Paper 3: Issues in calculating average effect sizes in meta-analyses

Dr. Rebecca Maynard, Nianbo Dong, and Irma Perez-Johnson
University of Pennsylvania

This session focused on the rationale for calculating average effect sizes, described which measures should be averaged, and discussed ways to create a meaningful average.

Basics of Computing Average Effect Sizes

There are several reasons to compute average effect sizes: (a) to provide a summary estimate of multi-dimensional outcomes especially in areas like education and child care (Do programs improve economic well-being, where the study reports findings for multiple indicators of cognitive well-being?); (b) to generate global estimates of impacts across population subgroups or replications (Do programs work across multiple settings or for different population groups, where results were reported out at the subgroup rather than the aggregate level?); and (c) to support a global statement of intervention effectiveness (Do home visitor programs improve outcomes for teen mothers, where findings are reported for several distinct outcomes?).

Researching the benefits of preschool provides an illustration of potential reasons to average effect sizes. Often times, studies begin with narrowly defined measures of outcomes (e.g., assessing a particular reading, math, or social skill). Then, researchers move on to examine more broadly defined outcomes, such as overall reading skills (which include elements of vocabulary, decoding, and phonemic awareness, for example). Next, researchers may move beyond examining intervention effects on reading to consider impacts on overall academic achievement. Finally, researchers may evaluate impacts over a range of academic skills, social skills, and physical development. The questions of greatest interest to policy makers and/or practitioners often pertain to the more aggregated measures—within and across studies. What are the outcomes for particular outcome domains and, within a domain, what are the outcomes across studies?

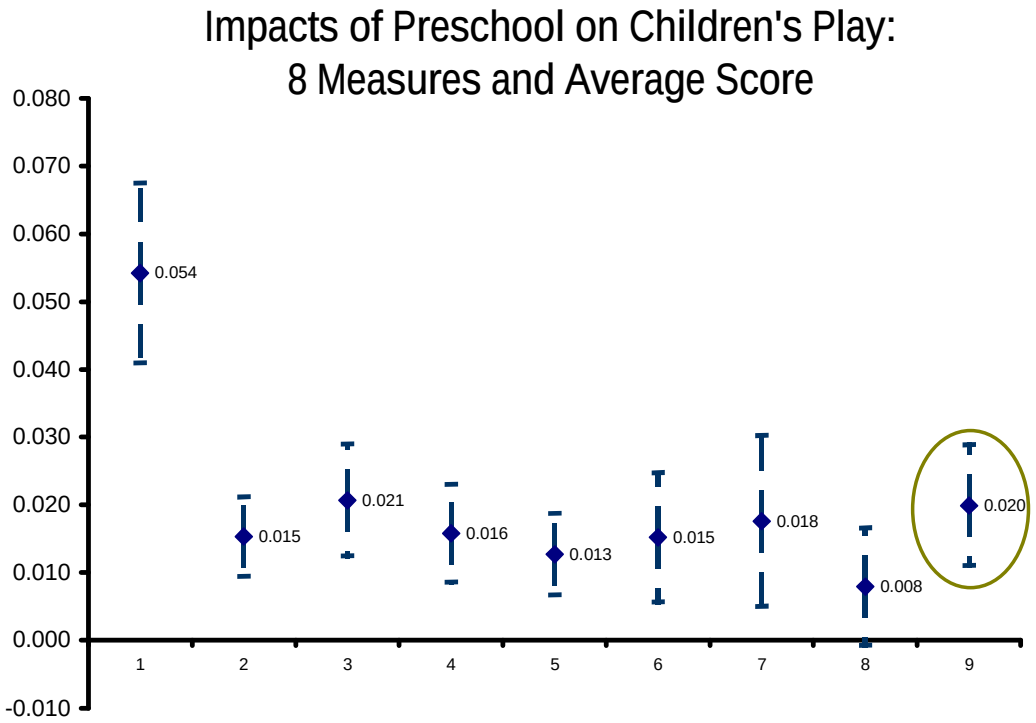
Computing average effect sizes is quite straightforward, once a decision is made regarding the reporting units and assuming the necessary data are available. However, making the necessary decisions entails considerable judgment and, sometimes, important unknowns, and often times important data are missing.

Natural or standardized units. The first issue confronted by the researcher who is presented with impact estimates for multiple outcomes is whether it would be better to report outcomes in "natural units" (e.g., dollars, weeks worked, or ITR test scores) or standardized units (standard deviation units). In many cases, natural units are easier to interpret and no more difficult to average. Even in cases where natural units are easier to interpret, it may desirable to convert to standardized means and mean differences if studies have used different measurement tools (for example, standardized tests normed to different populations).

Computing effect sizes and standard errors for individual estimates. In order to compute standardized effect sizes it is necessary to have sample means (adjusted for covariates, when available) and unadjusted standard deviations of the outcome measures for the control group or for the pooled study sample. (Statisticians do not agree about whether the control group standard

deviation or the pooled standard deviation is preferred. However, in cases of well-matched comparison groups, the choice will not matter much. Conceptually, you would prefer the population standard deviation, anyway.) The standard error of the standardized effect size should be computed using a pooled variance estimate that takes account of sample allocation to treatment condition and that is adjusted (where available) for covariates.

Deciding what to average. The decision regarding what to average will depend on the goal of the analysis as well as the statistical properties of the data. It is largely a conceptual issue to decide what outcomes should be averaged within a study—what measures fall into a common outcome domain—for example, academic skills versus school engagement. A decision to average effect sizes across studies, in contrast, should be driven by conceptual factors (Are the studies addressing common or different interventions, for example?) and by statistical realities—Is there evidence of heterogeneity among the study findings? Are the samples independent?



Source: ECLS-K, unweighted independent samples

Example 1: multiple estimates for related outcomes within a study. The following example presents multiples estimates for related outcomes within a single study using eight items on parent-children play activities. These are point estimates of mean differences in each of the eight measures for children who did and did not attend preschool. All but one of the eight mean differences is statistically significant (as indicated by the confidence interval crossing the horizontal axis). In this particular case, the researchers designed these eight items as a coherent set to measure different aspects of parent-child play. Thus, it probably makes sense to use some averaging of effect sizes and then analyze the aggregate.

In this example, the average effect size is computed as the simple average of the individual effect sizes and the confidence interval around that mean effect size is computed as the average effect size, plus and minus 1.96 * the average standard error for the individual estimates. (A similar approach to pooling results might also be appropriate in cases where multiple studies of an issue have been conducted using the same sample and data—for example, the case of multiple studies of the effects on preschool readiness all using the ECLS-K data.)

Example 2: Interpreting standardized measures. Poor and nonpoor children in the ECLS-K sample exhibit different degrees of variability in their test performance on IRT scaled tests. As seen in the table below, while the mean difference in test scores between children who do and do not attend preschool is larger for nonpoor than for poor children (3.63 versus 2.50). However, when expressed in standardized effect sizes computed using the sample standard deviations, as is typical in a Meta analysis, the estimated effect sizes are similar for both groups (.4 standard deviations). This inconsistency occurs because the variance in the test scores is smaller for the subsample of poor than for nonpoor students. Consequently, when the mean differences based on the IRT scores are converted to standardized effect sizes using standard deviations for the study samples, the effect size for the poor students is over-stated and that for non poor students understated. (See the last two columns in the table below).

The benefits of preschool for academic achievement

		Post-test Means			Effect Size Based on:	
	N	Intervention Group	Control Group	Difference	**Sample SD**	Pop SD
Study 1: Nonpoor	11,264	29.59	25.96	3.63	**0.40**	0.42
Study 2: Poor	2,250	23.78	21.28	2.50	**0.40**	0.29
Pooled Effect (micro data)	**13,514**	**29.08**	**24.88**	**4.20**	**0.49**	**0.49**

Source: Unweighted tabulations of the ECLS-K data.

Example 3: Average effects over multiple studies. The table below presents estimates of the benefits of preschool for non-overlapping samples of children—those from nonpoor and those from poor families. In this case, the average effect size can be computed in one of two ways—by computing a simple average of the effects, or by weighting inverse of the squared standard error of the effect size as the weight, as illustrated in the example below.

The benefits of preschool

	N	ES	SE_{ES}	W	ES*W	t
Study 1: Nonpoor	11,264	0.402	0.019	2912.9	1172.2	21.7
Study 2: Poor	2,250	0.403	0.054	346.2	139.4	7.5
Pooled Effect	**13,514**	**0.402**	**0.018**	**3259.0**	**1311.6**	**23.0**

Source: Split sample of ECLS-K, unweighted
Note: Pooled effects computed using formulas in Lipsey and Wilson, Practical Meta Analysis (2001). See also below.

(The simple average would be more commonly used in averaging results across programs or test sites.) This method of combining effect sizes can be applied to outcomes reported in natural units or in standardized units. In neither case will the results will necessarily be consistent with those derived from micro data. For example, in this example, the micro level data yield an average standardized mean difference of .47 with a t-statistic of 27. In part this is due to the fact that the standard deviations of the outcome measures differ substantially between the two samples and, in part, it is because the intervention and control groups are very unequal in sizes.

It is important to note that the above method of pooling results across samples should is not valid in cases where samples are overlapping. For example, one set of researchers might use a sample to report findings of the effects of childcare for children from poor families, a second set to report the effects for low-performers, and a third set to report the effects for high achievers when they went into preschool. Such examples of results from overlapping samples from the same data set are common in areas where much of the research is based on national public use datasets.

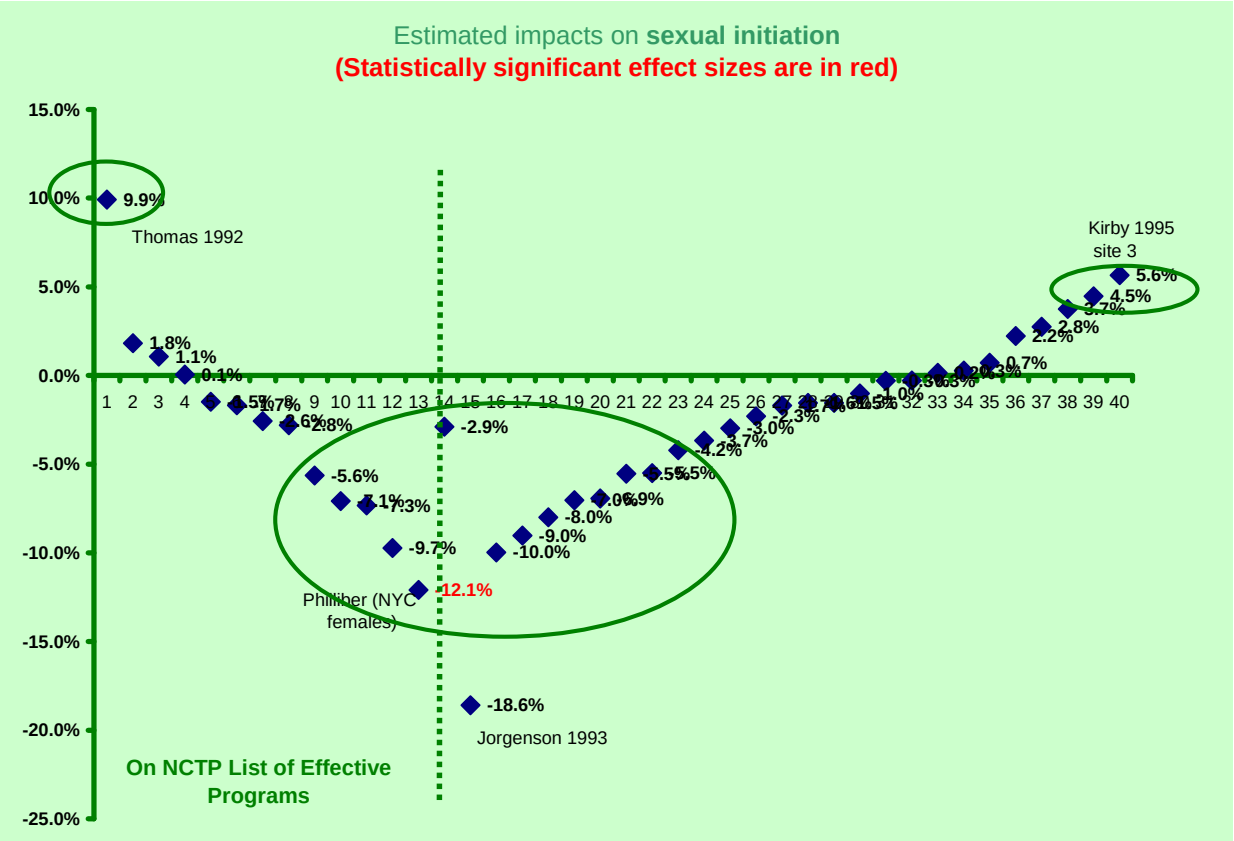
Example 4: Pooling results across multiple trials. The third example illustrates how effect sizes can be combined across multiple, studies. Before averaging of effects across studies, it is important to address four questions: (1) what specific outcome measures would have policy relevance; (2) what types of interventions are sufficiently similar to warrant averaging effect sizes across them; (3) will policy makers and/or practitioners find it useful to have effect sizes averaged across specific population groups (for example, males and females or middle school and high school youth)?; and (4) what is available the universe of *independent* effect sizes?

The example presents the results impacts of teen pregnancy prevention programs. The review presents findings at three levels—(1) the estimated effects of programs on each of three distinct outcomes (sexual activity; behavior that risks pregnancy; and pregnancy) for specific study samples, regardless of the study sample, program type, or setting; (2) average effect sizes across programs of a particular type; and (3) average effect sizes across all program, regardless of type.

The first graph below shows the individual program effects on sexual initiation (measured in natural units, since percentage point differences are easier to interpret than are odds ratios). Each blue diamond represents one of the 39 estimated impacts on sexual initiation from one of the 20

randomized control studies included in the review. The chart is divided in two by the dotted green line. The left side of the chart shows studies on a list of "effective programs" published by the National Campaign to Prevent Teen Pregnancy. All of the point estimates on the right side are programs that were evaluated, but are not on the list. Effect sizes in red are statistically significant; others are not. The three most notable features of these results are (1) the impact estimates vary considerably in size; (2) most are not statistically significant; and (3) there is not obvious difference in the pattern of results for the programs selected for the "effective program list" and those not. (The effective program list included studies that showed statistically significant, favorable effect sizes for any of a wide number of outcomes and it included studies that reported upwardly biased significance levels.)

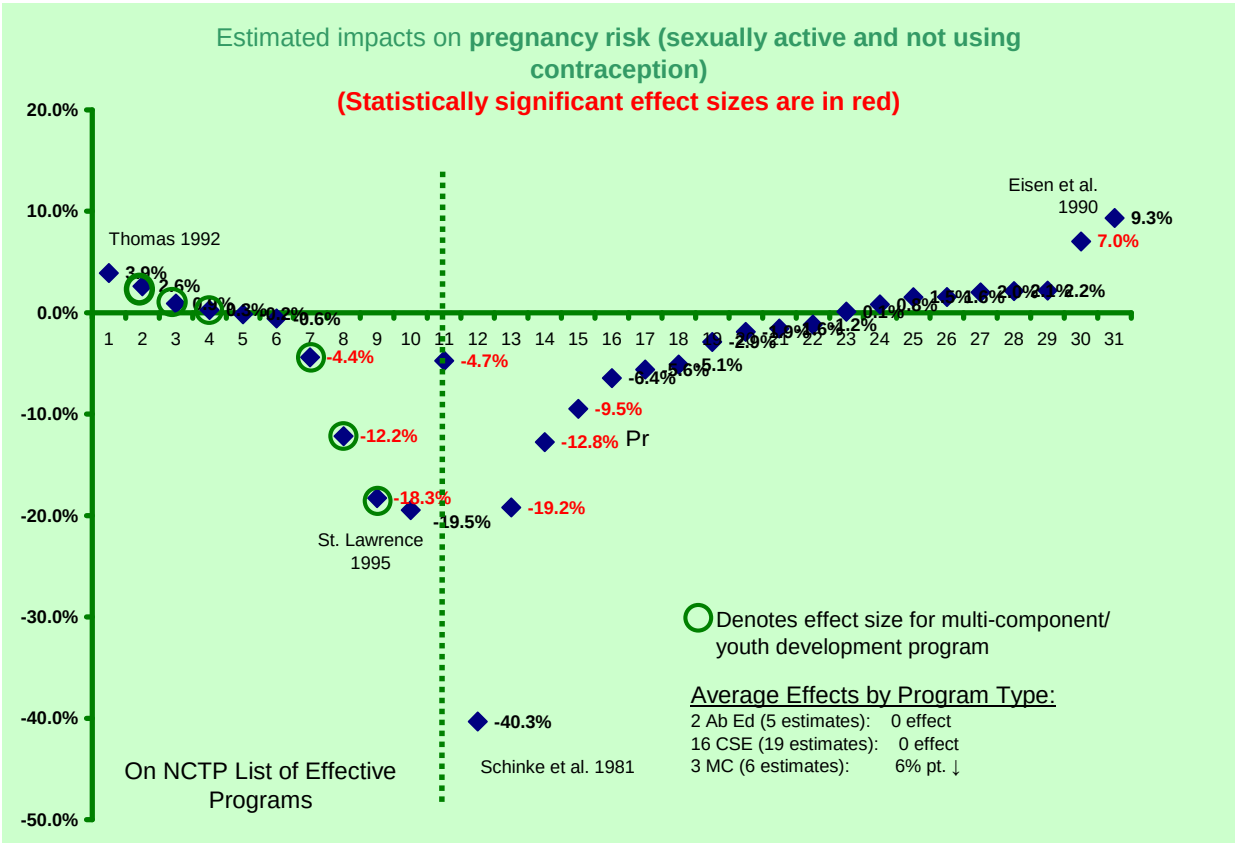
What does averaging mean in a case like this? In this case, the average overall effect size is -1.2 percentage points with a confidence interval ranging from +.4 to -2.8 percentage points. One could reasonably question the decision to combine all of these effect sizes on a single graph as well as any decision to create an overall average impact, given that the underlying programs varied substantially in their strategies, their targeting, and their per person cost. However, in this case, seemingly similar patterns of estimated effect sizes (some positive and some negative) most with relatively large confidence intervals was observed across all program types. Average impacts by program type ranged between +1 percentage point and -2 percentage points for the three main program types—abstinence focused programs, sex education with a contraception component, and multi-component/youth development programs. Importantly, had the research reported only the average effect sizes, it would have been easy to argue that the results could be masking evidence of effectiveness for particular types of programs. So, presenting the disaggregated findings is important for transparency.



Source: Scher et al. (2006), Campbell Collaboration
http://www.campbellcollaboration.org/frontend2.asp?ID=87

The next graph highlights the estimated impacts on pregnancy risk from multi-component programs—those effect sizes circled in green. In this example, there are more statistically significant differences (denoted in red) than were observed for sexual activity, and most of the significant findings (all but one) are in favorable directions. However, there is also one very large statistically significant difference favoring the control condition.

Using only the average effect size estimate for pregnancy outcomes would mask not only the variability in the point estimates, but also the variability in the point estimates across different types of programs. Hidden in this chart and in the average are three important facts: (1) two Abstinence Education (AbEd) programs (5 estimates) have zero effect; (2) 16 Comprehensive Sex Education (CSE) programs (19 estimates) have zero effect; and (3) 3 Multi-Component (MC) programs (6 estimates) show a fairly sizable favorable average effect (6 percentage points).



Source: Scher et al. (2006), Campbell Collaboration
http://www.campbellcollaboration.org/frontend2.asp?ID=87

Applying Standard Formulas to Compute Average Effect Sizes

The following table illustrates average (natural unit) effect sizes for the three key outcomes examined in the review of the teen pregnancy prevention research. Over all studies measuring each of the particular outcomes, the average effect size is small and not significantly different from zero as evidenced by the fact that the half-confidence intervals are larger than the effect sizes.

Small differences; None statistically significant

Outcome	# of studies & independent estimates	Sample size	Mean Outcomes: Intervention group	Mean Outcomes: Control group	Estimated Impacts: Difference in means	Estimated Impacts: 1/2 CI
Sexual Experience	21 studies; 40 estimates	37,705	37.9%	39.1%	-1.2%	+/- 1.6%
Pregnancy Risk	24 studies; 34 estimates	33,405	13.7%	15.0%	-1.3%	+/- 1.7%
Pregnancy	13 studies; 25 estimates	19,012	8.2%	8.6%	-0.4%	+/- 1.1%

These average estimates were generated using the standard formulas used in Meta analysis and summarized in Lipsey and Wilson (2001). Those formulas used for continuous outcomes are as follows:

Compute the effect sizes	Mean Difference/SDV_{pooled}
Compute the weight (w)	$w = \frac{1}{se^2}$
Compute the Weighted Average Effect Size	$\overline{ES} = \frac{\sum (w \times ES)}{\sum w}$
Standard Error of the Average Effect Size	$se_{\overline{ES}} = \sqrt{\frac{1}{\sum w}}$
t-Statistic for the Average Effect Size	$t = \frac{\overline{ES}}{SE_{\overline{ES}}}$

These formulas are easy to apply and yield estimates with reasonable statistical properties under well-defined conditions, However, they do not take account of some of the complexities encountered in the evaluation literature. For example, they do not take account of unbalanced sample designs or of the fact that standard deviations may vary considerably across the study samples. The consequences of the failure to account for these features of the studies was illustrated in the case of effect sizes for IRT test scores across samples of poor and nonpoor children. In addition, the formula for the standard error of the average effect size needs to be adjusted to take account of situations where the effect size estimates have been regression adjusted. It is common to encounter situations where some effect sizes have been regression adjusted and others have not. Applying the formulas above works well in cases where none of the results have been adjusted for covariates. In particular, it is important to "shrink" the standard error of an effect size average to account for the fact that some estimates are more precise that would be estimated based on the sample standard deviation and sample size alone.

Concluding Remarks

There can be real benefits to using standardized mean differences and to averaging effect sizes within and/or across studies. However, there also are limitations. One is that mean differences that have been standardized using sample standard deviations are sensitive to whether or not the samples are representative of similar populations.

In addition to raising a general caution about averaging effect sizes across studies that have been standardized based on standard deviations for heterogeneous samples, there also is a more general caution. Before averaging effect sizes, think carefully about the relevance effect size measures going into the average and the policy and practical significance of the estimates that will be generated.