What Constitutes Strong Evidence of Program Effectiveness?

By **Sputnik Contributor** on October 8, 2012 9:10 AM

Note: This is a guest post by Jon Baron, President of the Coalition for Evidence-Based Policy, and former Chairman of the National Board for Education Sciences

Bob Slavin's recent blog posts on **Bad Measures** and **Brief, Small, and Artificial Studies** provide a valuable discussion of how evaluation studies - even those using random assignment - can often fall well short of "rigorous." This post seeks to address a related question: what constitutes strong evidence of effectiveness? By strong evidence, I mean evidence that provides confidence that a program would improve important educational outcomes if implemented faithfully in a similar population.

Evidence standards articulated by the **Institute of Education Sciences** (IES), **National Academy of Sciences**, and other respected scientific bodies underscore that strong evidence usually requires well-conducted randomized controlled trials. Whether the findings from these trials constitute strong evidence, however, also depends on factors such as the following.

The studies demonstrate effects on final, policy-important outcomes and not just intermediate outcomes that may or may not lead to final outcomes.

Example - welfare/employment. The Labor Department's New Chance demonstration - a program providing educational, job training, and other services to young welfare mothers - was found in a large randomized trial to produce a sizable increase of 12 percentage points in the mothers' receipt of a GED (an intermediate outcome). By contrast, the study found no effects on employment, earnings,

welfare dependency, or ability to read - i.e., any of the final, policy-important outcomes that were hoped for.

Example - preschool education. HHS's randomized trial of Head Start found a sizable effect on the program goal of increasing preschoolers' ability to identify letters and words (an intermediate outcome), but no significant effects on their actual reading ability or other educational outcomes at the end of first grade (the more final, important outcomes).

The studies show that effects are sustained long enough to constitute meaningful improvement in educational or other key outcomes.

Example - workforce development. The federal Job Corps program - which provides education and job training to disadvantaged youth - was prematurely declared a success based results from a large, randomized trial showing an 8% increase in earnings at the three-year follow-up, and a projection that such effects would continue over time. A later follow-up found that the earnings effects did not persist, and instead had faded to zero at the five-year point and thereafter. As a result, the program's cost was found to greatly exceed its benefits.

The studies show that the effects are sizable (and not just statistically significant).

Example - early childhood home visiting. Healthy Steps - which provides home visiting services to families with a newborn child - was found in a large randomized trial to produce a statistically-significant increase in the percent of mothers bringing their child to a well-child doctor visit at one month of age. However, the effect size was small: 97% of mothers receiving Healthy Steps did a one-month doctor visit, versus 95% of control group mothers. The effect reached statistical significance only because the study had a very large sample - over 2100 mothers. (Large samples are capable of detecting small effects that may not be of practical importance.)

The effects have been replicated across different studies and/or study sites, and in real-world educational settings.

Example - K-12 education. Project CRISS - a teacher professional development program for improving adolescent reading - was considered highly promising based on a small randomized trial that met

What Works Clearinghouse standards, and was therefore selected by IES for a replication trial. The initial study found a large increase in students' reading comprehension as measured on a researcher-designed test (36 percentile points). By contrast, the more definitive, IES-sponsored replication trial - conducted in 38 high-poverty public schools, with a student sample 10 times as large as the initial study - found no effect on reading comprehension as measured on a well-established, standardized test. (This example is one of many in which IES-sponsored trials have overturned initial findings of effectiveness in small randomized trials or quasi-experiments.)

Of the educational programs often described as "evidence-based," only a small subset meet the above conditions for strong evidence (an earlier **post** lists websites where they can be found). Although few in number, these programs - if implemented effectively on a large scale - could produce important gains in areas such as reading achievement, college attendance, and workforce earnings.
Programs with preliminary or moderate - as opposed to strong - evidence, offer substantially less confidence that they produce meaningful effects compared to schools' usual practices. However, in many areas of education - such as dropout prevention - they may be the best option available to education officials, because programs with strong evidence do not yet exist. In these cases, we believe that program implementation should include a rigorous evaluation wherever feasible, to determine whether the program really works and build a more reliable foundation for future decision-making.

-Jon Baron

The Coalition for Evidence-Based Policy is a nonprofit, nonpartisan organization whose mission is to increase government effectiveness through the use of rigorous evidence about "what works."