Paper 2: How measurement of outcomes affects the interpretation and understanding of effect sizes

Margaret Burchinal, University of North Carolina

Effect sizes describe the magnitude of findings from statistical analyses. They provide a common metric and allow for easy comparison of findings across studies and across outcomes. Effect sizes are recognized standards that are not influenced by sample size, but the impact of measurement on estimated effect sizes is too often ignored. This paper describes why effect sizes are effected by measurement precision, briefly discusses measurement precision across age and developmental domain, and describes how different types of effect sizes are effected by measurement precision.

Effect sizes and measurement precision. Classical Test Theory can be use to understand why measurement error impacts effect sizes, Within the context of Classical Test Theory, the score of the i^{th} person on the j^{th} occasion of measurement is expressed as:

$$Y_{ij} = T_{ij} + E_{ij}$$

This score for that individual at that time point (Y_{ij}) can be from a scale, test, or observation, and is composed of two parts—the true score (T_{ij}) and error (E_{ij}). The error values are derived from multiple sources, including individual sources such as lack of attention, reliability of the scale, and validity of the scale. The variability of a measure, Var(Y), is also related to variability in both true scores, Var(T), and error scores Var(E):

$$Var(Y) = Var(T) + Var(E),$$

so that error variance is $\geq$ 1 – reliability. This index of reliability is usually larger than the reported instrument reliability because it includes variability from other "error" sources that can not be or are not directly measured.

Either the variance or standard deviation of a variable (Y) is used in computing most indices of effect size. The standard deviation is linked to the variability of that variable (i.e, it is the square root of the variance in the simplest effect size models). The effect size becomes smaller as the error variance becomes larger. This is because the variance of that measure is larger when the error variance is larger and either the standard deviation (SD) or the variance (Var) is used in the denominator in computing the effect size, as shown below:

Standardized mean difference: $d = \frac{(Mean_1 - Mean_2).}{\text{pooled SD}}$

For a correlation (or partial correlation): $r = \frac{Covariance(X,Y)}{Var(X)\ Var(Y)}$

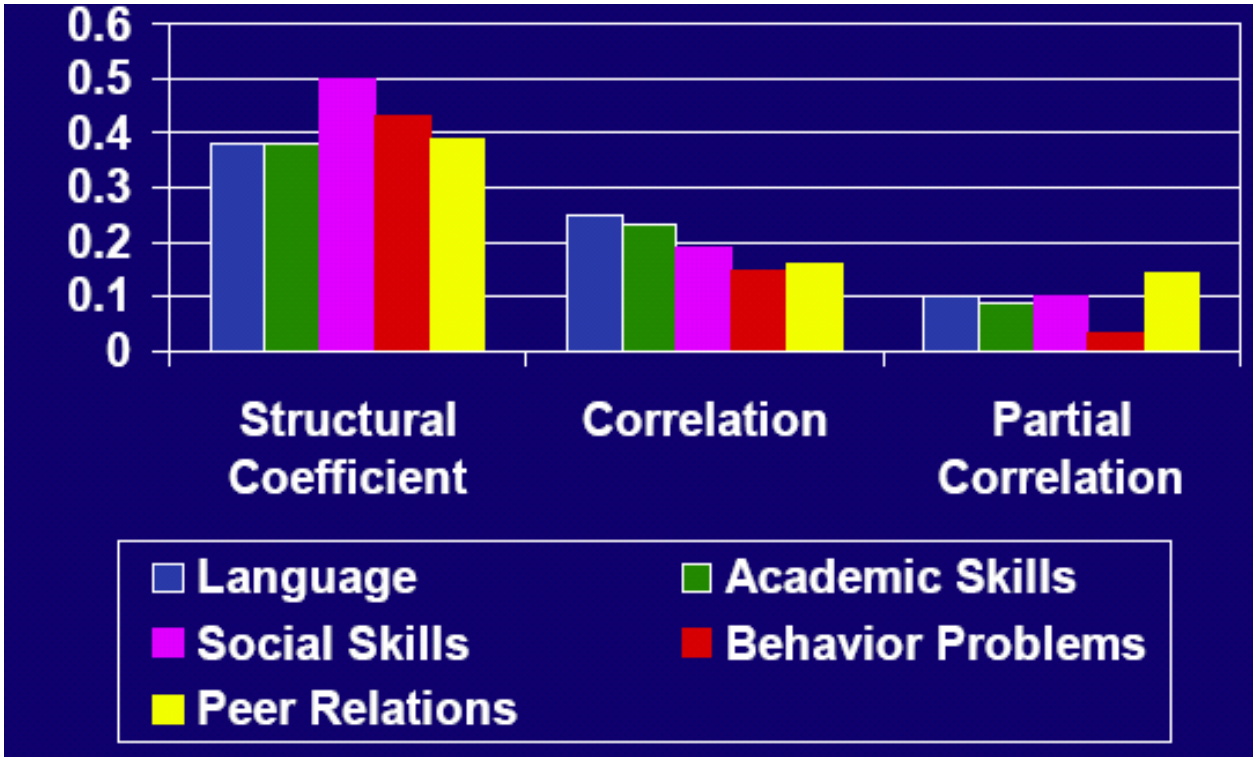
Variance Accounted for: $R^2 = \frac{Var(\text{predicted } Y)}{Var(Y)}$

Odds Ratio is less obvious: $OR = \frac{p_1 / (1-p_1)}{p_2 / (1-p_2)}$

Precision in measurements across age and development domain In general, there are sizable differences in measurement precision. When assessing infants, measures are much less reliable because very young children have a more limited skill repertoire and there is greater variability within the individual. Reliability and validity increase with age—as children get older they tend to become more cooperative and skills become more established. By age 3 to 5, test-retest reliability is relatively good. In addition, there are differences in precision across developmental domains. Cognitive, language, and academic skills can be measured with good reliably due to well-developed standardized tests.

In contrast, measurement in almost all other developmental domains tends to be less precise. Ratings scales or observations are typically used for examining social skills and behavior problems. When using rating scales, there may be reasonable test-retest by age 4, but discrepancies contributing to error variability due to informant bias. When using observations, it is difficult to achieve high reliability in terms of test-retest without observing for extended periods of time, especially when looking at relatively infrequent behaviors such as aggression.

Impact of measurement precision on different types of effect sizes. On the continuum of effect sizes, the most generous calculation is with structural coefficients. This method corrects for estimated error, ignores selection bias, and computes a correlation between variables. Another generous method is zero-order correlation which is unadjusted for covariates and ignores selection bias. A less biased method is partial correlation. This method adjusts for observed covariates and ignores selection bias due to unmeasured variables. The following chart displays the order of child care quality effect sizes based on the computation method.



Source: NICHD Study of Early Child Care

In conclusion, measurement limits our ability to detect associations, and thereby effect sizes. Measurement precision has a similar impact on the estimation of effect sizes and test statistics. When measurement precision is relatively low, moderate or large effect sizes cannot be expected.