A PRIMER ON EFFECT SIZES

Let us get a sense of what an effect size means. There are two common ways to calculate an effect size: first, when two groups are compared—such as comparing a class receiving a literacy program with a similar class not receiving this program—and second, over time—such as comparing the performance of a group of students at the outset and again at the end of a series of literacy instruction. In both cases, the effect size represents the magnitude of the difference—and of course the quality of the comparison, the measuring instruments, and the research design to control extraneous factors are critical.

An effect size of $d = 0.0$ indicates no change in achievement related to the intervention. An effect size of $d = 1.0$ indicates an increase of one standard deviation on the outcome (e.g., reading achievement), a $d = 1.0$ increase is typically associated with advancing children's achievement by two to three years, and this would mean that, on average, the achievement of students receiving the treatment would exceed that of 84% of students not receiving the treatment. Cohen (1988) argued that an effect size of d = 1.0 should be regarded as a large, blatantly obvious, and grossly perceptible difference, and as an example, he referred to the difference between the average IQ of PhD graduates and high school students. Another example is the difference between a person at 5'3" (160 cm) and one at 6'0" (183 cm)—which would be a difference visible to the naked eye.

We do need to be careful about ascribing adjectives such as *small*, *medium*, and *large* to these effect sizes. Cohen (1988), for example, suggested that $d = 0.2$ was small, $d = 0.5$ medium, and $d = 0.8$ large, whereas it is possible to show that when investigating achievement influences in schools, $d = 0.2$ could be considered small, $d = 0.4$ medium, and $d = 0.6$ large (Hattie, 2009). In many cases, this attribution would be reasonable, but there are situations where this would be too simple an interpretation. Consider, for example, the effects of an influence such as behavioral objectives, which has an overall small effect of $d = 0.20$, and reciprocal teaching, which has an overall large effect of $d = 0.74$. It may be that the cost of implementing behavioral objectives is so small that it is worth using them to gain an influence on achievement, albeit small, whereas it might be too expensive to implement reciprocal teaching to gain the larger effect.

The relation between the notions of magnitude and statistical significance is simple: Significance = Effect size × Study size. This should highlight why both aspects are important when making judgments. Effect sizes based on small samples or small numbers of studies may not tell the true story, in the same way that statistical significance based on very large samples may also not tell the true story (for example, a result could be statistically significant but have only a tiny effect size). Similarly, two studies with the same effect sizes can have different implications when their sample sizes vary (we should place more weight on the one based on the larger sample size). The most critical aspect of any study is the convincibility of the story that best explains the data; it is the visible learning story that needs critique or improvement—to what degree is the story in this book convincing to you?

Figure 1.1