Statistical Problems with Meta-Analysis

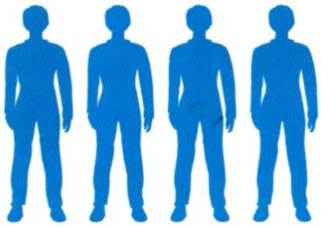
Roy CARR-HILL

We explained in the previous issue that, in order to go beyond 'vote counting' (of the number of artiles or studies or or against an hypothesis), to a comparison of the magnitude of intervention effects on a particular outcome, there must be (i) a common outcome concept or construct, plus (ii) a common scale or metric in which effect sizes are measured, and (iii) data from interventions conducted with relevantly similar samples. It is possible to compare the effects of very different interventions on the same outcome using a common scale (as in Kremer et al 2013) and in practice, policy-makers may legitimately wish to compare the effectiveness of alternative means towards a particular educational end. However, the final stage in meta-analysis - pooling - requires (iv) that there also be a common intervention (a defined intervention-outcome pair). The first two were examined in the previous issue; here we examine the third and fourth assumptions.

Samples and populations

The SR method requires that the studies be drawn from an appropriate sample or population. By standardising effect sizes and placing them along the same scale, the argument is that it is possible to compare any outcome, regardless of diversity in measurement within the original studies. However, a fundamental assumption of this method is that *differences in standard deviations among studies reflect differences in measurement scales, rather than real differences in variability among study populations* (Higgins & Green, 2011). This may be the most difficult assumption to satisfy.

These validity concerns can be illustrated through the use of a simple example (McKenzie). Imagine that an NGO conducted an intervention with a homogeneous group of students. The students took a test before and after the intervention. On the post-test, the average score on the test was a 50%, with a standard deviation of 5%. There was a gain of 1% across the sample. At the same time, the government conducted the same intervention with a heterogeneous group of students. The mean on the post-test with this second group of students was also 50%, but there was a standard deviation of 20%. The gain was 2% across the sample. Calculating the SMD for these two studies gives an SMD of 0.2 for the first study and 0.1 for the second study, meaning that the calculation of the SMD *inflates the effectiveness* of the intervention for the homogeneous group. Although an overly simplistic example, this scenario clearly demonstrates the risks in conducting reviews based solely on a comparison of SMD values.

	Homogeneous Group of Pupils	**Heterogeneous Group of Pupils**
		
	Baseline	
Mean test score	50%	50%
SD	5%	20%
	Endline	
Mean test score	51%	52%
SD	5%	20%
SMD	**0.2**	**0.1**

Differences in the distribution of scores can also be the result of pre-existing differences between study populations (e.g. learning gains within an illiterate population will necessarily be much more dramatic than learning gains within a relatively advanced population, regardless of the effectiveness of the intervention). Thus, the example cited in Kremer et al (2013) of an intervention with a homogeneous population of illiterate girls, showed huge effects because any improvement was many times the standard mean difference.

Either way, the standard deviations of multiple tests taken in multiple contexts are not really 'standard' in any sense of the word (Singh, 2015).

Pooling of Effects

Pooling relies on at least one additional assumption – typically a random effects model is used to combine data and it is assumed that the studies are based on samples which are effectively random draws from the relevant population. But the 'draw' of studies in one of the most well-known meta analyses of de-centralisation contained13 studies, 8 of which were from Mexico.

BUT it also relies on a directly comparable intervention one can compare the effects of two interventions on the same outcome but for pooling we must believe it to be the *same* intervention

Even when outcome measures are directly comparable, interventions frequently are not. In the case of school feeding, for example, the intervention might be considered sufficiently similar across contexts to allow comparison and synthesis of effects in studies with comparable outcomes, but often interventions are more complex and systemic. Reforms such as decentralisation, for example, are inextricably linked with the systems to which they belong; so that 'the same intervention' has only a very broad interpretation, arguably too broad to warrant pooling of studies.

These validity concerns are further compounded when the population of studies is taken into consideration. In addition to assuming comparability of measures and treatments across studies, meta-analysis assumes that the results being pooled are random draws from the population concerned. However, this assumption is also often violated, particularly in international development reviews. Certain kinds of educational interventions are more popular in some contexts than others. As a result, there is a degree of non-randomness in the population of studies available for any review. To take our review as an example, the majority of included studies (12 out of 26) focused on interventions in Latin America. This is unsurprising, given that Latin American countries were amongst the first lower-income contexts to attempt to decentralise their education systems. Making any worldwide generalisations on the basis of a largely Latin American sample is problematic and yet this is the practice that must be adopted in most international development reviews.

First, it requires that the included studies be investigating comparable interventions/ treatments. Although these may vary in intensity or duration, the treatments themselves must be comparable in order to validly combine the data. In order for pooling to be valid, the *treatments* (as well as the outcomes) must be comparable. To refer once again to the origins of the method in medicine, it is clear that most medical reviews combining the results of studies investigating the impact of a particular drug on a particular health condition include studies which assume identical definitions of the health condition, which rely on standardised measures and which are likely to vary only in terms of the characteristics of the sample. However, education reviews rarely meet these conditions and, as such, the required assumptions underpinning the use of meta-analysis are often violated.

In most social sectors, treatments vary widely, depending on context and population. This is certainly the case in education, where a treatment may, in principle, be similar but may differ dramatically in practice. Our recent review is a case in point, as we were tasked with synthesising the literature on school-based decision-making reforms. Although, on paper, all of the studies we examined did consider the impact of moving decision-making authority to the level of the school in a particular context, the particular nature of the reform differed widely across the studies. Some programmes gave decision-making authority to a school committee, comprising both school officials and community members; others gave decision-making authority only to the school itself. Some included financial incentives or training; others did not. Some gave schools both financial and human resource decision-making powers; others only gave financial authority; others still gave schools limited authority over curriculum and pedagogy decisions. It is hard to argue that these interventions are comparable ‘treatments’, and yet meta-analysis requires that we treat them as ‘sufficiently similar’. The validity of such an approach is questionable, particularly when combined with the concerns about incomparability outlined in the previous section. In practice, most education reviews using meta-analysis are forced to pool unlike intervention-outcome pairs, despite the fundamental assumption of comparability underpinning the method.

References

Roy Carr-Hill, R., Rolleston, C., Pherali, T. and Schendel, R. with Peart, E. and Jones E. (2016) The Effects of School-Based Decision Making on Educational Outcomes in Low and Middle Income Contexts: A Systematic Review, International Institute for Impact Evaluation (for DFID)

Kremer, M., Brannen, C. & Glennerster, R. (2013). "The challenge of education and learning in the developing world." *Science* 340(6130): 297-300