POWER, EFFECT SIZE, AND MEASUREMENT

STATISTICAL POWER

Statistical power is "the probability of rejecting a null hypothesis that is, in fact, false" (Williams, 1986, p. 67). Put more simply, statistical power is the probability of finding relationships or differences that in fact exist (Cohen, 1988).

In our fish story, it is the probability of finding minnows in Lake Alice, if they are in fact there. In terms of beta (the probability of a Type II error), statistical power = 1 - beta.

Statistical power is a function of "the preset significance criterion [alpha], the reliability of sample results, and the effect size [the actual size of the difference or strength of the relationship]..." (Cohen, 1988, p. 4).

Considering complex interrelationships of the above criteria, one can say that

The researcher can easily set alpha, but cannot easily set beta. Alpha and beta are directly, but not perfectly related. Lowering alpha increases beta and lowers the power. Increasing alpha decreases beta and increases power.

Statistical power is then related to:

- Sample size
- Effect size
- Statistical design (including number of groups, 1- vs. 2-tailed tests)
- Significance criteria

EFFECT SIZE

Effect size (ES) refers to the amount of common variance between the independent variable(s) (IV) and the dependent variable(s) (DV), or the degree to which changes in the IV(s) result in changes in the DV(s).

For example, if I am interested in the differences in competitive closure rate between rehabilitation counselors with master's degrees in rehabilitation counseling and those with bachelor's or unrelated master's degrees, my effect size would be the size of the difference between the means of the two groups. Or, if I wanted to test a specific intervention for students with learning disabilities, and I had a test, which I believed measured the effectiveness of my intervention; then my effect size might be the difference in test scores between an experimental group that received the intervention and a control group that did not receive the intervention. Similarly, if I wanted to examine the impact of a specific course on research anxiety, effect size could be the differences in the mean scores of research anxiety between an experimental group who completed the course and a control group who did not.

Here is a large problem: Effect size depends on what measure we use to operationalize the construct. For example, effect size depends on the net we use, the test we select, etc. Actual effect sizes may be much larger than observed effect sizes. What might be

considered a moderate to large effect in a laboratory situation may appear as a small effect in the real world where you can't control numerous sources of extraneous variance, e.g., variability in individual characteristics, treatment implementation, environmental characteristics (Cohen, 1988).

Small effect sizes are common and should be expected in ex post facto and quasi experimental situations (Cohen, 1988).

RELATIONSHIP OF MEASUREMENT, RESEARCH DESIGN, AND STATISTICAL POWER

This is just a conceptual introduction. We will return to validity of measurement in a future lecture.

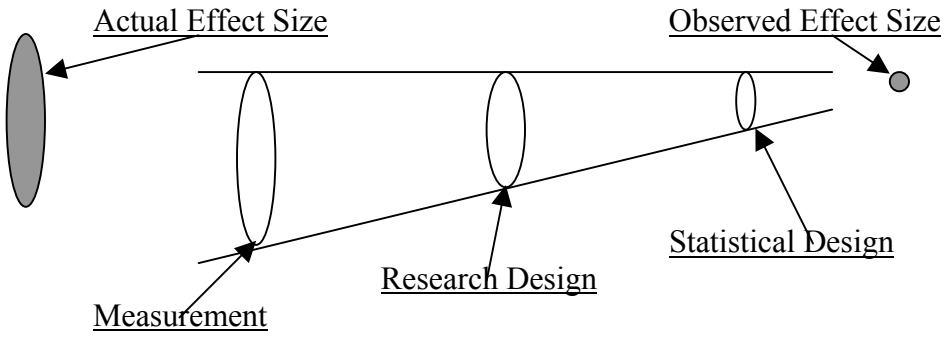
All research depends on an operational definition of the constructs of interest. In intervention research, the operational definitions of both the treatments and the outcomes influence effect size. As we are all aware, there are a variety of frames of reference regarding interventions and outcomes.

Consider the chapter 1 of the elephant fable with the researchers who mapped different parts of the elephant. Their descriptions of the elephant differed considerably. What we see in research depends, at least in part, on what facet(s) of the construct of interest is (are) operationalized by our outcome measure(s). It is always better to look at the construct in more than one way (more than one facet) in order to limit threats to validity from mono-operational bias. In other words, looking at the elephant from different angles can improve the degree to which our descriptions of the elephant actually describe the elephant.

Now, consider measuring the same elephant with portable X-Ray machines. Pictures of each part of the elephant are taken and then compared with each other. Not only do these pictures not resemble each other, but they also don't resemble the descriptions provided by the previous group of researchers. This chapter of the elephant fable indicates how what we see is indicated by our method of observation or measurement. Again, a researcher interested in a deeper understanding of the elephant may choose multiple methods of measurement in order to avoid threats to validity from mono-method bias.

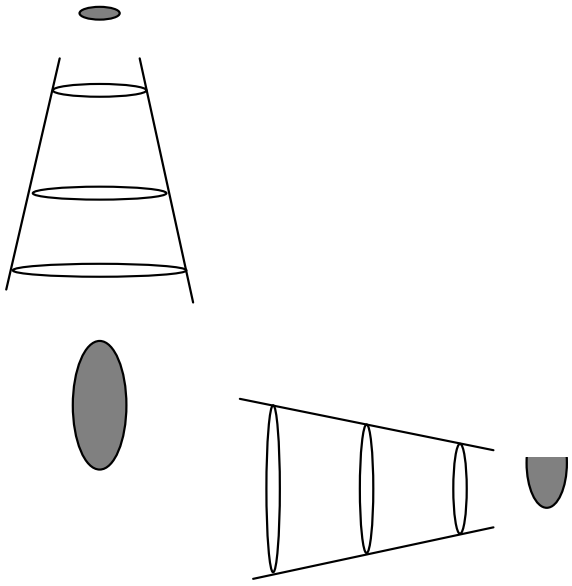
The relationship of measurement, research design, and statistical power means that large treatment effects can actually be observed as small effects. In other words, even if an intervention is very effective, measurement and design complications may make the effect appear small and thus require high statistical power for detection.

The following telescope model depicts the interrelation. The effect is obscured when we only look at part of the construct of interest. The apparent effect size is then attenuated by the extent to which our operational definitions (including our measurement techniques) do not reliably and validly capture the construct of interest (i.e., intervention effectiveness).



Apparent effect size is further attenuated when research design does not fully filter out extraneous sources of variation (e.g., counselor or client differences). Violations of assumptions of statistical procedures can further attenuate effect size. Interestingly, problems in research design and statistical design can also introduce sources of Type 1 error (e.g., dust on the lens or false positive results).

The relationship of effect size, measurement, and design is further complicated by the frame of reference or angle from which one approaches or operationalizes the construct. This complication is illustrated in the following figure.



Validity is a key element of the relationship of effect size, measurement, and design. Clearly, qualitative methods can further valid operationalization of constructs. Multiple operational definitions and multiple methods as recommended by Cook and Campbell (1979) can enhance the validity of research, including counseling effectiveness research. Further, units of measurement should be carefully considered in planning research. Researchers considering the social and cultural context of behavior have questioned the reductionist tradition of separating acts, actors, and audiences, as well as the tendency to study behaviors without consideration of social and cultural mediation (see e.g., Trueba,

Rodriguez, Zou, & Cintron, 1993; Wertch, 1991). Such questions clearly pose a challenge for effectiveness research.

Preanalysis Statistical Power Estimation

Preanalysis statistical power estimation is a recommended technique. The following steps will allow you to consider statistical power in research planning.

1. Estimate effect size from past research and the type of experimental design planned. When you are unsure, underestimate effect size so as to overestimate power. Also, in quasi-experimental or ex post facto circumstances, it is usually best to estimate a small effect size unless otherwise indicated.
2. Decide on exact statistical test and significance criterion.
3. Determine acceptable level of power, .80 is nice but .70 may be acceptable in some circumstances.
4. Use power tables for that statistical test or an appropriate computer program to determine the number of subjects required for the specified significance criterion and desired level of power.
5. If you have a fixed number of subjects, consider adjusting the significance criterion (alpha) or statistical design if necessary to obtain adequate power. Recall, the .05 significance criterion is not sacred, especially when it results in a power of less than .30 (i.e., less than a 30% chance of finding differences that actually exist). (Szymanski & Parker, 1992)

ALPHA INFLATION

Multiple comparisons can increase alpha, the probability of a Type I error. Recall the fish tank. As we learned from the Szymanski and Parker (1992) reading, the probability of a Type I error escalates with the number of comparisons made in the study. The experiment-wise alpha is computed as:

$$1-(1-\text{alpha})^{n}$$

As we discussed, one way to guard against alpha inflation is to use a Bonneferoni-type procedure and to split alpha by the number of comparisons. There are a variety of such procedures that can be used (see e.g., Marasciulo & Serlin, 1988) according to the relative importance of the tested hypotheses.

The problem with reducing alpha is that it inflates beta. In situations in which alpha inflation is accepted due to a problem with power, one must look to replications for confidence in the findings.

Again, let us consider Monet.

One study, alone, tells us little. However, one study, considered in relationship to others, tells us about patterns or trends in the relationships among variables.

MULTIPURPOSE POWER TABLES *

Table 1: **Power for Specific Statistics and Effect Sizes**

	Statistics			
Effect Sizes	t	r		F
A. Small	0.20	0.10	0.10	0.10
B. Medium	0.50	0.30	0.30	0.25
C. Large	0.80	0.50	0.50	0.40

Table 2: **N Required to Detect <u>Medium</u> Effect at .05 (two-tailed test)**

	Statistic			
Power	t	r		F(1)
0.20	10	15	<25	11
0.30	20	25	25	17
0.40	25	35	30	24
0.50	30	40	45	31
0.60	40	55	55	39
0.70	50	65	70	50
0.80	65	85	90	62
0.90	85	115	120	84
	N/G	N P's	Tot N	N/G

Table 3: **N Required to Detect <u>Small</u> Effect at 0.05 (two-tailed test)**

	Statistic			
Power	t	r	X^2	F(1)
0.20	65	125	125	65
0.30	105	200	200	105
0.40	150	300	300	150
0.50	200	400	400	200
0.60	250	500	500	250
0.70	300	600	600	300
0.80	400	800	800	400
0.90	550	1000	1000	550
	N/G	N P's	Tot N	N/G

*F(1) = F with 2 Groups
N/G = N per group
N P's = number of pairs of scores
Tot N = total N required

*from Professor Randall Parker and excerpted from Rosenthal & Rosnow (1984)