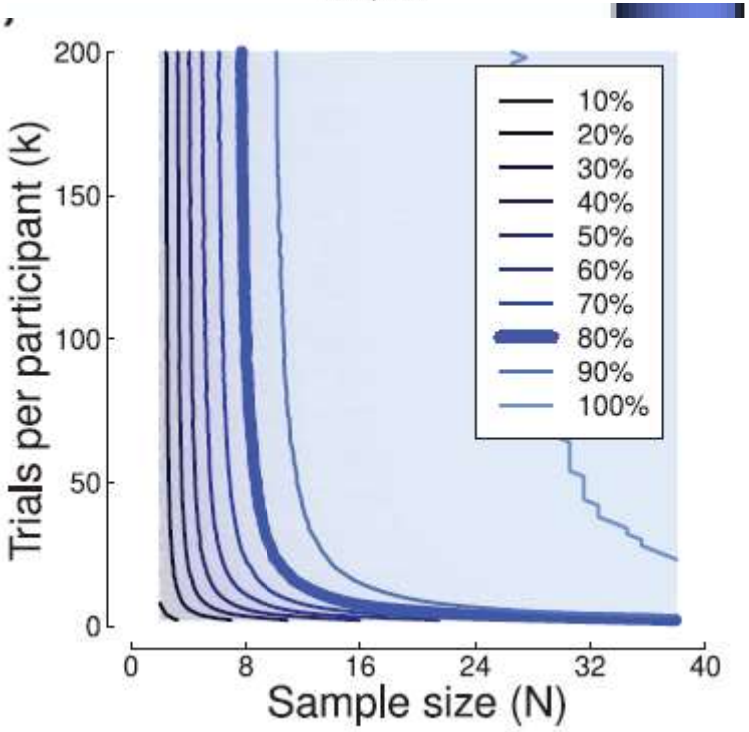


Fig. 1. Illustration of how power changes as a function of *I*, the number of participants, and *K*, the number of observations per condition per participant. The graph in (a) shows the trade-off between these two design parameters when population variability (σ_β = 28 ms) is small relative to trial variability (σ = 300 ms). The graph in (b) shows the trade-off when population variability (σ_β = 3) is large relative to trial variability (σ = 1). The plotted point on each curve indicates the values of *I* and *K* that provide 1,000 observations.

American Psychological Association
Psychological Methods
© 2020 The Author(s)
ISSN: 1082-989X
http://dx.doi.org/10.1037/met0000337

Power Contours: Optimising Sample Size and Precision in Experimental Psychology and Human Neuroscience

Daniel H. Baker, University of York
Greta Vilidaite, University of Southampton
Freya A. Lygo and Anika K. Smith, University of York
Tessa R. Flack, University of Lincoln
André D. Gouws and Timothy J. Andrews, University of York



Power, Dominance, and Constraint: A Note on the Appeal of Different Design Traditions

Advances in Methods and Practices in Psychological Science
2018, Vol. 1(1) 19–26
© The Author(s) 2018
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/2515245917745058
www.psychologicalscience.org/AMPPS
SAGE

Jeffrey N. Rouder[1,2] and Julia M. Haaf[2]
[1]Department of Cognitive Sciences, University of California, Irvine, and [2]Department of Psychological Sciences, University of Missouri

The signal and the noise

$$SE = \sqrt{\frac{S^2}{n}}$$

Distinguish conceptually between unnecessary ("added noise") and necessary ("natural") variance

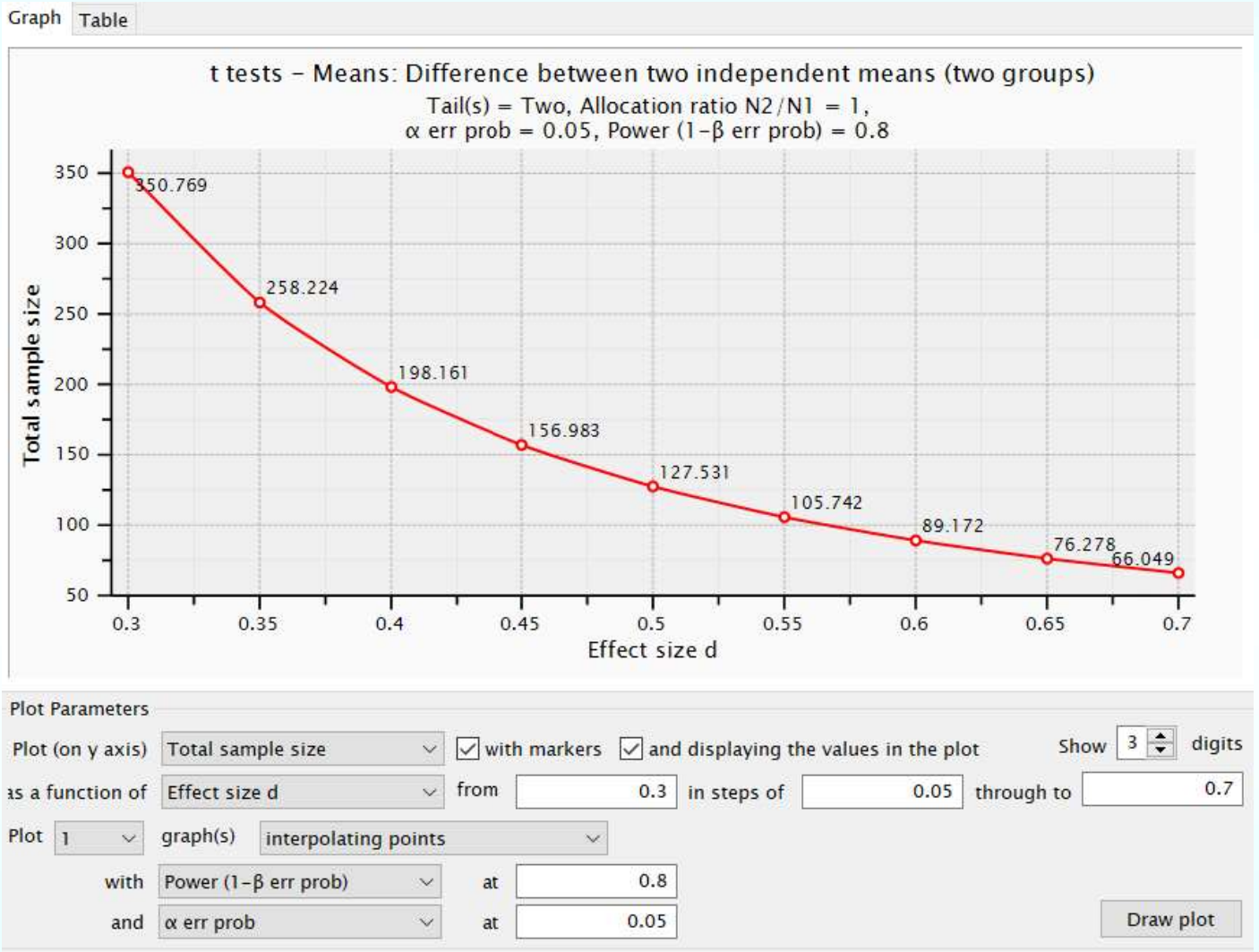
Improve your design. Optimize it. Think carefully about it. Few extra hours spent on this can be worth hundreds of extra participants (and avoid frustrations…)

Reduce the **noise**! Increase the **signal**!

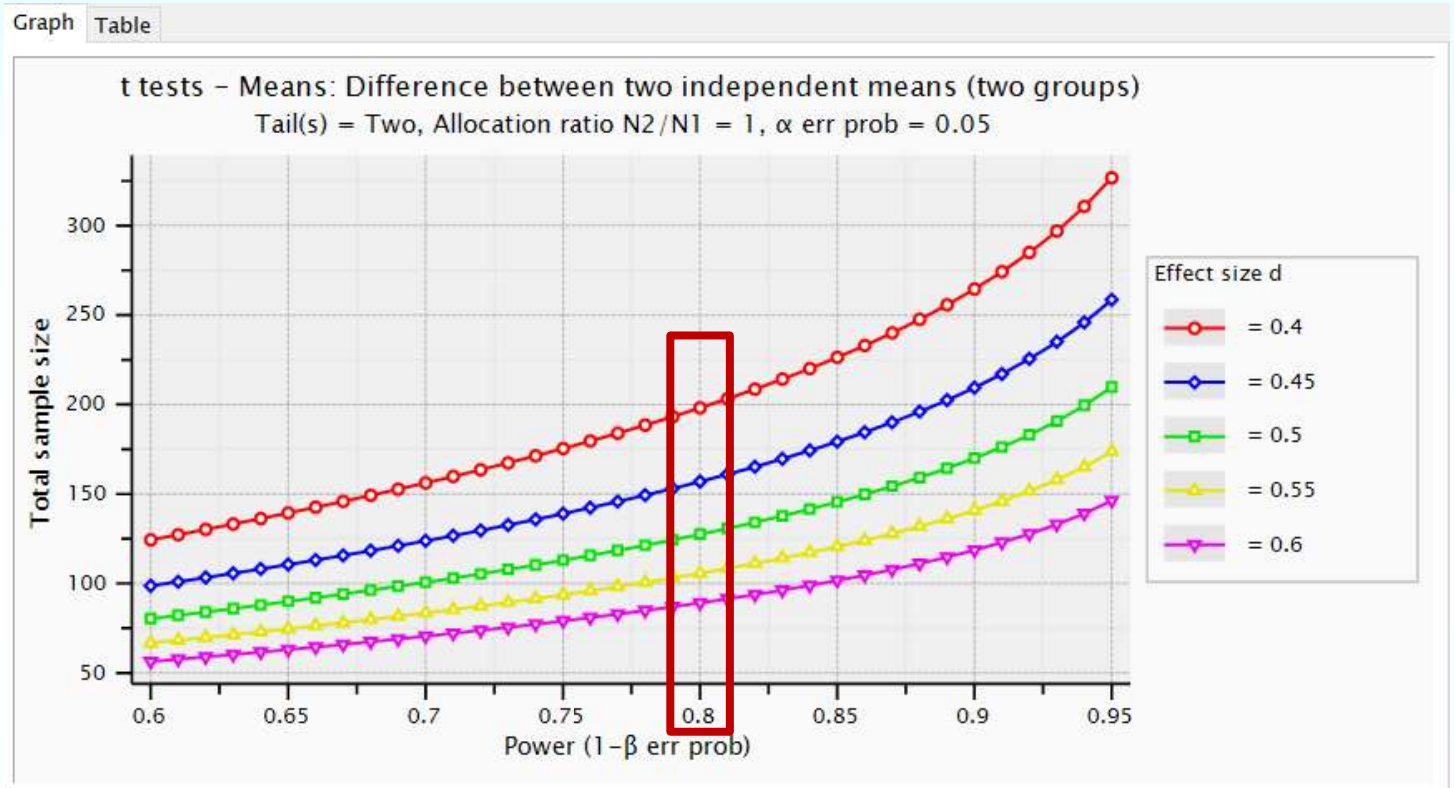
Problems in power analysis

- **Main error**: post-hoc power (calculated after the results) is trivial and misleading. **Sensitivity analysis** is much better
- **Main problem**:
  - One key element of power analysis for planning studies is the Effect Size (ES)
  - We cannot know the ES. If we knew it, we wouldn't need to run the study…
  - At best we can guess/estimate ES from a meta-analysis or from previous studies, often based on a hunch. **Uncertainty** of the estimate.
  - What happens if the ES estimate is incorrect?

Uncertainty of ES



Uncertainty of ES



Uncertainty of ES

Graph | Table

t tests - Means: Difference between two independent means (two groups)
Tail(s) = Two, Allocation ratio N2/N1 = 1, α err prob = 0.05

		Effect size d = 0.4	Effect size d = 0.45	Effect size d = 0.5	Effect size d = 0.55	Effect size d = 0.6
#	Power (1-β err prob)	Total sample size	Total sample size	Total sample size	Total sample size	Total sample size
16	0.750000	175.449	139.039	112.997	93.7315	79.0806
17	0.760000	179.664	142.369	115.695	95.9607	80.9535
18	0.770000	184.029	145.818	118.488	98.2689	82.8927
19	0.780000	188.556	149.395	121.385	100.663	84.9042
20	0.790000	193.261	153.112	124.396	103.151	86.9946
21	0.800000	198.161	156.983	127.531	105.742	89.1716
22	0.810000	203.275	161.024	130.804	108.446	91.4438
23	0.820000	208.626	165.252	134.228	111.276	93.8216
24	0.830000	214.241	169.689	137.822	114.246	96.3167
25	0.840000	220.153	174.359	141.605	117.372	98.9433
26	0.850000	226.397	179.293	145.601	120.675	101.718

Plot Parameters

Plot (on y axis) [Total sample size] ☑ with markers ☐ and displaying the values in the plot

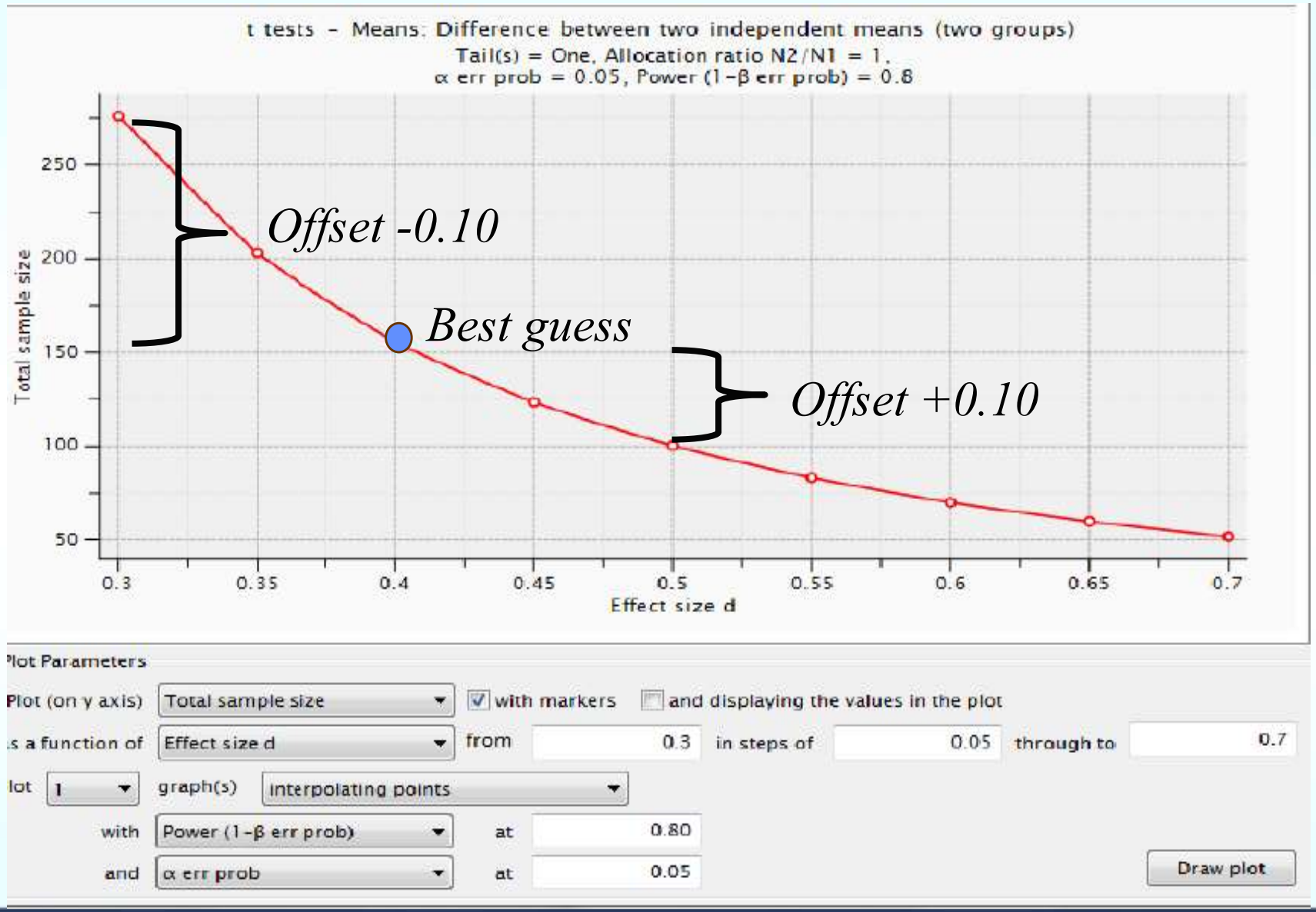
as a function of [Power (1-β err prob)] from [0.6] in steps of [0.01] through to [0.95]

Plot [5] graph(s) [interpolating points]

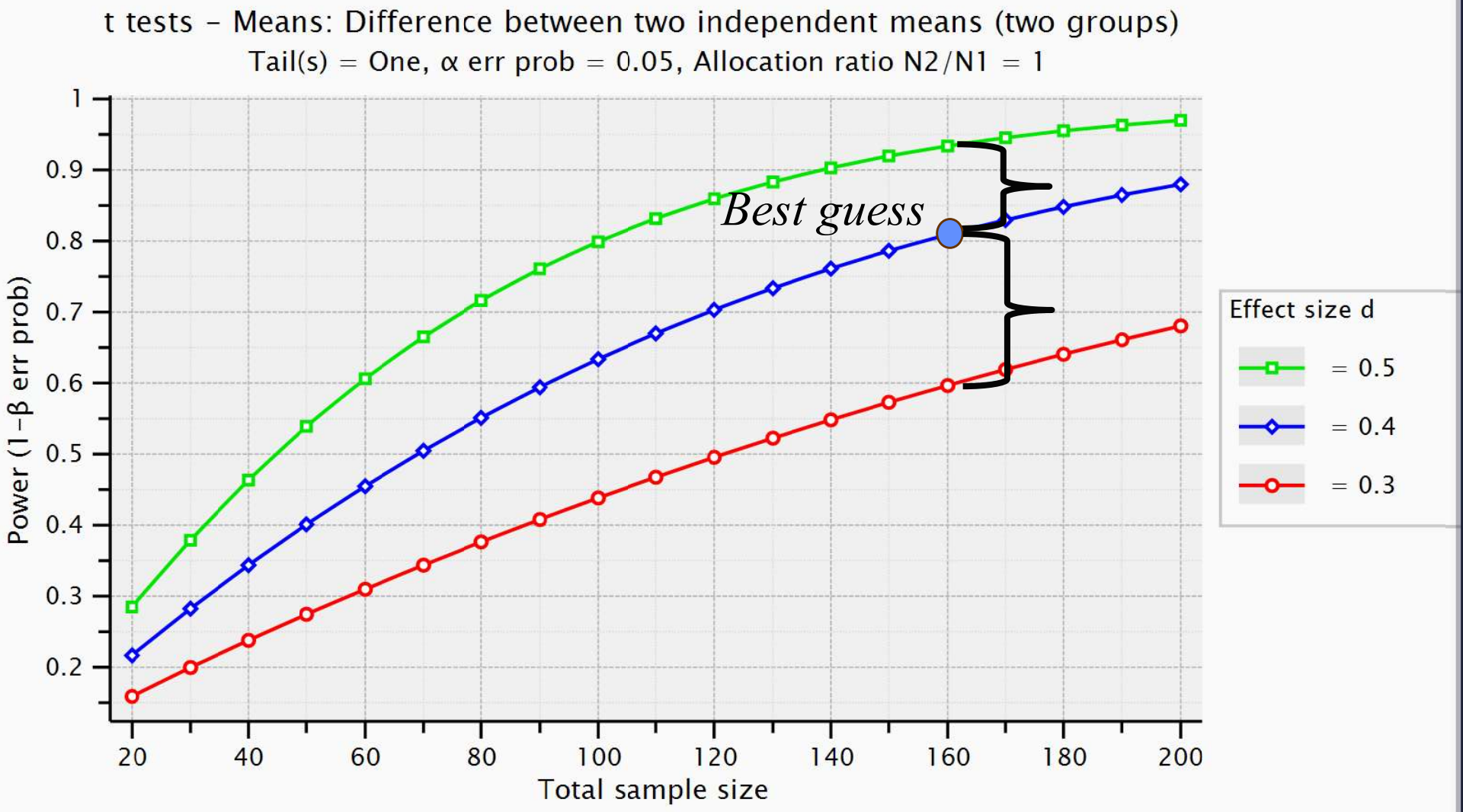
with [Effect size d] from [0.4] in steps of [0.05]

and [α err prob] at [0.05] Draw plot

Asymmetry of ES errors

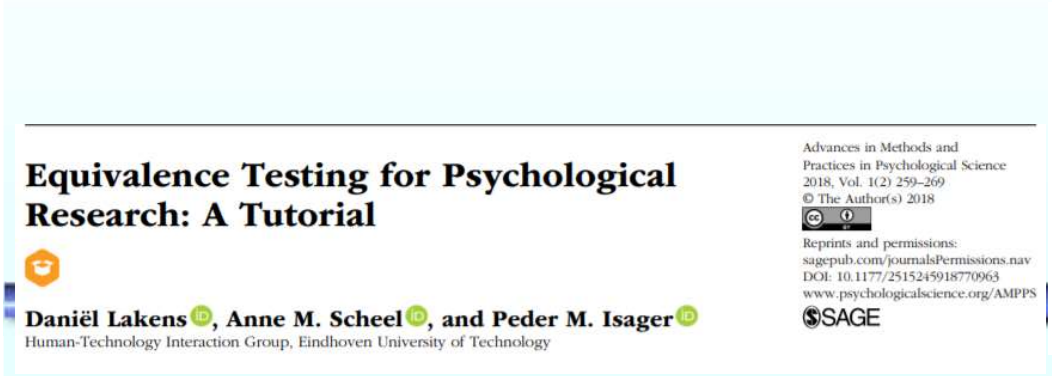


Asymmetry of ES errors



What to do then?

- Power depends on estimated ES (we don't know the "true" ES)
- ES over-estimation is more common (*optimistic bias*) and more influential than under-estimation (*asymmetric effect*)
- Should consider different scenarios rather than a single value
- Could consider minimum effect of interest (SESOI, Lakens, 2014)
- Could consider sensitivity analysis
- Could consider safeguarding yourself against "optimistic" ES estimates



Equivalence Testing for Psychological Research: A Tutorial

Daniël Lakens, Anne M. Scheel, and Peder M. Isager
Human-Technology Interaction Group, Eindhoven University of Technology

Advances in Methods and Practices in Psychological Science
2018, Vol. 1(2) 259–269

DOI: 10.1177/2515245918770963
www.psychologicalscience.org/AMPPS
SAGE

Safeguard Power as a Protection Against Imprecise Power Estimates

Marco Perugini, Marcello Gallucci, and Giulio Costantini
University of Milan-Bicocca, Italy

Perspectives on Psychological Science
2014, Vol. 9(3) 319–332

DOI: 10.1177/1745691614528519
pps.sagepub.com
SAGE

Summing up Power Analysis I

- Power analysis is one important way to efficiently plan a study
- Try to power your study adequately
- A main problem is to best guess a predicted ES
- Beware of the uncertainty of ES estimates and the asymmetric impact of ES estimate errors
- Wise to consider uncertainty in the ES estimate (e.g., by running different scenarios)
- Think in terms of range of values rather than a specific value

What does it mean to have enough power?

The Crest-tailed Mulgara is a species of marsupial that was recently rediscovered living in an area where it had been presumed extinct for about 100 years (Credit: Reece Pedler)

Power as fuel in the tank

- Have enough fuel to find what you are looking for (hoping that it is there) in a place at a distance that you hope have guessed reasonably well

Summing up Power Analysis II

- Increasing power means to decrease inference errors in general (direct effect on false negatives and indirect effect on false positives)
- More complex designs (e.g., multi-level) are challenging
- Sometimes no algorithmic solutions are available
- A simulation approach can be a valid solution (see Giulio's seminar tomorrow)

Some readings for some advanced issues

Contrast, regression, moderation, and mediation effects

- Perugini, M., Gallucci, M., & Costantini, G. (2018). A Practical Primer To Power Analysis for Simple Experimental Designs. *International Review of Social Psychology, 31(1)*.

Within and Mixed ANOVA

- Guo, Y., Logan, H. L., Glueck, D. H., & Muller, K. E. (2013). Selecting a sample size for studies with repeated measures. *BMC medical research methodology*, *13*(1), 100
- Rouder, J. N., & Haaf, J. M. (2018). Power, dominance, and constraint: A note on the appeal of different design traditions. *Advances in Methods and Practices in Psychological Science, 1*, 19–26.
- Web app: GLIMMPSE (https://glimmpse.samplesizeshop.org)

Mixed/Multilevel Models

- Judd, C. M., Westfall, J., & Kenny, D. A. (2016). Experiments with more than one random factor: Designs, analytic models, and statistical power. *Annual Review of Psychology.* Web app: https://jakewestfall.shinyapps.io/two_factor_power/ See also Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: a tutorial. *Journal of Cognition*, *1*(1).
- Kelcey, B., Xie, Y., Spybrook, J., & Dong, N. (2020). Power and sample size determination for multilevel mediation in three-level cluster-randomized trials. *Multivariate Behavioral Research*
  https://www.causalevaluation.org/power-analysis.html

Simulation based power analysis

- Gelman, A., Hill, J. (2006) *Data analysis using regression and multilevel/hierarchical models*. Cambridge: Cambridge University Press.

Advanced models and exemplary R code

- Liu, X. S. (2014). *Statistical Power Analysis for the Social and Behavioral Sciences: Basic and Advanced Techniques*. New York: Routledge.

UNIVERSITÀ DEGLI STUDI DI MILANO BICOCCA

(Some)

Tips for getting it right

Back to the problem

- As scientists, we all want to get something right
- If we get it right, it is replicable and will be replicated
- But what does it mean “to get it right”?
- So, what can we do to increase our chances?

Some pointers

1st pointer: Power

- Design your study with adequate power (***probability of finding an effect if it does exist***)

- Underpowered studies produce conflicting evidence and false negatives **but also false positives** (Maxwell, 2004; Ioannidis, 2005)

- Direct effect on **False Negatives** but also indirect effect on **False Positives**
  (False Discovery Rate /True False Positives)

Why many effects are not replicated?

- A mix of different factors and possible explanations
- Two main factors

  a) Low power and **b) Publication bias**

- Under these conditions, it is predictable that the literature will contain many false positives (results that seems significant but are not) and artificially boosted effect sizes
- Hence effects will be difficult to replicate

Power and Design

- Everything else being equal, power is affected by the design of your study
- **Within Ss** are more powerful than Between Ss designs

INTERNATIONAL REVIEW OF SOCIAL PSYCHOLOGY

Perugini, M., et al. (2018). A Practical Primer To Power Analysis for Simple Experimental Designs. *International Review of Social Psychology*, 31(1): 20, 1–23, DOI: https://doi.org/10.5334/irsp.181

RESEARCH ARTICLE

A Practical Primer To Power Analysis for Simple Experimental Designs

Marco Perugini, Marcello Gallucci and Giulio Costantini

journal of cognition

Brysbaert, M. 2019 How Many Participants Do We Have to Include in Properly Powered Experiments? A Tutorial of Power Analysis with Reference Tables. *Journal of Cognition*, 2(1): 16, pp. 1–38. DOI: https://doi.org/10.5334/joc.72

REVIEW ARTICLE

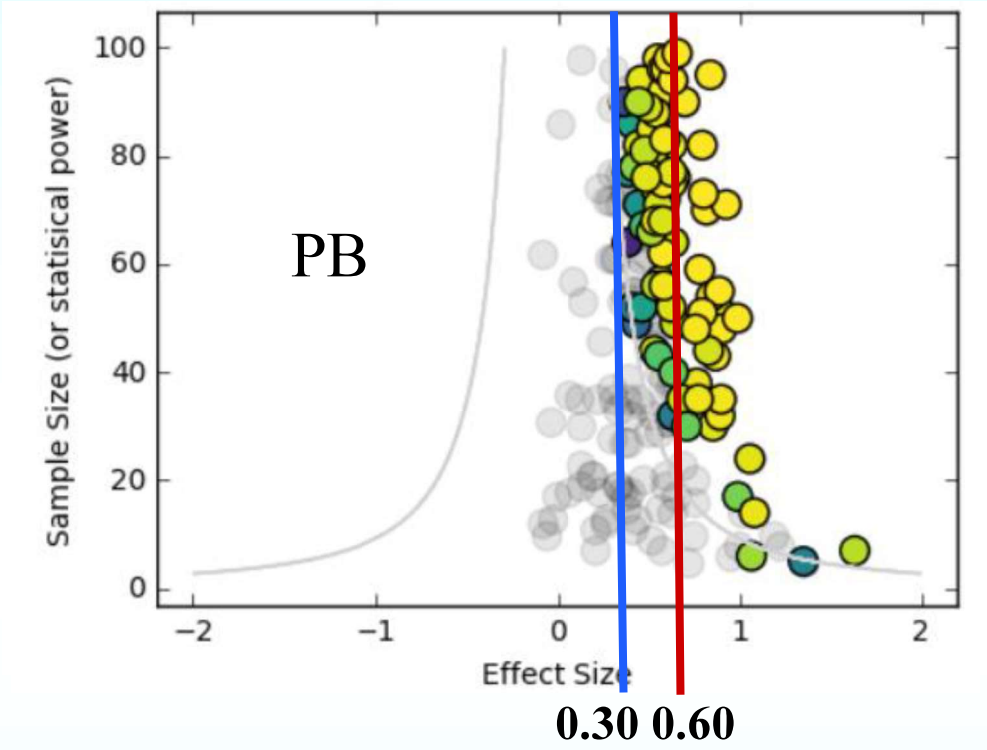
How Many Participants Do We Have to Include in Properly Powered Experiments? A Tutorial of Power Analysis with Reference Tables

Marc Brysbaert
Department of Experimental Psychology, Ghent University, BE
marc.brysbaert@ugent.be

1.b Publication bias

- Tendency to publish mainly significant results (and to submit for publication mainly studies with significant results)
- There are sometimes understandable reasons (unclear evidence, contradictory support, pilot studies, tentative paradigms, etc.)
- But often is a by-product of confirmation/positivity biases and insufficient culture of cumulative knowledge in a scientific field

Publication bias



The ES will be overestimated. How much depends on the extent of PB and on the prevalence of small samples.

A reader will think that Cohen's d=0.60 but in fact is d=0.30

Publication bias, Effect Sizes, underpowered studies

ES: Cohen's d=0.60 (vs. d= 0.30)

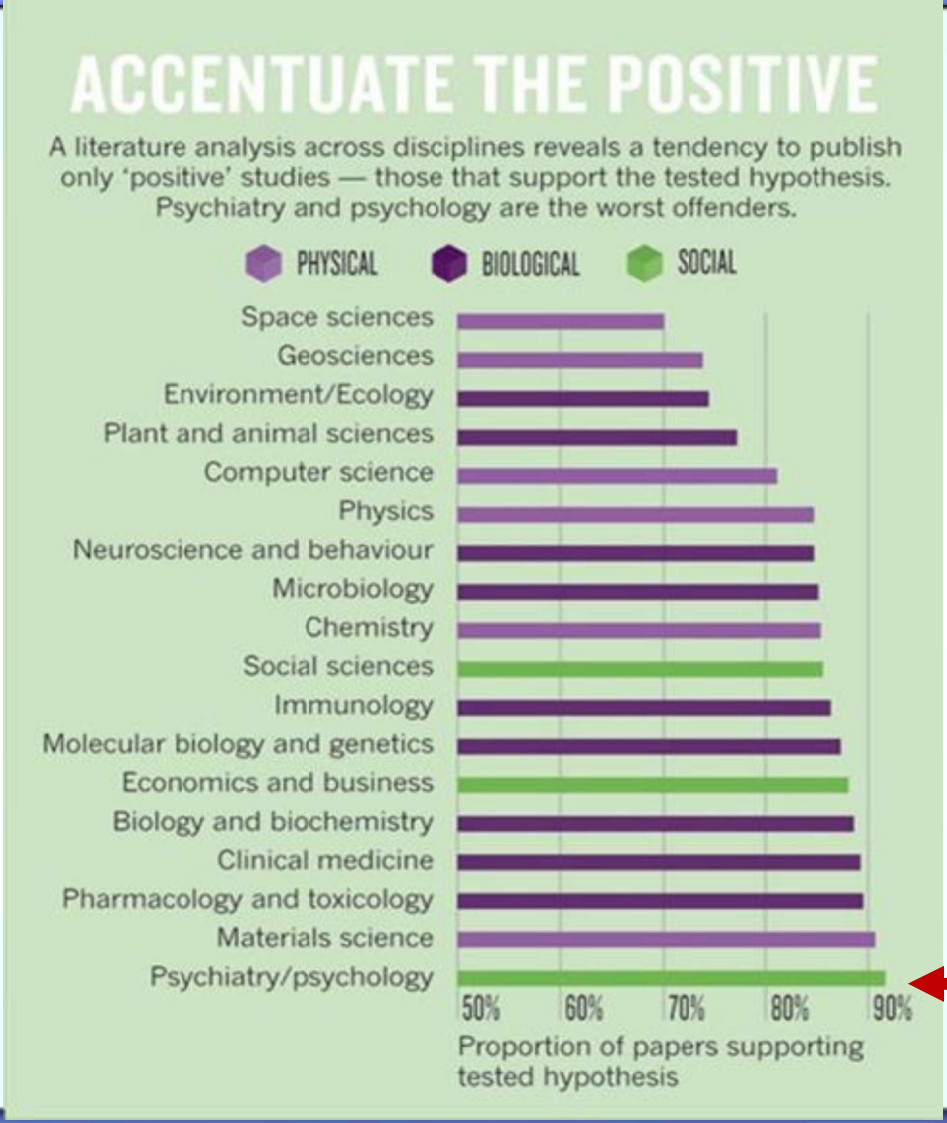
N for power:

80%	**90%**
72 Ss (vs. 278)	98 Ss (vs. 382)

Suppose we run a study with 98 Ss.
Expected power is 0.90 but **real power will be 0.43**

Vicious cycle: PB leads to overestimated ES leading to underpowered studies leading to non replicated effects, **even assuming that the effects are true and the researchers do not "cheat"**

Is there publication bias in science?



YES

(Fanelli, 2010)

Publication bias, Effect Sizes, sample sizes

Without publication bias, there should be **no** relation (r=0)

September 2014 | Volume 9 | Issue 9 | e105825

OPEN ACCESS Freely available online

PLOS | ONE

Publication Bias in Psychology: A Diagnosis Based on the Correlation between Effect Size and Sample Size

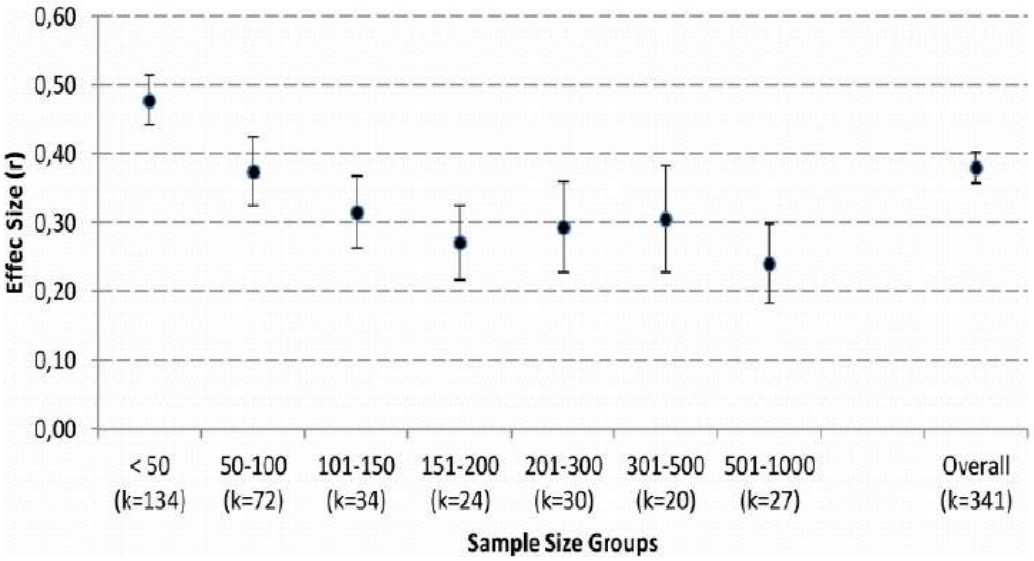
YES

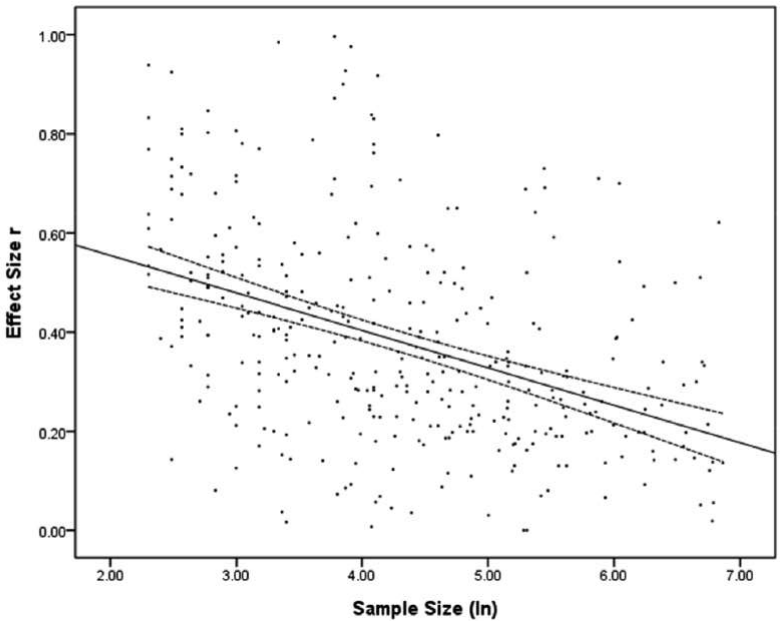
Anton Kühberger[1,2]*, Astrid Fritz[3], Thomas Scherndl[1]

1 Department of Psychology, University of Salzburg, Salzburg, Austria, 2 Centre for Cognitive Neuroscience, University of Salzburg, Salzburg, Austria, 3 Österreichisches

Methods: We investigate whether effect size is independent from sample size in psychological research. We randomly sampled 1,000 psychological articles from all areas of psychological research. We extracted *p* values, effect sizes, and sample sizes of all empirical papers, and calculated the correlation between effect size and sample size, and investigated the

r = .54 !





Publication bias, Effect Sizes, sample sizes

Without publication bias, there should be **no** relation

Neuroinform (2012) 10:67–80
DOI 10.1007/s12021-011-9125-y

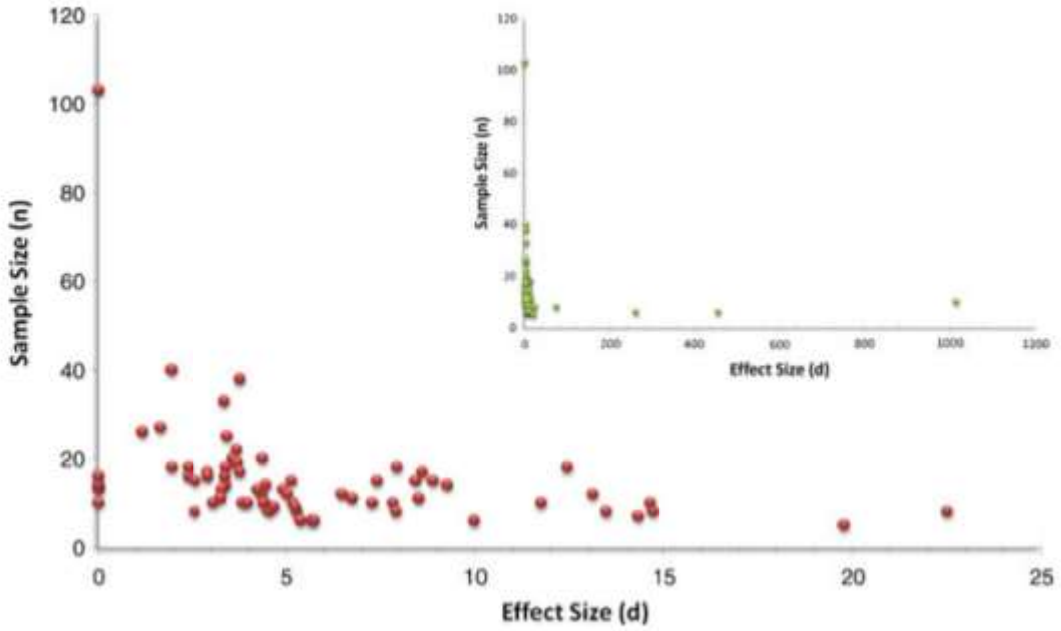
ORIGINAL ARTICLE

Publication Bias in Neuroimaging Research: Implications for Meta-Analyses

Robin G. Jennings · John D. Van Horn

YES

Fig. 3 Funnel plot of Cohen's d by sample size for studies without extreme values (n=70). While a 'large' Cohen's d value is usually d>0.8, most of our values are between 1 and 25, with funnel plot asymmetry due to the heavy right-tail evident here. **Inset:** Funnel plot of Cohen's d by sample size for each study (n=74), showing the four extreme outlier values

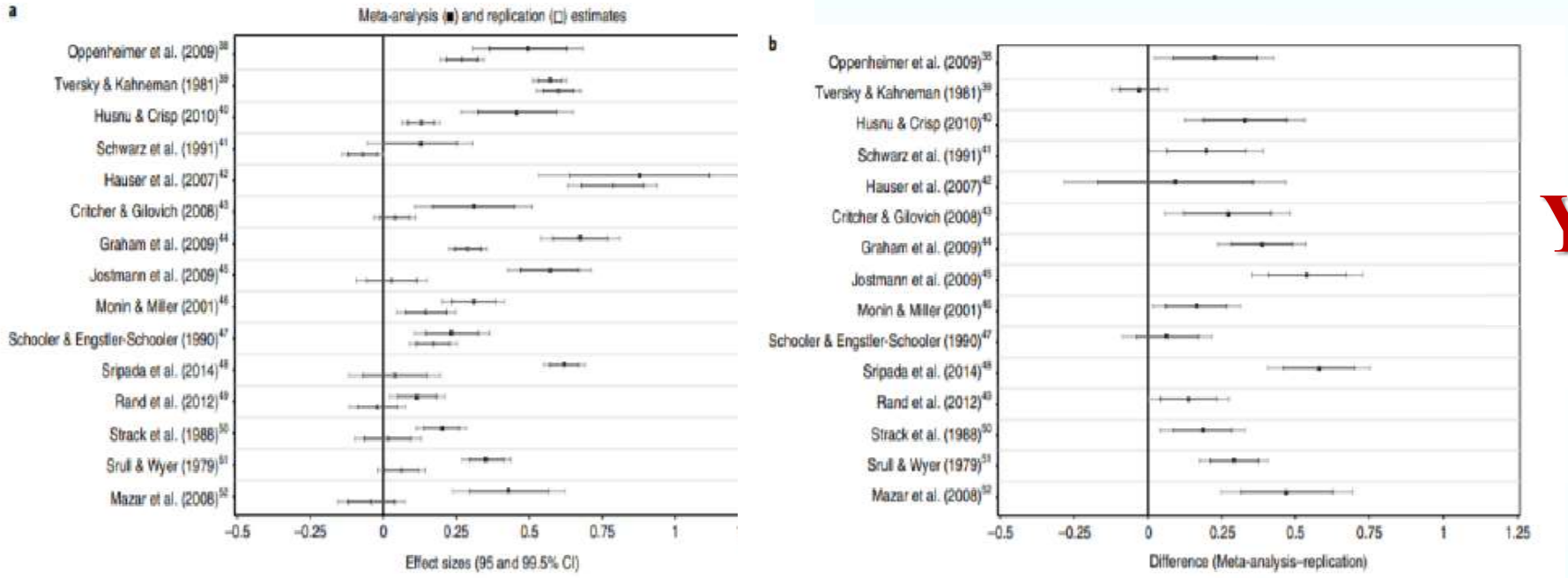


Meta-analysis vs. Replication

nature human behaviour
ARTICLES
https://doi.org/10.1038/s41562-019-0787-z
Corrected: Author correction

Comparing meta-analyses and preregistered multiple-laboratory replication projects

Amanda Kvarven[1,3], Eirik Strømland[1,3] and Magnus Johannesson[2*]



YES

7/15 (47%) replicated significant effect

12/15 (80%) smaller ES

Average ES (Cohen's d): 0.163 vs. 0.423

2nd pointer: Confirmatory studies

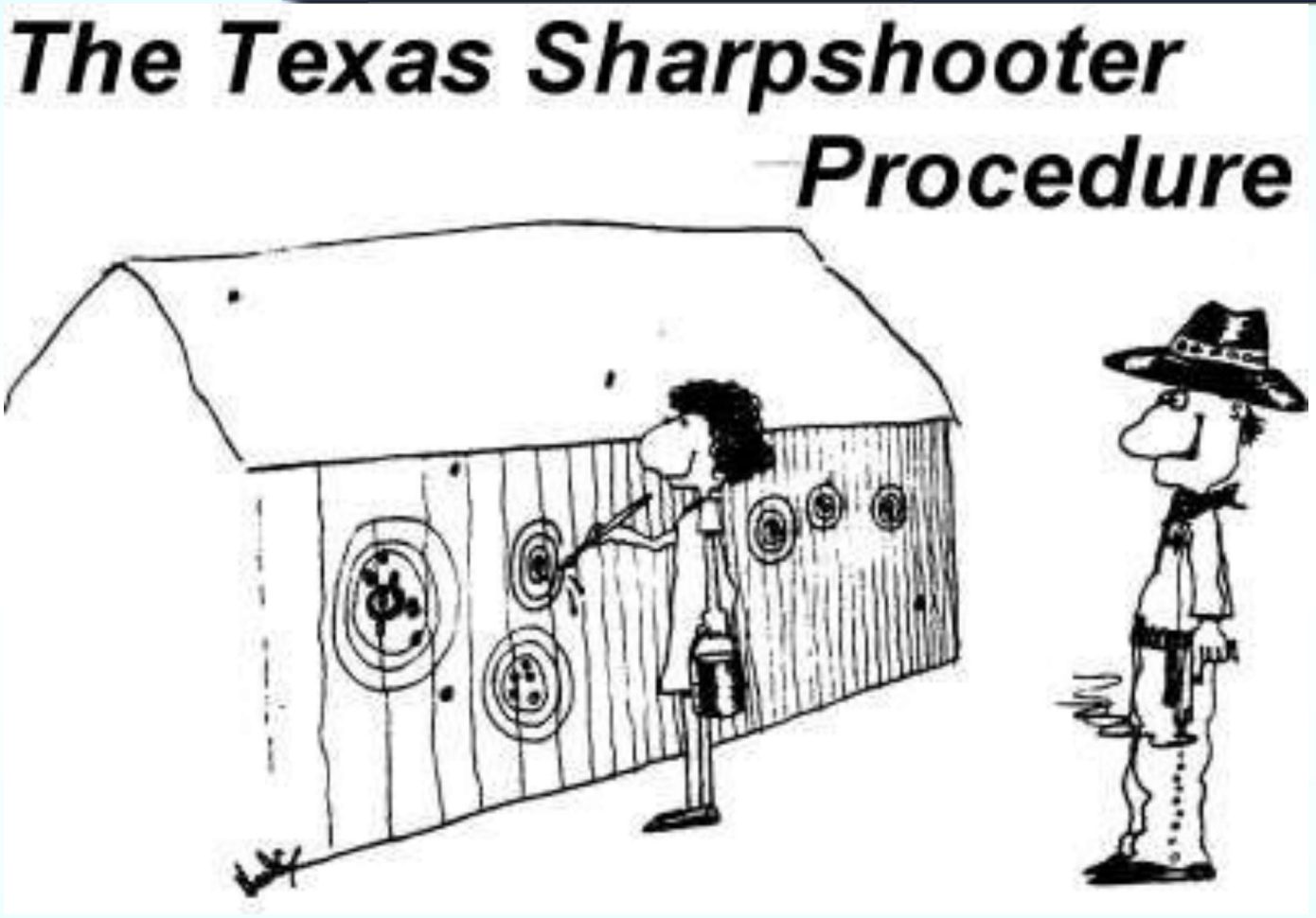
- Distinguish between exploratory and confirmatory studies
- If you find a “surprising effect”, confirm it with another well powered study before building on it
- Results can be significant simply out of random sampling
- Post-hoc ≠ pre-hoc
- Consider pre-registration (and other “badges” too)

Badges

- Encouraged practice for some journals (e.g., Psychological Science)

- More info at https://osf.io/tvyxz/wiki/home/ and https://osf.io/8uz2g/
- See also as https://aspredicted.org/

Why pre-registration?



I always avoid prophesying beforehand, because it is a much better policy to prophesy after the event has already taken place (Winston Churchill, 1943)

Why pre-registration?

2600–2606 | PNAS | March 13, 2018 | vol. 115 | no. 11

The preregistration revolution

Brian A. Nosek[a,b,1], Charles R. Ebersole[b], Alexander C. DeHaven[a], and David T. Mellor[a]

[a]Center for Open Science, Charlottesville, VA 22903; and [b]Department of Psychology, University of Virginia, Charlottesville, VA 22904

Progress in science relies in part on generating hypotheses with existing observations and testing hypotheses with new observations. This distinction between postdiction and prediction is appreciated conceptually but is not respected in practice. Mistaking generation of postdictions with testing of predictions reduces the credibility of research findings. However, ordinary biases in human reasoning, such as hindsight bias, make it hard to avoid this mistake. An effective solution is to define the research questions and analysis plan before observing the research outcomes—a process called preregistration. Preregistration distinguishes analyses and outcomes that result from predictions from those that result from postdictions. A variety of practical strategies are available to make the best possible use of preregistration in circumstances that fall short of the ideal application, such as when the data are preexisting. Services are now available for preregistration across all disciplines, facilitating a rapid increase in the practice. Widespread adoption of preregistration will increase distinctiveness between hypothesis generation and hypothesis testing and will improve the credibility of research findings.

3rd pointer: Meta-analytic

- Use meta-analytic **mind-set**
- Do not over-interpret significant or non-significant results in single studies
- The dance of p values
- There must be some studies that fail to replicate a real effect!
- Example with my own research