

AEA/CDC Summer Evaluation Institute

Offering 49: Exploring Effect Size and Measures of Association

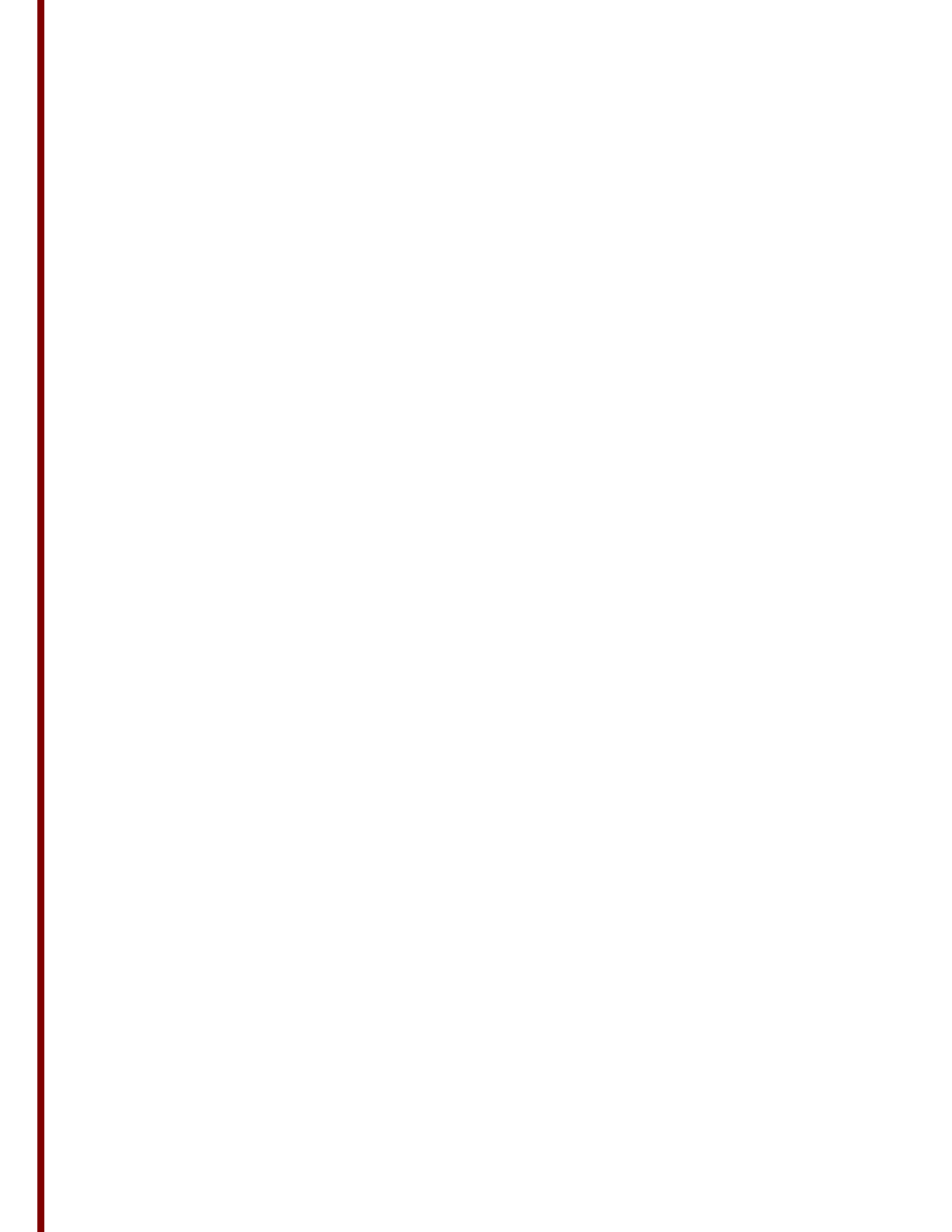
Description: Answer the call to report effect size and association measures as part of your evaluation results. This workshop will improve your capacity to understand and apply a range of measures including: standardized measures of effect sizes proposed by Cohen, Glass, and Hedges; Eta-squared; Omega-squared; the Intraclass correlation coefficient; and Cramer's V. Through mini-lecture and demonstration you will improve your understanding of the theoretical foundation and computational procedures for each measure. The session will include: definitions of and procedures for computing a range of effect size and association measures, a presentation that examines the relationships among the common measures, and description of computation of selected confidence intervals for effect sizes and association measures. You will receive SPSS and SAS software program codes for performing many of the computations related to the measures and common confidence intervals.

Audience: Attendees with an understanding of basic statistics through regression and working in any context

Jack Barnette, PhD has served as a faculty member at Penn State University, University of Virginia, University of Memphis, University of Alabama at Tuscaloosa, University of Iowa, and is now Senior Associate Dean for Academic Affairs and Professor of Biostatistics at the University of Alabama at Birmingham. He has served as an APHA Statistics Council Member and Section representative to the APHA Action Board. Presently, he is chairing the ASPH biostatistics competency workgroup and is co-chair of the ASPH Biostatistics/Epidemiology Section. He has more than 30 years experience in teaching, advising students, and applying research, evaluation, and statistical methods to a wide variety of educational and public health projects. He has conducted evaluations of projects funded by CDC, HRSA, SAMHSA, NHLBI, and NIOSH. He serves on three of the ASPH/CDC Preparedness Exemplar Groups: Education and Evaluation Methods, Certificate Programs, and University-based Student Preparedness. He has been conducting research on the use of effect sizes and measures of association for the past eight years and he has presented pre-sessions on this topic at the last four AEA annual meetings. He holds the PhD in Educational Research and Development from Ohio State (1972).

Offered (Two Rotations of the Same Content - Do not register for both):

- Monday, June 23, 9:25 – 12:45 (20 minute break within)
- Tuesday, June 24, 9:25 – 12:45 (20 minute break within)



Effect Size and Measures of Association
2008 Summer Evaluation Institute
Sponsored by The American Evaluation Assoc. and
The Centers for Disease Control and Prevention,
June 23 and 25, 2008

J. Jackson Barnette, PhD
Professor of Biostatistics
School of Public Health
University of Alabama at Birmingham

Four Sessions in this Workshop

1. Overview of Effect Size Measures
2. Technical Definitions and Equations for Computation
3. Computational Examples of Commonly Used Effect Size and Association Measures and Available Confidence Intervals
4. A brief look at the relationships of these

Plus an open discussion of some applications

Dr. Jack Barnette Eval Inst 2008 - Effect Size 2

Session 1
Title of this Session

Overview of effect size measures
or
“When a good p is not quite enough”

Dr. Jack Barnette Eval Inst 2008 - Effect Size 3

Outline for Session 1

1. A brief history of effect size evolution
The significance testing approach
The need for more than significance testing
Focus on finding other measures that provide more information
2. General categories of "effect size"
 - a. Confidence intervals
 - b. Percent variance-accounted-for statistics
 - c. Effect sizes in standardized units of difference
3. The call in the literature
 - a. Examples from publication manuals
 - b. Examples from journal guidelines
5. Issues that need to be addressed
6. The future of effect size reporting

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

4

A brief history of effect size evolution

Three general areas of discussion:

1. Significance Testing, the p -value
2. The desire for more
3. Focus on finding other measures that provide more of a picture of the effects or differences

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

5

References

There are a lot of references to this history
I have summarized the history as reported in:

Kline, R. (2004). *Beyond significance testing: Reforming data analysis methods in behavioral research*. Washington: APA.

and

Nix, T. & Barnette, J. (1998). The data analysis dilemma: Ban or abandon. A review of null hypothesis significance testing. *Research in the Schools*, 5(2), 3-114.

Carl Huberty (UGA Prof. Emeritus) has published a lot on this topic

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

6

Significance Testing, the p -value

Clearly, hypothesis testing has been a major activity of research for almost a century
Null Hypothesis Significance Testing (NHST) has been around a long time
Was a hybrid of two different schools of thought with proponents (Fisher and K. Pearson) who had a very contentious relationship (they hated each other)

Dr. Jack Barnette Eval Inst 2008 - Effect Size 7

Significance Testing, the p -value

Ronald Fisher (father of ANOVA), in the 20's, described the testing of a null hypothesis and used p -values to do that, but he did not set the 0.05, 0.01 criteria we use today
Neyman and Pearson, in the 30's, described the expansion of Fisher's model by adding the notion of the alternative or research hypothesis
The use of the p -value as compared with standards of 0.05 and 0.01 followed soon after that.

Dr. Jack Barnette Eval Inst 2008 - Effect Size 8

Significance Testing, the p -value

Neyman and Pearson proposed the fixed alpha approach and also described the notion of power and Types of error (Type I and Type II)
This approach really caught on and was the primary data analysis framework from the 40's into the 60's
Tremendous growth in use of these methods in published research, 85% in 60's and 90% since

Dr. Jack Barnette Eval Inst 2008 - Effect Size 9

Significance Testing, the p -value

There was always some controversy about this methodology

Some reasons for this were:

1. Misunderstandings of p -values
2. Didn't seem to promote replication
3. Many p -values are not very meaningful
4. p -values so specific that they don't tell researchers what they need to know

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

10

Significance Testing, the p -value

5. statistical significance provides no information about size of effects

The p -value just wasn't enough information to meet the needs to know more about the research results

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

11

The Desire for More

In the 70's Gene Glass proposed the use of meta-analysis that required information on effects beyond or other than the p -value

The "meta-analysis thinking" was a strong influence on the development of effect size and strength of association measures

In early 90's there were suggestions to include effect sizes in reporting (APA Style Manual, 1994) but these did not catch on, encouragement did not work

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

12

The Desire for More

In 1999, the Task Force on Statistical Inference (TFSI) was formed to report on the controversy about significance testing and to promote the use of alternative methods

Some wanted a ban on NHST

I was included in this group; sort of not my true feelings on the matter

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

13

The Desire for More

The TFSI had a strong influence on bringing the practice of reporting effect sizes more into the "thou will" level than "thou should consider it"

We'll see that in a bit when we look at the wording in the 2001 APA Style Manual

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

14

Focus on Finding other Measures

Many of the methods we use for effect size reporting have been around a long time
 r and η a real long time ago, pre-Fisher

Fisher described eta-squared or the correlation ratio as a measure of variance accounted for

Cohen's d about 1962, Glass effect size about 1976, and Hedges effect size about 1981

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

15

Focus on Finding other Measures

The TFSI also described the different effect sizes that could be used.
There are some that are used fairly extensively and those will be the topic of this workshop

Dr. Jack Barnette Eval Inst 2008 - Effect Size 16

It Just Made Sense

The problem with the *p*-value is that it gave virtually no information that could be used to make decisions about how to use the results in "real world" rather than probability terms
We needed information related specifically to the variables, how they worked, and how they were a part of the broader context
We needed practical and clinical significance as well

Dr. Jack Barnette Eval Inst 2008 - Effect Size 17

It Just Made Sense

We needed information that would tell us things like:
What change in reading level will we see if we use this program?
Is it cost effective to use this program?
Is this program effective in changing behavior of this group?
Have competencies been developed by this group?

Dr. Jack Barnette Eval Inst 2008 - Effect Size 18

It Just Made Sense

Effect sizes permit assessment of change or difference on the scale tied to the real world
All *p*-values do are to indicate that a difference could be or should not be attributed to chance
As I tell my students, statistical methods have no sense of the reality of the numbers they use
Only the user can put in into context

Dr. Jack Barnette Eval Inst 2008 - Effect Size 19

It Just Made Sense

The three general types of effect size/association measures are designed to help the user relate outcomes to the real world context
Confidence intervals and standardized effect sizes are in the metric of the dependent variable in units of standard deviation of difference
Association measures are a little less useful since they are in the form of proportion of variance of the dependent variable accounted for

Dr. Jack Barnette Eval Inst 2008 - Effect Size 20

Importance of Effect Size

- Knowledge of the magnitude of a treatment effect is qualitatively different than knowing if the effect is real.
- Real effects may be very important or very unimportant.
- Effect sizes force us to think beyond statistical significance.
- The reporting and interpreting are being required by more and more journals (more to come on this).

Dr. Jack Barnette Eval Inst 2008 - Effect Size 21

The Concept of Effect Size

A common definition of Effect size (ES) is that it is a family of indices that measure the magnitude of a treatment effect.

Many believe ES estimates to be independent of sample size (unlike statistical tests). ES measures are the common currency of meta-analysis studies that summarize the findings from a specific area of research.

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

22

How do we use "Effect Sizes" in Research and Evaluation?

- Determine sample size
- Assess practical significance (meaningfulness of findings)
- Provide an alternative view of effect in addition to significance tests
- Use as the metric in meta-analyses for combining studies

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

23

General Categories of "effect size"

Four general categories of effect size

1. Confidence intervals
2. Odds Ratios and Relative Risks
3. Effect sizes in standardized units of difference
4. Variance-accounted-for statistics

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

24

Confidence Intervals

Confidence intervals are usually not included in categorizations of effect sizes, but since they are also ways of describing the magnitude of effects they probably should be included

In fact, we are now seeing a call for confidence intervals around observed standardized effect sizes and measures of strength of association

Why not have confidence intervals of confidence interval limits? No reason we couldn't.

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

25

Effect Sizes in standardized units of difference

Some of the topics we will discuss are referred to as standardized effect sizes

An effect size would be the observed difference on means, proportions, etc.

For them to be useful for comparative purposes they need to be standardized

Sort of like z scores, they are an observed difference in units of standard deviation

Often the "standardized" part is dropped

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

26

Effect Sizes in standardized units of difference

These can be positive or negative, but we usually have them in absolute difference form

We will discuss three of these, spending the most time on the one seen most often

Cohen's d (most common one)

Glass's g'

Hedges's g

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

27

Variance-Accounted-for Statistics

These statistics are very similar (or even the same) to the notion of the correlation coefficient (r) and the coefficient of determination (r^2). The correlation measures the linear relationship between two variables and the squared correlation is a measure of the proportion of variance that is common to the two variables

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

28

Variance-Accounted-for Statistics

This same notion is found in the variance-accounted-for statistics. They cover the range of 0 where none of the variation in the dependent variable is attributable to the variation of the treatment variable to 1 where all of the variation in the dependent variable is attributable to the variation of the treatment variable

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

29

Variance-Accounted-for Statistics

r^2 is one such statistic
There are several others that we may or may not recognize as variance-accounted-for statistics
Cronbach's Alpha and Cohen's Kappa are also examples
Others we use in effect size assessment are usually referred to as strength of association or strength of relationship measures

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

30

Variance-Accounted-for Statistics

The examples we will discuss are:
Eta-squared, η^2 , also known as the
correlation ratio
Omega-Squared, ω^2
Intraclass Correlation, ICC or ρ_I
Cramer's V for chi-square applications

Current Journal Requirements

Many journals are requiring the reporting of
effect sizes or measures of strength of
association along with significance tests.

Reviewers and editors are sending
manuscripts back for revision if they do not
provide such information.

The APA Style Manual

Many publication manuals including the
APA have recognized the importance of
effect size and are encouraging or
requiring it be addressed for
manuscripts to be published.

APA on Effect Sizes and Confidence Intervals

“When reporting inferential statistics (e.g. *t* tests, *F* tests, and chi-square), include information about the obtained magnitude or value of the test statistic, the degrees of freedom, the probability of obtaining a value as extreme or more extreme than the one obtained, and the direction of the effect.”

--p. 22.

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

34

APA on Effect Sizes and Confidence Intervals

“The reporting of confidence intervals (for estimates of parameters, for functions of parameters such as differences in means, and for effect sizes) can be an extremely effective way of reporting results. Because confidence intervals combine information on location and precision and can often be directly used to infer significance levels, they are, in general, the best reporting strategy.

continued on next page

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

35

APA on Effect Sizes and Confidence Intervals

The use of confidence intervals is therefore strongly recommended. As a rule, it is best to use a single confidence interval size (e.g. a 95% or 99% confidence interval) throughout the course of the paper.”

--p. 22

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

36

APA on Effect Sizes and Confidence Intervals

“Effect size and strength of relationship.”

Neither of the two types of probability value¹ directly reflects the magnitude of an effect size or the strength of a relationship. For the reader to fully understand the importance of your findings, it is almost always necessary to include some index of effect size or strength of relationship in your Results section.

continued on next page

¹ $p < \alpha$ or $p = .12$ for example, not a part of quote

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

37

APA on Effect Sizes and Confidence Intervals

You can estimate the magnitude of the effect or the strength of the relationship with a number of common effect size estimates, including (but not limited to) r^2 , η^2 , ω^2 , R^2 , Φ^2 , Cramer's V, Kendall's W, Cohen's d and κ , Goodman-Kruskal's λ and γ , Jacobson and Truax's (1991) and Kendall's (1999) proposed measures of clinical significance, and the multivariate Roy's Θ and the Pillai-Bartlett V.

continued on next page

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

38

APA on Effect Sizes and Confidence Intervals

As a general rule, multiple degree-of-freedom effect indicators tend to be less useful than effect indicators that decompose multiple degree-of-freedom tests into meaningful one degree-of-freedom effects—particularly when these are the results that inform the discussion.

continued on next page

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

39

APA on Effect Sizes and Confidence Intervals

The general principle to be followed, however, is to provide the reader not only with enough information about statistical significance but also with enough information to assess the magnitude of the observed effect or relationship.”

-- p. 25-26

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

40

The Source

American Psychological Association. (2001). Publication manual of the American Psychological Association (5th ed.). Washington, DC: Author.

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

41

Educational and Psychological Measurement Guidelines

Two aspects of these issues from *EPM*:

1. Call for effect sizes through APA Guidelines
2. Call for confidence intervals for score reliability coefficients

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

42

Educational and Psychological Measurement Guidelines

Fan, X & Thompson, B. (2001). Confidence intervals about score reliability coefficients, please: An *EPM* guidelines editorial. *Educational and Psychological Measurement*, 61, 517-532.

“This guidelines editorial also promulgates a request that *EPM* authors report confidence intervals for reliability estimates whenever they report score reliabilities and note what interval estimation methods they have used.”

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

43

Assistance for Doing This from ME

Article in *EPM*:

Barnette, J. (2005). ScoreRel CI: Software for computation of confidence intervals for commonly used score reliability coefficients. *Educational and Psychological Measurement*, 65,980-983.

Software referred to in the article will be given to participants on request

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

44

Required Reporting in Journals

A large number of journals now require the reporting of effect sizes especially when significance testing is used

If you are submitting a manuscript, read the guidelines carefully to see the journal's requirements

Failure to do this, if expected, is easy to spot and will often require revision if the ms is publishable otherwise

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

45

Unresolved Issues

There needs to be greater understanding of what they can do and what they can't do
Misconceptions of interpretation
Identifying relationships and conversions
Getting a clear understanding of how effect size confidence intervals are computed
Assessing how useful effect size confidence interval can be

Dr. Jack Barnette Eval Inst 2008 - Effect Size 46

The Future of Effect Size Reporting

They are here to stay
They clearly provide a different perspective on difference or effect that is useful
There will be more calls or even requirements for reporting effect sizes
Not sure where we will come down on effect size confidence intervals, at this point they don't seem to have much utility

Dr. Jack Barnette Eval Inst 2008 - Effect Size 47

Session 2

Technical Definitions
And Equations for
Computation

Technical Definitions and Equations

This session will concentrate on defining the various effect size and association measures. Will look at effect size/association measures for probabilities, means and proportions (based on Chi-square analysis) Later, we'll have computed examples and look at the confidence intervals for these and a few special topics

Effect Size / Association Measures

Based on:

Probabilities (risks or odds)
Means (*t* tests and ANOVA)
Proportions (chi-square types of tests)

Association Measures Based on Probabilities

Relative Risk

Odds Ratio

We will not be concentrating our session on these, but wanted to indicate they are often used as effect size/strength of association measures in meta-analysis

Relative Risk and Odds Ratio

- Relative Risk and Odds Ratios are types of ratios that are used extensively
- There is some tendency to think these are the same and sometimes the one that is calculated is actually the other one
- They are expected to be similar but they may not be
- They are not interchangeable

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

52

Contingency Table (Frequencies in Cells)

	Condition		Total
	Present	Not Present	
Risk Factor Present	A	B	A+B
Risk Factor Absent	C	D	C+D
Total	A+C	B+D	N

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

53

Relative Risk

A way of defining this is:

$$RR = \frac{P(A) \text{ in Treated or Exposed Group}}{P(A) \text{ in Non-Treated or Non-Exposed Group}}$$

The ratio of the probability of some event happening such as death or disease in the group who receives the treatment or is exposed to the probability of that event happening in the group that does not receive treatment or exposure

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

54

Relative Risk

Relative Risk is the ratio of those exposed who have the outcome condition to those who are not exposed but who have the outcome or condition

$$\text{Relative Risk (RR)} = \frac{A/(A+B)}{C/(C+D)}$$

If RR is equal to 1.00 the same risk is in both the exposed and unexposed groups

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

55

Relative Risk

If RR is less than one, the exposed group has a lower likelihood of having the condition as compared with the non-exposed group

If RR is greater than one, the exposed group has a higher likelihood of having the condition as compared with the non-exposed group

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

56

Odds Ratio

The Odds Ratio is a ratio of two odds:
We would have the odds from two groups to compare for the odds ratio:

$$\text{Odds Ratio} = \frac{\text{Odds for Group 1}}{\text{Odds for Group 2}}$$

$$\text{Odds Ratio} = \frac{\text{Odds of Being Exposed in Group with Condition}}{\text{Odds of Being Exposed in Group without Condition}}$$

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

57

Odds Ratio

The Odds Ratio is the ratio that a patient or subject in an experimental group is exposed to the risk factor relative to a patient or subject in the comparison group is exposed to the risk factor

$$\text{Odds Ratio (OR)} = \frac{A/C}{B/D} = \frac{AD}{BC}$$

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

58

Confidence Intervals

- For both the relative risk and the odds ratio, the distributions around these values are not normal, but the log of the distributions is distributed normally
- We find the standard error in log units, then determine lower and upper limits of the log of these values and then we convert them back to the original scale

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

59

Strength of Association using Odds

For both the Relative Risk and Odds Ratio, no association is represented by 1.

The greater the departure from 1, the stronger the relationship

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

60

Use as Effect Size

- Relative Risk and Odds Ratio are often used in the health sciences, not so much in evaluation
- We can easily find confidence intervals for them (free program available from JJB)
- They are on a common metric so use in meta-analysis is clean and logical
- Converting to other measures of effect size can be problematic, have no direct equations for doing this

Dr. Jack Barnette Eval Inst 2008 - Effect Size 61

Data Set for Facial Injuries Example, with (Cell Labels)
data from "A Case-Control Study of the Effectiveness of Bicycle Safety Helmets in Preventing Facial Injury," by Thompson, Rivara, and Wolf, American Journal of Public Health, Vol. 80, No. 12).

	Facial Injuries Received	All Injuries Nonfacial	Total
Helmet Worn	30 (A)	83 (B)	113 (A+B)
No Helmet Worn	182 (C)	236 (D)	418 (C+D)
Total	212	319	531

Dr. Jack Barnette Eval Inst 2008 - Effect Size 62

Relative Risk and Odds Ratio with Confidence Intervals

We can use the program written by Barnette to find all of these values

We'll see how this is done in Session 3

Dr. Jack Barnette Eval Inst 2008 - Effect Size 63

Effect Size Measures Based on Means

These provide estimates of differences in units of standard deviation.

Most common ones are:

- Cohen's Effect Size
- Glass's Effect Size
- Hedges's Effect Size

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

64

Cohen's Effect Size (1969)

Cohen's d was the first commonly recognized effect size. It represented mean differences in units of common population standard deviation.

Population Form

$$\delta = \frac{\mu_1 - \mu_2}{\sigma}$$

Statistic Form

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s}$$

Cohen used this as the basis his research on power. σ and s represent the total score standard deviation

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

65

Glass's Effect Size (1978)

Glass proposed a modification of the Cohen d where the common standard deviation was replaced with the standard deviation of the control group.

$$g^1 = \frac{\bar{X}_{Experimental} - \bar{X}_{Control}}{S_{Control}}$$

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

66

Hedges's Effect Size (1981)

Hedges felt a better estimate of effect size might be based on a pooled (variance) standard deviation rather than the standard deviation of one of the groups.

$$g = \frac{\bar{X}_1 - \bar{X}_2}{S_{Pooled}}$$

This pooled variance term is the same one that would be used in the *t* test

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

67

Use of these Three

Of these, *d* (Cohen) tends to be used most often

Glass's can vary considerably from the other two since it is based on only one group (the Control Group)

Hedges's will tend to be slightly lower (more conservative as compared with Cohen's *d*), get closer together as sample size increases

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

68

The Effect Sizes

There are multi-sample versions of all of these.

They are the same in form and will provide relatively similar results depending on how the standard deviation estimates may vary.

Again, they provide a measure of difference relative to units of standard deviation.

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

69

Interpreting the Effect Sizes

A d or g' or g of .25 indicates that the range of difference among the means is one-fourth of the size of the standard deviation.

A d or g' or g of 1.2 indicates that the range of difference among the means is one and two-tenths of the size of the standard deviation.

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

70

Cohen's Standards

There is always the issue of "how large an effect size is meaningful?" It is suggested that this be based on other similar finding in the literature and/or what would be a meaningful difference for pedagogic, treatment, fiscal, etc. reasons.

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

71

Cohen's Standards

Cohen needed to base his research on power on some effect sizes so he pretty much arbitrarily chose three values that had been used extensively as standards for effect sizes:

0.2 is a "small effect"

0.5 is a "medium effect"

0.8 is a "large effect"

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

72

Cohen's Standards

Cohen warned about using these in practice, but these seem to have become the default values for a great deal of research.

The major problem with this is that effect sizes are influenced by the number of samples and the sample sizes.

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

73

Cohen's Standards

Thus, using these same standards across many different research arrangements is not a good practice. They may be attained just by chance in several situations.

In many cases, we don't have estimates of effect sizes for use in power determinations so these are often used, i.e. "will be looking to find a moderate effect size so we use $d=0.5$ "

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

74

Checking up on Standards

Here are some results from a Monte Carlo simulation study where 5,000 standardized effect sizes (comparing means) were generated at random from a unit normal distribution using 2 through 10 samples with samples sizes from 5 to 100 in steps of 5. Following are some of the results:

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

75

Observed Effect Sizes when $K= 2$

...and there is no "real" difference:

$n= 5$, mean= .56, $p>.2= .76$, $p>.5= .46$, $p>.8= .24$

$n= 25$, mean= .23, $p>.2= .49$, $p>.5= .08$, $p>.8= .01$

$n= 50$, mean= .16, $p>.2= .33$, $p>.5= .01$, $p>.8= .00$

$n=100$, mean= .12, $p>.2= .17$, $p>.5= .00$, $p>.8= .00$

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

76

Observed Effect Sizes when $K= 4$

$n= 5$, mean= .97, $p>.2= .99$, $p>.5= .85$, $p>.8= .60$

$n= 25$, mean= .41, $p>.2= .90$, $p>.5= .29$, $p>.8= .03$

$n= 50$, mean= .29, $p>.2= .74$, $p>.5= .06$, $p>.8= .00$

$n=100$, mean= .21, $p>.2= .49$, $p>.5= .00$, $p>.8= .00$

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

77

Observed Effect Sizes when $K= 10$

$n= 5$, mean= 1.40, $p>.2= 1.00$, $p>.5= 1.00$, $p>.8= .96$

$n= 25$, mean= .62, $p>.2= 1.00$, $p>.5= .76$, $p>.8= .14$

$n= 50$, mean= .44, $p>.2= .99$, $p>.5= .27$, $p>.8= .00$

$n=100$, mean= .31, $p>.2= .92$, $p>.5= .02$, $p>.8= .00$

Again remember NO REAL DIFFERENCE!!

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

78

Variance of Effect Sizes

Do means effect sizes vary by number of samples?

Do means of effect sizes vary by sample size?

Does the proportion of effect sizes achieving Cohen criteria differ by number of samples?

Does the proportion of effect sizes achieving Cohen criteria differ by sample size?

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

79

Variance of Effect Sizes

The answer to all of these questions is **YES!**
In fact there is great variation.

Should the same standards be used to judge effect sizes in all of these situations?

Don't think so, but that's the way it is done!

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

80

Another Possible Approach

Some work Jim McLean and I have worked on off-and-on is a method of predicting the values of d and η^2 that would occur by chance, permitting a comparison of observed with chance

A sort of "value-added" approach

Examples we have in practice that do a similar thing are Cohen's Kappa and Adjusted R^2

We'll see some of these results in the examples

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

81

Strength of Association Measures

These provide indications of the proportion of variance that can be attributed to the treatment

Most common ones are:

Eta-Squared, η^2

Omega-Squared, ω^2

Intraclass Correlation, ρ_i

Cramer's V

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

82

Eta-Squared (Pearson and Fisher)

Around the turn of the century, Pearson was using the correlation ratio (η). Fisher showed how it applied to the analysis of variance.

In squared form, it is the proportion of total variance attributed to the treatment.

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

83

Eta-Squared

A η^2 of 0.25 would indicate that 25% of the total variation is accounted for by the treatment variation.

$$\eta^2 = \frac{SS_{Treatment}}{SS_{Total}}$$

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

84

Eta-Squared

Positives: easy to compute, easy to interpret.

Negatives: it is more of a descriptive than inferential statistic, it has a tendency to be positively biased and chance values are a function of number and size of samples.
BUT IT'S NOW EXTENSIVELY USED AS AN INFERENTIAL STATISTIC in effect size reporting

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

85

Partial Eta-Squared

In a case of other than one-way ANOVA, often the error term for an effect is no longer MS_{Error}

In that case, partial η^2 is computed instead

$$\eta^2_{Partial} = \frac{SS_{Effect}}{SS_{Effect} + SS_{Error\ for\ Effect}}$$

It is interpreted in the same way as η^2

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

86

Relating η^2 to Effect Size (Cohen)

Cohen indicated an Eta-Squared of:

0.0099 relates to a Cohen "small effect" (0.2)

0.0588 relates to a Cohen "medium effect" (0.5)

0.1379 relates to a Cohen "large effect" (0.8)

Cohen may not have guided us well on this!

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

87

Relating η^2 to Effect Size d

You will see this equation referenced, attributed to Cohen (1988):

$$r = \frac{d}{\sqrt{d^2 + 4}}$$

Looks like converting from d to r (then squaring to get r^2) is pretty easy and straightforward

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

88

Relating η^2 to Effect Size d

And you will see the following equation, attributed to Friedman (1968):

$$d = \sqrt{\frac{2r}{1-r^2}}$$

DO NOT BE DECEIVED.

It is easy to think they are useful beyond that, but they are not. I'll demonstrate this later.

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

89

Relating η^2 to Effect Size (Cohen)

- Actually these relationships hold ONLY when $K=2$ and sample size is very large
- If $K > 2$ and/or sample size is small these relationships do not hold.
- Will demonstrate this later.

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

90

Omega-Squared (Hays, 1963)

When a fixed effect model of ANOVA is used, Hays proposed more of an inferential strength of association measure, referred to as Omega-Squared (ω^2) to specifically reduce the recognized bias in η^2 .

It provides an estimate of the proportion of variance that may be attributed to the treatment in a fixed design. $\omega^2 = .32$ means 32% of variance attributed to the treatment.

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

91

Omega-Squared

ω^2 is computed using terms from the ANOVA

$$\omega^2 = \frac{SS_{Between} - (k-1)MS_{Error}}{SS_{Total} - MS_{Error}}$$

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

92

Omega-Squared

Positives and Negatives (Pun intended) of ω^2

Positives: it is an inferential statistic that can be used for predicting population values, easily computed, it does remove much of the bias found in η^2 .

Negatives: it can have negative values, not just rounding error type, but relatively different than 0. If you get one that is negative, call it zero.

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

93

Omega-Squared Values

Values associated with levels of ω^2
According to Kirk (1995). Note these are the same
values Cohen gave for eta-squared.
Can this be true, equating η^2 and ω^2 ?

0.010 is a "small association" (0.2)
0.059 is a "medium association" (0.5)
0.138 is a "large association" (0.8)

How can this be? IT CAN'T BE

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

94

Intraclass Correlation

Omega-squared is used when the
independent variables are fixed.
Occasionally, the independent variables
may be "random" in which case the
intraclass correlation is used to assess
strength of association.

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

95

Fixed vs. Random Effects

There are labels applied to the nature of the
treatment or independent variables.

Fixed effect – a variable that has levels that
stand alone, not assumed to represent the
population of all possible independent
variables.

Random effect – a variable that has levels that
are assumed to be representative of the
population of possible independent variables.

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

96

Intraclass Correlation

Values to determine the ICC come from the ANOVA.
If unequal group sizes use the harmonic mean in place of n

$$ICC = \frac{MS_{Treatment} - MS_{Error}}{MS_{Treatment} + (n-1)MS_{Error}}$$

Dr. Jack Barnette Eval Inst 2008 - Effect Size 97

Intraclass Correlation

Can also find from:

$$ICC = \frac{F - 1}{(n - 1) + F}$$

The ICC is a variance-accounted-for statistic, interpreted in the same way as is Omega-Squared. It also has the same strengths and weaknesses.

Dr. Jack Barnette Eval Inst 2008 - Effect Size 98

Cramer's V

A measure of strength of association used when the statistical test is based on a chi-square distribution.

It is a proportion of variance accounted for statistic.

Dr. Jack Barnette Eval Inst 2008 - Effect Size 99

Cramer's V

Mean Square Contingency Coefficient (Cramer)
The measure of association for contingency tables

$$\phi^2 = \frac{\phi^2}{\phi_{Max}^2} \text{ estimated as } V = \frac{\chi^2}{Nq} \text{ where } q = \min.\{r-1, c-1\}$$

Has a range of 0 to 1 and is an indication of the proportion of common variance for two metric or ordinal variables using the chi-square statistic computed for a contingency table

Computational Examples

We'll take a look at how the common ones are computed using data

Then we'll look at how we determine confidence intervals for some of these using SPSS and SAS

Session 3

Computational Examples of Commonly Used Effect Size and Association Measures and Available Confidence Intervals

This Session

We will look at examples of computing the common effect size/strength of association measures

We'll include confidence intervals of some of these: odds ratio and relative risk using my program and d , η^2 , and Cramer's V using SAS macros.

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

103

Odds Ratio and Relative Risk

We'll look at an example of the use of odds ratios and relative risks since these could be considered types of effect sizes and they are often used in meta-analysis

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

104

Data Set for Facial Injuries Example, with (Cell Labels)
data from "A Case-Control Study of the Effectiveness of Bicycle Safety Helmets in Preventing Facial Injury," by Thompson, Rivara, and Wolf, *American Journal of Public Health*, Vol. 80, No. 12).

	Facial Injuries Received	All Injuries Nonfacial	Total
Helmet Worn	30 (A)	83 (B)	113 (A+B)
No Helmet Worn	182 (C)	236 (D)	418 (C+D)
Total	212	319	531

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

105

View of Barnette Software for Finding Relative Risk and Odds Ratios with Confidence Intervals, Facial Injury Data

Facial Injury → Condition or Outcome

		Yes	No	Total	
Helmet ↓	Exposure	Yes	30	83	113
	or	a	b	a + b	
	Treatment	No	182	236	418
		c	d	c + d	
	Total	212	319	531	
		a + c	b + d	N	

Dr. Jack Barnette Eval Inst 2008 - Effect Size 106

View of Barnette Software for Finding Relative Risk and Odds Ratios with Confidence Intervals, Facial Injury Data

Enter Confidence Interval of Interest (0.90, 0.95, 0.99 or other), CI Probability=

	0.9500
p	0.0250 0.9750
z	-1.9600 1.9600

Resulting Distribution assuming normal distribution of error

Dr. Jack Barnette Eval Inst 2008 - Effect Size 107

View of Barnette Software for Finding Relative Risk and Odds Ratios with Confidence Intervals, Facial Injury Data

$$RR = \frac{a/(a+b)}{c/(c+d)} = \frac{0.2655}{0.4354} = \boxed{0.6097}$$

$$OR = \frac{a/b}{c/d} = \frac{0.3614}{0.7712} = \boxed{0.4687}$$

Use this one since a retrospective study

Dr. Jack Barnette Eval Inst 2008 - Effect Size 108

View of Barnette Software for Finding Relative Risk and Odds Ratios with Confidence Intervals, Facial Injury Data

Confidence Intervals

Lower Limit Upper Limit

0.4403 ≤ RR true ≤ 0.8444

Sign. RR since 1 is not in CI

Use this one since
a retrospective study

0.2958 ≤ OR true ≤ 0.7425

Sign. OR since 1 is not in CI

Dr. Jack Barnette Eval Inst 2008 - Effect Size 109

Effect Size Confidence Intervals

Who can argue against the value of confidence intervals?

There has been new attention to the desire to have confidence intervals around effect size measures (d and η^2 in particular).

Effect Size Confidence Intervals

- We are all familiar with confidence intervals using means, mean differences, proportions, and proportion differences.
- They have one commonality and that is that they all use a model around zero as the parameter (referred to as based on 0 as the centrality parameter)

Dr. Jack Barnette Eval Inst 2008 - Effect Size 111

Effect Size Confidence Intervals

- But, in the case of an effect size, we are looking at the confidence interval around the value of the effect size rather than 0.
- While the probability model we use will be t or F , it is not based on the expected value of the mean which is 0 for t or $df_e/(df_e - 2)$ for F
- It will be a noncentral t or F
- This makes the mathematics VERY complicated

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

112

Effect Size Confidence Intervals

- In the case of the central t or F , we have tables and easy to use programs to find points on the distributions
- In the case of the noncentral t or F , there would be an infinite number of such distributions, one for every possible value of the parameter (effect size) making tables of little value

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

113

Effect Size Confidence Intervals

- So here's what one would have to do:
- Determine the parameter estimate (the observed or corrected effect size)
- Using a computer program, find the confidence interval probability limits, say you want the $CI_{0.95}$, then you need to find the $p_{0.025}$ and $p_{0.975}$ points on the noncentral t or F

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

114

Effect Size Confidence Intervals

Then you need a computer program to find the inverse, converting the probability points from the statistic distribution back to the score scale that the effect size is on.

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

115

Effect Size Confidence Intervals

However, the effect size has a different value than the expected mean related to the mean of the *F* distribution, these values will be different and these will be the ones used to get points back to the original scale metric.

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

116

Effect Size Confidence Intervals

- Say we have a central *F* distribution with 2, 29 degrees of freedom, we would have an *F* probability distribution with a mean of 1.074
- If we wanted a 0.95 CI then the respective values corresponding with 0.025 and 0.975 limits would be 0.0253 and 4.201. These values border the middle .95 of the area

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

117

Effect Size Confidence Intervals

- The special section of *Educational and Psychological Measurement*, 61, August 2001, looks at the basis for computing these confidence interval and provided references and programs that may be useful.
- Many of the confidence intervals presented as examples seem so large that we wonder how useful these will be.
- We will look more closely at that occurrence

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

118

Smithson's Utilities

It is not an easy task to find the confidence intervals.

Michael Smithson has provided macros for use in SPSS and SAS programs to help, but there is a reasonable amount of work needed to use them

Go to:

<http://www.anu.edu.au/psychology/people/smithson/details/CIstuff/CI.html>

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

119

Confidence Interval Resources

- The Smithson SAS routines work fine
- I've had some trouble with the SPSS routines and they are a little harder to use
- There were some missing codes
- Here are a few other URLs with descriptions and programs:
 - <http://www.latrobe.edu.au/psy/esci/>
 - <http://www.cem.dur.ac.uk/ebeuk/research/effectsize/ESbrief.htm>

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

120

Effect Size Confidence Intervals

- We will see there needs to be much more work in this area for this to be easily used by the typical evaluator or researcher.
- You may be in for a **surprise**.
- Now to the examples of computation of effect sizes and some confidence intervals
- We don't have ways of finding confidence intervals for some of the effect size and association measures we use

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

121

t Test Example 2 Sample Independent

Mean Comparison, K= 2 Comparison

Group	n	Mean	SD	SE	CI _{0.95}
C	30	13.83	2.001	0.365	13.09 to 14.58
E	30	15.27	3.073	0.561	14.12 to 16.41
Total	60	14.55	2.671	0.345	13.86 to 15.22

$t_{obs} = 2.141, df= 58, p= 0.037$

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

122

t Test Example

ANOVA Summary Table

Source	df	SS	MS	F	p
Bet.	1	30.817	30.817	4.583	0.037
Error	58	390.033	6.725		
Total	59	420.850			

Note: $t^2 = F$

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

123

t Test Example

Cohen's *d*

$$d = \frac{\bar{X}_1 - \bar{X}_2}{s} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{MS_{Error}}} = \frac{15.27 - 13.83}{\sqrt{6.725}} = \frac{1.44}{2.593} = 0.555$$

Chance predicted values from Barnette/McLean for *d*:

$$K=2, n=30 \quad d_{chance} = \frac{1.625}{\sqrt{59}} = \frac{1.625}{7.681} = 0.212$$

Dr. Jack Barnette Eval Inst 2008 - Effect Size 124

d Confidence Interval Using SAS

d CI_{0.95} Computed Using Charles Roe's Program CI_{0.95} = 0.03 to 1.06

d CI_{0.95} Computed Using Smithson's SAS Macros

1. Use Nonct.sas to find limits of non-central t

Input: Obs	t	df	conf
1	2.141	58	0.95

(lower and upper limits of NonCentral t with 58 df)

Output: prlow	prupp	ncplow	ncpupp
0.975	0.025	0.13398	4.13027

2. Use these limits in t2d2samp.sas

Input: Obs	t	df	conf	prlow	prupp	ncplow	ncpupp	n1	n2
1	2.141	58	0.95	0.975	0.025	0.13398	4.13027	30	30

Output:	Obs	t	df	conf	ncplow	ncpupp	n1	n2	d	dlow	dupp
1	2.141	58	0.95	0.13398	4.13027	30	30	0.55280	0.034593	1.06643	

d = 0.553, *d* CI_{0.95} : 0.035 to 1.066 (carried out further)

Dr. Jack Barnette Eval Inst 2008 - Effect Size 125

d Confidence Interval

Notice how large this CI is?

*d*CI_{0.95} : 0.035 to 1.066

Where are Cohen's standards?
Are you surprised?

Dr. Jack Barnette Eval Inst 2008 - Effect Size 126

t Test Example

Glass's g'

$$g' = \frac{\bar{X}_{Experimental} - \bar{X}_{Control}}{s_{Control}} = \frac{15.27 - 13.83}{2.001} = \frac{1.44}{2.001} = 0.720$$

Hedges's g

$$g = \frac{\bar{X}_1 - \bar{X}_2}{s_{Pooled}} = \frac{15.27 - 13.83}{2.671} = \frac{1.44}{2.671} = 0.539$$

Dr. Jack Barnette Eval Inst 2008 - Effect Size 127

Comparing Effect Sizes

Type	Error Term	df	Value
Cohen	$\sqrt{MS_{Error}}$	$J(n-1)$	0.555
	Total SS/df		
Glass	$s_{Control}$	$n_c - 1$	0.720
Hedges	s_{Pooled}	$n_1 + n_2 - 2$	0.539

Pooled Sample 1 + Sample 2 Sum of Squares

Notice why Cohen d and Hedges g are so close?
Lots of folks compute Hedges's g and call in Cohen's d

Dr. Jack Barnette Eval Inst 2008 - Effect Size 128

The Dependent t (CAUTION)

In the often used t test comparing means from the same groups like a pre-test, post-test design, we need to be careful about finding the appropriate effect size

Say we have the following situation ($n=10$):

	Mean	SD
Pre	8.60	1.663
Post	10.10	1.713

$s_d = 1.434, r = 0.640$

Dr. Jack Barnette Eval Inst 2008 - Effect Size 129

The Dependent t (CAUTION)

The t test result with $df= 9$ is 3.308, $p= 0.009$
If we did a repeated measures ANOVA,

$$F = \frac{MS_{time}}{MS_{residual}} = \frac{11.250}{1.028} = 10.946, p = 0.009$$

Note that $t^2 = 3.308^2 = 10.943 = F$
As it should be (with a tad of rounding error).
But what about the effect size to use?

The Dependent t (CAUTION)

One of the measures we would use would be the
partial Eta-squared

$$\eta_{partial}^2 = \frac{SS_{time}}{SS_{time} + SS_{residual}} = \frac{11.250}{11.250 + 9.250} = 0.549$$

But what about a standardized effect size like d or
 g' or g ?

The Dependent t (CAUTION)

The temptation would be to use a Glass type
effect size with the pre-test SD as the
denominator

$$g' = \frac{\bar{X}_{post} - \bar{X}_{pre}}{s_{pre}} = \frac{10.10 - 8.60}{1.663} = 0.902$$

This will be a negatively biased estimate of the
effect size because the correlation of the pre
and post-test scores is not accounted for

The Dependent t (CAUTION)

Remember the correction for the standard error of mean differences in the dependent case?

$$s_{\bar{X}_d} = \sqrt{s_{\bar{X}_{pre}}^2 + s_{\bar{X}_{post}}^2 - 2r_{pre-post} * s_{\bar{X}_{pre}} * s_{\bar{X}_{post}}} = 0.453$$

This is the error term for the dependent t , but it is not the SD of the differences which is 1.434

So, one option would be a g (pooled) approach, which would be:

$$g = \frac{\bar{X}_{post} - \bar{X}_{pre}}{s_d} = \frac{10.10 - 8.60}{1.434} = 1.046$$

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

133

The Dependent t (CAUTION)

Another approach would use the results from the repeated measures ANOVA

This would be a d type effect size

$$d = \frac{\bar{X}_{post} - \bar{X}_{pre}}{\sqrt{MS_{residual}}} = \frac{10.10 - 8.60}{\sqrt{1.028}} = 1.479$$

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

134

The Dependent t (CAUTION)

So, which to use in this case?

I would not use the unadjusted Glass g ' here since the correlation of the pre and post tests is not accounted for

Either the g or d would be justified

d , however, may be removing too much of the variance, so g might be the best choice here, but a justification could be made for using d

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

135

Variance Accounted for Measures

These provide estimates of the proportion of variance of the dependent variable accounted for by the manipulation of the independent variable

A rather arbitrary standard is that 0.20 in a variance accounted for statistic indicates "practical significance"

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

136

t Test Example

Variance Accounted for Measures

Eta-Squared

$$\eta^2 = \frac{SS_{Between}}{SS_{Total}} = \frac{30.817}{420.850} = 0.0732$$

7.32% of Variance Accounted for
Chance predicted values from Barnette/
McLean for η^2 :

$$K=2, n=30 \quad \eta^2_{chance} = 0.5573 n^{-1.0250} = 0.0171$$

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

137

Confidence Interval for η^2 Using SAS

Use Smithson's NoncF2.sas macro

Input:

F	df1	df2	conf
4.583	1	58	0.95

Output: (NonCentral F limits)

Obs	F	df1	df2	conf	prlow	prupp	ncplow	ncpupp
1	4.583	1	58	0.95	0.975	0.025	0	17.0574

η^2	Lower	Upper
rsq	rsqlow	rsqupp
0.073231	0	0.22136

Estimate of η^2 is 0.0732 (same as computed above)

CI_{0.95}: 0 to 0.2214

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

138

η^2 Confidence Interval

Notice how large this CI is?

$\eta^2 \text{ CI}_{0.95} : 0 \text{ to } 0.2214$

Where is commonly used standard?
Are you surprised?

Dr. Jack Barnette Eval Inst 2008 - Effect Size 139

Test Example Variance Accounted for Measures Omega-Squared if Ind. Var. is Fixed

$$\omega^2 = \frac{SS_{\text{Between}} - (k-1)MS_{\text{Error}}}{SS_{\text{Total}} - MS_{\text{Error}}} = \frac{30.817 - (2-1)6.725}{420.850 - 6.725} = 0.0582$$

5.82% of the Variance Accounted for

Dr. Jack Barnette Eval Inst 2008 - Effect Size 140

Test Example Variance Accounted for Measures Intraclass Correlation if Ind. Var. is Random

$$ICC = \frac{MS_{\text{Treatment}} - MS_{\text{Error}}}{MS_{\text{Treatment}} + (n-1)MS_{\text{Error}}} = \frac{30.817 - 6.725}{30.817 + (30-1)6.725} = 0.1067$$

10.67% of the Variance Accounted for
ICC tends to be higher than Omega-squared, by a factor of two in many cases, but the difference disappears as Omega-Squared approaches 1

Dr. Jack Barnette Eval Inst 2008 - Effect Size 141

ANOVA Example, $K=5$

Grp	n	Mean	SD	SE	$CI_{0.95}$
1	30	14.47	2.933	0.535	13.37 to 15.56
2	30	12.07	2.303	0.421	11.21 to 12.93
3	30	15.27	3.073	0.561	14.12 to 16.41
4	30	14.97	3.023	0.552	13.84 to 16.10
5 (C)	30	12.80	2.124	0.388	12.01 to 13.59
Tot	150	13.91	2.965	0.242	13.47 to 14.35

Dr. Jack Barnette Eval Inst 2008 - Effect Size 142

ANOVA Example, $K=5$

ANOVA Summary Table

Source	df	SS	MS	F	p
Bet.	4	36.907	59.227	8.004	0.000
Error	145	1072.967	7.400		
Total	149	1309.874			

Dr. Jack Barnette Eval Inst 2008 - Effect Size 143

ANOVA Example, $K=5$

Cohen's d

$$d = \frac{\text{Range}}{\sqrt{MS_{Error}}} = \frac{15.27 - 12.07}{\sqrt{7.400}} = \frac{3.20}{2.720} = 1.176$$

Chance predicted values from Barnette/McLean for d :

$$K=5, n=30 \quad d_{\text{chance}} = 5.23883 / \sqrt{149} = 0.4292$$

Dr. Jack Barnette Eval Inst 2008 - Effect Size 144

ANOVA Example, $K=5$

Glass's (need to have a control group to use this)

$$g' = \frac{Range}{s_{Control}} = \frac{3.20}{2.124} = 1.507$$

Hedges's

$$g = \frac{Range}{s_{Pooled}} = \frac{3.20}{2.965} = 1.079$$

Dr. Jack Barnette Eval Inst 2008 - Effect Size 145

Problem with d in $K>2$ Situations

d when used with more than two groups looks across the range of means rather than each adjacent pairwise difference

There is no standard for comparison of this value

We can't use the typical d to η^2 conversion

In the more than 2 group situation, regardless of sample size, $d=0.5 \neq \eta^2=0.0588$

Dr. Jack Barnette Eval Inst 2008 - Effect Size 146

Problem with d in $K>2$ Situations

In the $K>2$ case, it is recommended that d be computed for 1 df pairwise comparisons, in our case there will be

$$c = \frac{K(K-1)}{2} = \frac{5(5-1)}{2} = \frac{20}{2} = 10$$

But, we probably don't need to do all 10 of them

Dr. Jack Barnette Eval Inst 2008 - Effect Size 147

Pairwise *d* Values

One approach to doing this which would cut down on the total number needed to do would be to conduct a post-hoc pairwise analysis such as Tukey's HSD and only compute *d* for the pairs that are significantly different

We'll do that for our example

Dr. Jack Barnette Eval Inst 2008 - Effect Size 148

Pairwise *d* Values

Tukey's HSD would be:

$$HSD = q_{\alpha, J, df_e} \sqrt{\frac{MS_E}{n}} = q_{0.05, 5, 145} \sqrt{\frac{7.400}{30}} = 3.91 * 0.497 = 1.94$$

Any pairwise difference of 1.94 or greater would be significant at $p < 0.05$. This would be:

$$\mu_2 \neq \mu_3 \quad \mu_2 \neq \mu_4 \quad \mu_2 \neq \mu_1 \quad \mu_3 \neq \mu_5 \quad \mu_4 \neq \mu_5$$

Dr. Jack Barnette Eval Inst 2008 - Effect Size 149

Pairwise *d* Values

Following would be a Table of the individual *d* values and their η^2 equivalents

Pair	Diff.]	<i>d</i>	η^2
2-3	3.20	1.176	0.263
2-4	2.90	1.066	0.227
2-1	2.40	0.882	0.167
3-5	2.47	0.908	0.175
4-5	2.17	0.800	0.142

Dr. Jack Barnette Eval Inst 2008 - Effect Size 150

ANOVA Example, K=5

Eta-Squared

$$\eta^2 = \frac{SS_{Between}}{SS_{Total}} = \frac{236.907}{1309.874} = 0.1809$$

Chance predicted values from Barnette/McLean for η^2 :

$$K=5, n=30 \quad \eta^2_{chance} = 0.8408 n^{-1.0113} = 0.0270$$

Dr. Jack Barnette Eval Inst 2008 - Effect Size 151

Confidence Interval for η^2

Use Smithson's NoncF2.sas macro

Input:

F	df1	df2	conf
8.004	4	145	0.95

Output: (NonCentral F limits)

Obs	F	df1	df2	conf	prlow	prupp	ncplow	ncpupp
1	8.004	4	145	0.95	0.975	0.025	10.4157	56.0627

η^2	Lower	Upper
rsq	rsqlow	rsqupp
0.18087	0.064929	0.27207

Verifies value of η^2 found above
 $CI_{0.95}$: 0.0649 to 0.2721

Dr. Jack Barnette Eval Inst 2008 - Effect Size 152

η^2 Confidence Interval

Notice how large this CI is?

$$\eta^2 CI_{0.95} : 0.0649 \text{ to } 0.2721$$

Where is commonly used standard?
Are you surprised?

Dr. Jack Barnette Eval Inst 2008 - Effect Size 153

ANOVA Example, $K=5$

Use Omega-Squared if Ind. Var. is Fixed

$$\omega^2 = \frac{SS_{Between} - (k-1)MS_{Error}}{SS_{Total} - MS_{Error}} = \frac{236907 - (5-1)7.400}{1309873 - 7.400} = 0.1592$$

15.92% of the Variance Accounted for

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

154

ANOVA Example, $K=5$

Use Intraclass Correlation if Ind. Var. is Random

$$ICC = \frac{MS_{Treatment} - MS_{Error}}{MS_{Treatment} + (n-1)MS_{Error}} = \frac{59.227 - 7.400}{59.227 + (30-1)7.400} = 0.1893$$

18.93% of the Variance Accounted for

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

155

Another Alternative in this Case

This example had a group called a "Control"
We could have conducted a post-hoc follow-up using Dunnett's Test and used Glass's g' with those mean differences that were significant. We'd need to figure out how to convert those to variance accounted for statistics and I haven't seen an equation that does that, but I'm sure it could be derived

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

156

Factorial Designs

A factorial design permits us to look at how combinations of treatments might affect mean scores of groups

Of primary importance is finding out if combinations of treatments affect the mean scores of subjects exposed to the combinations

Not going to illustrate this in this workshop, but will in the AEA Pre-session

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

157

Randomized Block Designs or Repeated Measures Designs

What if we had a randomized block (with one subject per cell) or repeated measures design (with more than one subject per cell).

Will not illustrate these in this session, but will in the AEA Pre-session

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

158

Chi-Square Example

Contingency Table (Frequencies), $N=143$

	Poor	Fair	Good	Excellent
Program A	4	6	17	11
Program B	2	3	28	26
Program C	7	12	15	12

$$\chi^2_{\text{obs}} = 16.221, df = 6, p = 0.013$$

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

159

Chi-Square Example

Cramer's V (Variance Accounted for)

$$V = \frac{\chi^2}{Nq} \quad \text{where } q = \min \{r-1, c-1\}$$

$$V = \frac{\chi^2}{Nq} = \frac{16.221}{143 * 2} = 0.0567$$

5.67% of variance accounted for

Dr. Jack Barnette Eval Inst 2008 - Effect Size 160

Confidence Interval for Cramer's V

Use Smithson's Noncchi.sas macro to find upper and lower limits for the NonCentral Chi-square

Input: chi df conf
 16.221 6 0.95

Output:

Obs	chi	df	conf	prlow	prupp
1	16.221	6	0.95	0.975	0.025

Lower and Upper Limits of NonCentral χ^2

ncplow	ncpupp
0.77087	29.7175

Dr. Jack Barnette Eval Inst 2008 - Effect Size 161

Confidence Interval for Cramer's V

Then use this output in Smithson's CramersV.sas macro to find Cramer's V confidence interval

Input:

Obs	chi	df	conf	prlow	prupp
1	16.221	6	0.95	0.975	0.025

ncplow	ncpupp	samp	rows	cols
0.77087	26.7175	143	3	4

Output:

Obs	chi	df	conf	ncplow	ncpupp
1	16.221	6	0.95	0.77087	26.7175

samp	rows	cols	Vsq	Vsqlow	Vsqupp
143	3	4	0.056717	0.023674	0.11440

Dr. Jack Barnette Eval Inst 2008 - Effect Size 162

Confidence Interval for Cramer's V

Best estimate of Cramer's V is 0.0567, with
a $CI_{0.95}$: 0.0237 to 0.1144

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

163

Multiple Dimension Chi-Square

To use Cramer's V , you would need to break
down the structure into all the possible
two-dimension tables and compute
Cramer's V for each one, just like we do in
the other $K > 2$ situations

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

164

Summary of Some Findings

We can compute values for the commonly
used effect size and association measures
We can also use software to find confidence
intervals in most cases for d , η^2 , and V
We observe that these confidence intervals
are very wide, some (including me) would
say almost too wide to be useful

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

165

AERA 2006 Paper that Discusses This

Barnette, J. J. & McLean, J. E. (2006).
Confidence intervals of common effect
sizes: What are they good for? San
Francisco: CA, Annual Meeting of the
American Educational Research
Association.

Available upon request

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

166

Session 4

Relationships Among the
Common Measures

Some Common Standards

- Cohen provided the following:
- $d= 0.2$ was related to η^2 of 0.0099 (~1%)
- $d= 0.5$ was related to η^2 of 0.0588 (~6%)
- $d= 0.8$ was related to η^2 of 0.1379 (~14%)
- Kirk (1995) indicated that these same values can be used to link ω^2 with d . I don't think so.

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

168

Some Common Standards

- Oh, that life could be so simple.
- But it's not!
- This relationship is only true when $K= 2$, and sample sizes are large (100+)

Dr. Jack Barnette Eval Inst 2008 - Effect Size 169

Relationships

When $K= 2$, it is very straightforward to convert any one of these to all of the others if the two sample sizes are known:

- t statistic
- p value
- d
- η^2
- ω^2
- ICC

These are all directly related, may not be linear, but they are directly related (exact predictability)

Dr. Jack Barnette Eval Inst 2008 - Effect Size 170

Converting Values

As indicated, if we know the sample sizes of two groups where we want to compare the means, we can convert one effect size to other values

There is a perception that the p -value and Cohen's d are independent, that they represent different things

In fact, they only represent values on a different metric, but they are exactly related

Dr. Jack Barnette Eval Inst 2008 - Effect Size 171

d as Related to Others when $K > 2$

- The relationship of *d* to the others when there are more than two groups is not specific and there are no specific equations to relate *d* to the other measures
- Thus, there is no way to specifically convert *d* to the others or the others to *d*
- Very limiting for meta-analysis if there needs to be a common metric
- Those equations that seem to work DO NOT
- A few examples follow

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

172

Relationship Summaries

- When $K = 2$, all four are related, not linear, but specifically related
- When $K > 2$ relationships of *d* with the other three are not specific
- When $K > 2$ relationships of Eta-squared, Omega-squared, and ICC are linear and predictable without error ($R^2 \sim 1$)

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

173

Converting Values

If you come to the AEA Pre-session Workshop, I'll provide a detailed description and demonstration of the relationships among the effect size/association measures
I'll just summarize the relationships here

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

174

CAUTION

- When you read that d is related to any of the other three measures and these are used to convert d to r^2 or convert r^2 to d be very cautious.
- Clearly these commonly used equations are useful ONLY when $K= 2$ and there are reasonably large groups.
- You will find this relationship used inappropriately in many cases of meta analysis and other applications.

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

175

A Few Summary Comments

- Effect sizes and strength of association measures provide useful additional information about differences on the metrics of standard deviation or proportion of variance accounted for
- These will be useful in reporting evaluation results as well as research results
- Need to be very careful about converting one measure to another, especially when $K > 2$

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

176

A few Summary Comments

- Confidence intervals around difference estimates remain a useful tool for examining the magnitude of difference
- However, confidence intervals around effect sizes and strength of association measures are relatively large and do not seem to provide much useful information, they appear much less useful than the confidence interval around the difference estimate

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

177

Participant Discussion

Are there any evaluation situations related to the use of effect size and association measures that you'd like to discuss?

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

178

Contact Information

J. Jackson Barnette, PhD
Professor of Biostatistics
School of Public Health
University of Alabama at Birmingham
barnette @ uab.edu
(205) 934 4552

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

179

Want More???

- AEA day-long workshop will probably be conducted at the AEA meeting in Denver in Nov.
- Contact me if you have questions or if I can provide additional information in your use of these measures
- Now to the Workshop Evaluation

Dr. Jack Barnette

Eval Inst 2008 - Effect Size

180
