Reporting Effect Sizes from Published Research: Lessons Learned

Mary Terzian, PhD, MSW
Vanessa Harbin, MPP

Third Annual Teenage Pregnancy Prevention Conference
May 21, 2013

Background

• HHS Teen Pregnancy Prevention Evidence Review
  – Systematic review of programs that impact teen pregnancy, STIs, and associated risk behaviors
  – Conducted for HHS by Mathematica and Child Trends

• Progress to date:
  – Initial findings released in spring 2010 in conjunction with TPP and PREP grant announcements
  – Updated findings released in spring 2012
  – New call for studies released in February 2013
Background

• Over 1,900 citations identified through literature search
• 200 studies met inclusion criteria, 88 studies received a high or moderate study rating
• 31 program models with statistically significant impacts on:
  – Sexual activity (initiation, frequency, number of partners)
  – Safe or unsafe sex (contraceptive use, unprotected sex)
  – Sexually transmitted infections (STIs); pregnancies; or births
• No effect size criteria applied in review
• Going forward, effect size will be reported as part of the review
Goals of this session are to:

- Familiarize audience with effect sizes and how to interpret them
- Provide an overview of our effect size report work and share some general findings.
- Summarize lessons learned and offer recommendations for calculating and reporting effect size.

**Presentation at a Glance**

Part 1: Effect Size 101

Part 2: Overview of PPRER Effect Size Report work and Findings

Part 3: Summary of Lessons Learned and Recommendations
Part 1: Effect Size 101
What *is* an effect size?

• **An effect size is defined as:** “A type of quantitative representation [or description] of the *magnitude* of relations, differences, or comparisons that are in some way meaningful in the research design to which they are applied.” (Dr. Larry Hedges)

• **Effect size helps quantify the magnitude of an impact**
  – In the PPRER context, it’s about *how much* difference there was between groups (treatment and control, comparison groups)
Why are effect sizes so important?

• Effect sizes (in program evaluation context) tell us *how much* a program may be able to change an outcome.
  – Whereas statistical significance only tells us *whether* a program changes an outcome

• *Standardized effect sizes:*
  – Offer decision-makers a way to assess program effects across a variety of outcome measures
  – Are increasingly being requested by publication editors, funders, and policymakers
Ways of describing magnitude of an effect

• Effect sizes can be described in original, unstandardized units or in standardized units.
  – **Unstandardized effect sizes** - Simple to interpret and easy to calculate, but cannot be compared across different measures of the same outcome.
  – **Standardized effect sizes** - Less intuitive because they are expressed in statistical units (such as standard deviations or a ratio of odds), but can be compared across different outcomes or different measures of the same outcome.
Outcome measures and associated ES “ingredients”

• Outcome measures can be dichotomous (y/n) or continuous; therefore each are associated with different ES “ingredients”
  – Raw or adjusted proportions are produced by dichotomous measures (e.g., “Have You Ever Had Sex?”).
    • Example: At post-test, 30% of intervention group has had sex and 40% of control group has had sex.
  – Raw or adjusted means are produced by continuous measures (e.g., “How many sexual partners have you had in past month?”).
    • Example: At post-test, the intervention group has had sex with an average of 1 partner and the control group has had sex with an average of 1.4 partners.
Dichotomous measures and associated effect sizes

- Commonly reported ESs for dichotomous measures are percentage point differences and odds ratios
  - **Percentage point difference** is the proportion of the intervention group having the event of interest minus the proportion of the control group.
    - *Example*: 30% of intervention group reported ever having sex and 40% of the control group reported ever having sex; so, the percentage point difference = -10.
  - **Odds Ratio (OR)** is the odds for the intervention group having an event divided by the odds for the control group.
    - *Example*: The odds of the intervention group having sex = .43 (30/70) and the odds of the control group having sex = 0.67 (40/60); so the Odds Ratio, or OR = 0.64 (0.43/0.67)
Continuous measures and associated effect sizes

• Commonly reported ESs for continuous measures are simple mean differences and standardized mean differences
  
  – **Simple Mean Difference** is simply a difference between raw or regression-adjusted means.
    
    • *Example*: If the post-test mean of the intervention group = 1, and the post-test mean of the control group = 1.4, then the mean difference = -0.4.
  
  – **Standardized Mean Difference** is simply a difference between raw or regression-adjusted means, which has been translated into standardized units.
    
    • *Example*: This is most commonly expressed as Cohen's $d$: a standardized mean difference that is typically calculated by dividing the mean difference by the pooled standard deviation (but can be calculated using other “ingredients” if necessary).
Example of a Cohen’s $d$ equal to 1.0

Control group mean = 0

Intervention group mean = 1.0
Conventions for describing the strength of an ES

• **Cohen's d**
  - The traditional convention for Cohen's d:
    • Very Small: \( d \leq 0.20 \) and Small = 0.20 to 0.50; Moderate: \( d = 0.50 \) to 0.80 and Large: \( d = \geq 0.80 \).
      - Note: There is much uncertainty around this convention and most recommend that the practical value of a Cohen's d should be considered in context and not only using the metric above.

• **Odds Ratio**
  - Suggested rules of thumb to describe the strength of an OR exist, however they are not widely agreed on or reported.
    • OR = 1 (no difference); OR > 1 (treatment group odds higher); OR <1 (treatment group odds lower).
Interpreting ES

• Don’t just rely on the magnitude, consider the context
  – *Intervention context*: dosage, program type, program specificity
  – *Measurement context*: shorter-term vs. a longer-term follow-up point
  – *Outcome context*: effect sizes for some outcomes smaller than others
Summary Points

- Effect sizes are useful because they convey the magnitude of a program effect and have strong utility to policy and practice.
- Standardized effect sizes are useful because they allow a comparison of effect sizes for different outcome measures.
- Commonly reported standardized effect sizes are Cohen’s D and Odds Ratio.
- Standardized effect size can be calculated using different “ingredients”.
- Standardized effect size interpretation depends on context.
Part 2:
ES Report Work
Effect Size Review: Goal, objectives, and critical tasks

• **Goal:**
  – To provide more context on magnitude of effects for studies reviewed

• **Objectives**
  – Produce easy-to-understand report summarizing both unstandardized and standardized effect size information
  – Update program summary reports with unstandardized effect size information

• **Critical tasks**
  – Understand different methods of calculating effect size
  – Collect effect size “ingredients” from all 35 evaluations of 31 programs for all outcomes meeting moderate or high study rigor
  – Calculate standardized and unstandardized effect sizes for every effect
  – Report effect sizes in a way that is meaningful to a general audience
  – Continue to calculate ES for other programs as time/resources allow
Figure 1: Number of outcomes reviewed

- Started with 35 studies/publications (corresponding to 31 programs)
- Identified 265 unique outcomes
- Authors calculated std. ES for 61 outcomes (23%)
- Child Trends calculated std. ES for 77 outcomes (29%)
- Not enough information for 127 outcomes (48%)
Findings for review of outcomes

• Most outcomes were dichotomous
  – 150 were dichotomous (57%) and 115 were continuous (43%)

• Most outcomes were non-significant
  – 111 were significant (42%) and 154 were non-significant (58%)

• If ES is only reported for significant outcomes, you’re only telling part of the story
  – Just because something is not significant, does not mean the effect should be ignored
  – Important particularly in comparing across studies
Findings for review of effect sizes (1 of 2)

• Standardized effect sizes most often not reported for every outcome
  – Out of 35 studies, only 5 (14.3%) provided standardized ES for every outcome (significant and non-significant)

• Several different kinds of standardized effect size were reported
  – Odds Ratio, Risk Ratio, Ratio of Adjusted Means, Relative Risk, Event Rate Ratio, Cohen’s d
Findings for review of effect sizes (2 of 2)

- When standardized effect sizes not reported, there were often missing “ingredients”
  - For continuous outcomes, authors often left out information for:
    - Raw means or regression-adjusted means (missing for 70%)
    - Standard deviations
    - Sample sizes of intervention and control group
  - For dichotomous outcomes, authors often left out information for:
    - Raw or regression-adjusted proportions (missing for 75%)
Part 3: Summary of Lessons Learned and Recommendations
Summary Points (related to collecting ES information)

• Publications often do not provide effect sizes, but when they do provide effect sizes, they often:
  – Provide little or no information about calculation methods or formulas
  – Report ES for some outcomes but not others, and often only for the significant outcomes
  – Fail to discuss the practical significance of the ES magnitude

• Publications often neglect to report all ES “ingredients”
  – When ingredients are missing (e.g., regression-adjusted means or proportions and analytic sample sizes for the intervention and control groups), we cannot calculate the ES.
  – When ingredients are provided, it is sometimes difficult to discern what they are (e.g., if means or proportions provided are raw or regression-adjusted)
Recommendations based on lessons learned

• When calculating Cohen’s d, think about the “ingredients” in the numerator:
  – Regression-adjusted means are preferable to raw means, but there are cases where raw means are appropriate.
    • Develop a protocol for selecting either adjusted or raw means and communicate this in your report.
  – In evaluations with two or more measurements, your numerator can represent change-over-time difference or a point-in-time difference
    • Decide which parameter is most appropriate to communicate and make sure to note this in your report.

• If calculating ES for growth over time or for a study design with clustering, you must use a different formula (see next slide)
Recommendations based on lessons learned

- When calculating Cohen's $d$, be aware of adjustments in the formula for different methods:
  - In particular, use a modified ES formula when the program effect was estimated using the following analytic methods:
    - Cross-sectional multilevel models (e.g., HLM)
    - Generalized Estimating Equations or GEE models
    - Repeated measures ANOVA or ANCOVA
    - Growth curve models

*See handout for a list of specific formulas generated from the available literature.
Recommendations based on lessons learned

• When reporting ES information:
  – *Be thorough*
    • Calculate unstandardized and standardized effect sizes
    • Make sure to include all necessary “ingredients” so that readers can replicate your method
  – *Be transparent*
    • Include information about how effect sizes were calculated and protocol used for selecting effect size “ingredients”
  – *Be helpful*
    • Place the magnitude of the effect in context and, to the best of your ability, describe the practical significance
Questions?
For links to more information:

- **Review Findings:**

- **Review Protocol 2.0:**
Contact Information

For presentation:
Mary Terzian
mterzian@childtrends.org
240.223.9209

For pregnancy prevention evidence review:
PPRER@mathematica-mpr.com

“Educational or instructional materials referenced during presentations at the Ready, Set, Sustain: Continuing Our Success conference are for informational purposes only. Presenters' references to these materials do not constitute endorsement by the Office of Adolescent Health or U.S. Department of Health and Human Services. Any statements expressed are those of the presenters and do not necessarily reflect the views of the Department.”