

Commensurability of Outcomes Measures in Meta Analyses in Systematic Reviews outside Medical Care

Roy CARR-HILL

Abstract

Systematic reviews are increasingly being carried out in international development research around educational issues. Many such reviews use techniques of meta-analysis to combine the results from several different studies. Where, as is frequent, the included studies don't use precisely the same outcome measure, the various measures are transformed to a common scale using the 'standardised mean difference'. This paper questions shows how this technique is often practiced when the raw outcomes and their measurements are incommensurable and considers the implications for policy-makers and practitioners interested in using review results.

I. Introduction

Meta-analysis, although ignored by many policy analysts, is fundamental to quantitative systematic reviews, but have not themselves been reviewed critically except implicitly in that some of those commissioned to carry out a systematic review do not use those procedures (e.g. Westhorp, 2014).

The purpose of this paper is to provide a partial critique of the statistical approach of the meta-analysis being used in these systematic reviews (partial, because it does not address the *purely* statistical issues involved in standardisation, in weighting and in pooling, which will be considered in the next issue). Instead, the focus here is on the comparability of the outcomes and of the measures used for those outcomes in the studies whose results are then being combined ('pooled') in the systematic reviews.

These systematic reviews involve combining the results from several different studies to calculate ‘effect’ sizes and presenting those on what is called a ‘forest plot’¹. The specific procedure at issue in this paper is the common use of the standardised mean difference measure (in order to calculate ‘effect’ sizes) to make the outcomes of studies using different measurement outcomes appear to be using the same or even equivalent outcomes.

The main example use here is the recent systematic review we (Carr-Hill et al., 2015) carried out including only those studies that focussed on student outcomes, such as student attendance, drop-out and repetition, and a variety of test scores²; but other examples from criminology and health are also included in the next section.

I. II. How it all Started

First we describe the origin of the Cochrane Collaboration and the Campbell Collaboration

II.1 In Pharmaceutical and Medical Research

In pharmaceutical and medical research, both the outcomes and the procedures are usually precisely defined and these are measured using standardised instruments. Although there might be variations in terms of the age, ethnicity or gender composition of the populations used in those studies, there is no or little dispute about the outcomes measurements themselves or about the technical equivalence of the interventions; instead these variations are considered in terms of what is called heterogeneity analysis – a fancy term for examining whether the same effects remain for sub groups defined by age or social class etc. At the same time, Carr-Hill (1996) suggested that the dominance of such procedures and institutions such as the UK’s National Institute for Clinical and Health Excellence drawing mainly on systematic reviews could lead to the emasculation of anthropologists and sociologists from discussions about health service policies; and that is precisely what has happened to a large extent.

¹ The forest plot is a graphical display of estimated results from a number of scientific studies addressing the same question, along with the overall results.^[1] It was developed for use in medical research as a means of graphically representing a [meta-analysis](#) of the results of [randomized controlled trials](#).

² Specifically; Petrosino (), CCT review

These procedures are illustrated in the logo of the Cochrane Collaboration, founded in 1993 by Iain Chalmers, which illustrates a meta-analysis of the data from seven randomized controlled trials (RCTs), comparing one health care treatment with a placebo³ in a forest plot⁴. The logo shows the results of a [systematic review](#) and meta analysis on an inexpensive course of [corticosteroid](#) given to women about to give [birth](#) too early – the evidence on effectiveness that would have been revealed had the available RCTs been reviewed systematically. A reading of the titles of 20 most popular reviews (Box 1) makes it clear that nearly all the reviews are of this type with a clearly defined intervention and outcome; and, where there is ambiguity – e.g. a recent analysis of Interventions for preventing falls in older people living in the community - the review explores those ambiguities in detail.

Box 1: Twenty Most popular Cochrane Evidence Studies

Acupuncture for tension-type headache

Mothers' position during the first stage of labour

Vaccines to prevent influenza in healthy adults

Midwife-led continuity models versus other models of care for childbearing women

Gabapentin for chronic neuropathic pain and fibromyalgia in adults

Whole-body cryotherapy for preventing and treating muscle soreness after exercise

Screening for breast cancer with mammography

Interventions for preventing falls in older people living in the community

Vitamin C for preventing and treating the common cold

T-tube drainage versus no T-tube drainage after open common bile duct exploration

Corticosteroids for bacterial meningitis

Oral misoprostol for induction of labour

³ A placebo is a simulated or otherwise medically ineffectual treatment for a disease or other medical condition intended to deceive the recipient that they are being treated.

⁴

Loop diuretics for patients receiving blood transfusions

St. John's wort for treating depression.

Two different laparoscopic techniques for repairing a hernia in the groin

Exercise for depression

Doppler ultrasound of fetal blood vessels in normal pregnancies

Honey as a topical treatment for acute and chronic wounds

Statins for the primary prevention of cardiovascular disease

Water fluoridation to prevent tooth decay

The main point is that the comparisons are made in terms of precisely defined outcomes, directly linked to established measurement instruments and metrics. But when broader health care interventions are considered such as the possibility of delegation from doctors to nurses, even fervid advocates of Systematic Reviews were unable to find sufficient similarity between the interventions used and the outcomes measured to carry out statistical meta-analysis (Laurent et al., 2005)); UK policy had to rely on a downmarket observational multiple case study (Carr-Hill and Jenkins-Clarke, 2000?).

In order to compare 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). In this paper, we examine the first two assumptions.

II.2 Campbell Collaboration (C2)

The Campbell Collaboration - a sibling organisation to the Cochrane Collaboration - prepares, maintains and promotes the accessibility of systematic reviews in areas such as education, criminal justice, social

policy and social care. It grew out of an exploratory meeting in July 1999 and was founded in 2000.

An example of a C2 systematic review is one completed on so-called “scared straight programs”, in which kids at risk of committing a crime—and sometimes even those who are not at risk—hear from convicted felons who try to deter them from delinquent acts or crimes. These programs received a lot of press, are popular with many parents, and they seem to enjoy some approval among policymakers. The systematic review uncovered 200–300 articles from studies of these programmes. Only a small fraction turned out to be fair tests of the programme considered, and there was no discussion of the comparability of the programs. The nine randomized trials identified in the review showed, contrary to popular belief, that these programs actually enhanced the likelihood that kids would subsequently engage in delinquent behaviours or crimes.

Table 1: Scared Straight Programmes in the US

Institution	Subjects	Measure of Subsequent Offending
Michigan Dept. of Corrections: 1967	60 juveniles NFI tour of reformatory or no tour	Recidivism was measured as a petition in juvenile court for either a new offense or a violation of existing probation order. Large negative effect
Greater Egypt, Illinois, 1979	Mix of delinquents 13-18 yo, and at risk NFI	Participants were compared on their subsequent contact with police, on two personality inventories (Piers-Berne and Jesness). Small negative effect.
Michigan Jolt study 1979	227 delinquents NFI ; 5 hours in correctional facility	Participants were compared on a variety of crime outcomes collected from participating courts at three and six month follow-ups. Small negative effect
Virginia Insiders 1981	80 delinquents 13-20 NFI	A variety of crime outcome measures at 6, 9, 12 month intervals. Only positive findings, though not statistically significant.
Texas 1981	160 delinquents 15-17 (2+ offences) to four conditions (pris-	Vreeland examined official court records and self-reported delinquency after six months. The control participants did better than three treatment groups on official delinquency; little

	on+ counselling)	effect on intervention group.
New Jersey 1982	82 juveniles 11-18 yo not all delinquents NFI	Used official court records to assess the intervention. Highly negative effect. Queries over randomisation
SQUIRES programme California, 1984	108 delinquents (multiple prior)	Lewis compared participants on seven crime outcomes at twelve months.????
Kansas Program	52 delinquents 14-19	The investigators examined official (from police and court sources) and self-report crime outcomes at six months. No effect.
Mississippi Project Aware 1992	176 juveniles 12-16 NFI	Experimental and control groups were compared on a variety of crime outcomes retrieved from court records at 12 and 24 months.????

Reviewing the subjects and the measures of outcome in the table, we see that

- (a) Most used court records to provide outcome data (although jurisdictions vary in what counts as an offence or violation of parole);
- (b) the California study subjects had multiple priors, Texas 2+, otherwise NFI???
- (c) Where age ranges were given, no two were the same.

In respect of (b), the probability of ‘recidivism’ is well-known to increase with the number of previous recorded offences (REFS), so the California and Texas studies are definitely different. In respect of (c), it is well-known that the likelihood of offending during the teenage period rises by every single year of age (US DoJ, 2012); so that having different age groups means that the likelihood of offending is different among different age groups – and the likelihood of re-offending is also different.

In contrast, the review by Villetaz et al. (2006) is careful to discuss and distinguish between different types of recidivism outcome measures. Efforts were made to find more differentiated indicators of reoffending, such as new arrests, contacts with police, or self-report measures. For example, some studies have shown that the frequency of new offences decreases following any type of intervention (compared

with an equivalent pre-intervention period), and that arrest data may differentiate better between groups of offenders who were treated in different ways. This is particularly true in countries where re-incarceration (for parole violations) is more common than reconviction in case of a new offence, or in continental countries where a multitude of offences leads eventually to one single rather than several convictions (that will be recorded under the most serious offence). Some studies had also used self-report data in order to assess the outcome of different interventions. In order to assess improvement, they have tried to look not only at prevalence of reconviction (or percentage of those who re-offend), but also at “incidence” rates (i.e. frequencies of new offences per time unit). Nevertheless, they could not carry out meta-analysis because there was no way of establishing that the outcomes were actually comparable.

II.3. What Has been Learnt

The central point from this contrasting brief review of the procedures being used in systematic reviews in medical health services and criminal justice research is that comparing and standardising outcome measures between different studies is fair enough so long as there is an underlying common outcome construct and an interval scale (e.g. comparing imperial and continental measures of length, volume and weight) but this is not the case with the sentences for juvenile delinquents considered above so that any standardising procedure does not solve the problem.

III. And in Education?

III.1 Common outcome concept or construct

Valid comparisons of the effectiveness of interventions depend on the ability to compare effects on the same or a sufficiently similar outcome. Some educational outcomes may in principle be considered more analogous to such ‘objective’ outcomes, including for example attendance or absenteeism, enrolment and grade progression, although the concept and measurement of the latter is aligned to policy, this is not always consistent (see below).

III.1.1 Especially with Test scores

Perhaps the most common outcomes reported in education reviews are test-scores. The construct to be measured by a particular test might be, for example, 'grade 6 mathematics skills'. But an instrument (test) designed to measure this construct may nonetheless cover a broad or narrow range of content, may be long or short and may be more or less accessible or demanding (the latter clearly also depends on the sample of test takers, considered further below). There is therefore no automatic reason why the content of any two grade 6 maths tests should be considered 'sufficiently similar' to count these instruments as yielding comparable data; that is, unless the two tests can be shown explicitly to contain similar items.

If this is a problem in relation to two grade 6 tests, it is surely true of a grade 6 maths test and a grade 3 maths test and most of tests in other subjects etc. The PISA team of course go to great lengths to ensure comparability but some of the items may not be covered at all in the curriculum of some of the participating countries.

Although it might be possible to establish a consistent conceptual definition of a particular outcome (i.e. by defining mathematics ability in terms of skills in multiplication and division), the particular skills which should be included in the definition are likely to differ by population (i.e. multiplication and division in Grade 3 will differ substantially from multiplication and division in Grade 5). At the same time, whilst it is possible conceptually to define reading ability in terms of vocabulary, sentence construction, etc., it is (much) more difficult to tie those down to a specific set of instructional materials delivered in a particular grade.

For example, a part of the Young Lives study, a school survey in Vietnam measured learning attainment at two points at the beginning and end of Grade 5 – pupils gained 0.13SD in Vietnamese reading and 0.41 in maths (on an interval IRT scale) with similar inputs in terms of teaching hours and materials and with the same population – but what differs is not the effectiveness of education necessarily but the content of the tests – the G5 maths test can easily target precisely what should be learned in that year while in reading it is much more difficult to do so and the test necessarily covers a broader range of skills (so it is more difficult to make progress). In this case school effectiveness and test effects are very difficult to separate.

Even if a group of studies looks at the same outcome within comparable samples, it is unlikely that the studies will use the same test. Indeed, it is often the case that both the test *content* - in absolute terms and in terms of its breadth or scope - and the *scoring methodology* differ across studies. Tests may be criterion- or norm-referenced, and they may use raw or weighted scores. Furthermore, the relationship between the population and the measurement instrument used is likely to differ depending on the sample. A Grade 3 maths test can be easy or difficult, depending on either the items included or on the population to whom it is administered. If the sample largely consists of high-achievers, the difficulty level will differ from a scenario in which the same test is used with a sample largely consisting of drop-outs.

III.1.2 Other issues with Tests

The issue of diverse populations is particularly problematic in international development research, given that systematic reviews often attempt to combine results from studies completed in different countries around the world. This adds the complicating factor of language, as some students study in their native tongue while others study in a second or third language.

Moreover, reviews often go beyond comparing test data from different tests in the same subject and routinely combine studies which investigate different constructs from different academic subjects. For instance, in their recent review of educational interventions in low- and middle-income contexts, Snilstveit et al. (2016) investigated impact on three broad learning domains: “1) maths and language arts (local language and any official language(s) of country), (2) cognitive and problem solving skills, and (3) composite assessment scores from test scores in different subjects or other measures of skills and learning” (p. 17). In such reviews, the challenges of comparison are considerable, if not insurmountable.

Another example is the difficulty of making a sensible comparison between changes in an UWEZO⁵ test score out of 100 between those who are illiterates scoring 0 and those who are just literate scoring 5 (where would expect there to be an increase of at least one standard deviation on their scale), with those at the top of a scale who are scor-

⁵ UWEZO organises citizen-led surveys in Kenya, Uganda and Tanzania. UWEZO means capability in Swahili.

ing 90 rather than 85 on a maths or science test (where the increase in terms of standard deviations would be very small).

Finally, the use of IRT (Item Response Theory) as a statistical procedure (the PISA favourite) for comparing scores from two test scales might seem to fare better because they are supposed to be on an interval scale, but the problem remains that it depends on what material the test covers and who are the test takers in order to understand what a standard deviation in the test score means. If there is a narrow test of, for example, G4 maths content then one might expect a large amount of learning in a year but if it is a general test covering several grades then a standard deviation change on this test might represent several years of learning (whereas on the Grade 4 test one could theoretically learn everything in a year). Moreover, there are very strict assumptions involved in using IRT (Goldstein, 2004).

III.1.3 Specific Reviews of Tests in Languages

Examination of the 14 studies testing languages included in the review (see Table) shows that 8 were in Spanish, 3 were in English, 2 were in French and 1 in Bahasa; 3 were at school level, 1 in 'secondary' (unspecified) and 5 were at student level in one grade and the other 5 in multiple grades. Where grade(s) was(were) specified in the studies, one each were at Grades 1 and 2 level, six were at 3rd grade, three were at 4th grade, two were at 5th grade and three were at 6th grade; one was in secondary and one was unspecified (see Table 2). How these could have possibly been seen as even potentially comparable; the mixture should have been seen as blasphemous to the advocates of meta-analysis advocates

Authors	Country	Lang.	Year of Data	Grade
Bando (2010)	Mexico	Sp	2001-07	School-level
Beasley & Huillery	Niger	Fr	2008	Student level, G1,4,6
Blimpo & Evans (2011)	The Gam- bia	En	2008-11	3 rd and 5 th grade
Di Gropello & Mar- shall (2005)	Honduras	Sp	2003-03	Student level 3 rd grade
Duflo etal	Kenya	En	2005-06	Grades 2 and 3
Khatttri	Philippines	Sp	2002-03 and 2004-05	School level av. per-centile scores
King & Ozler (a)	Nicaragua	Sp	1995-97	Secondary
Lassibilie	Madagascar	Fr	2006-07	School level
Parker (2005)	Nicaragua	Sp	2002	3 rd , 6 th Grade
Pradhan	Indonesia	Ba		4 th & 6 th Grade
Rodriguez	Colombia	Sp	2002-03 and 2005	5 th Grade
Santibanez	Mexico	Sp	2007-08 and 2008-09	3 rd Grade
Sawada & Ragetz	El Salvador	Sp	1996	3 rd Grade
World Bank (a)	Sri Lanka	En	2006	4 th Grade

For full description of studies, see Carr-Hill et al (2015)

III.2 Common Scale or Metric

III.2.1 Introduction

When common outcome constructs are used for comparison of effect sizes, it is also essential to ensure that these constructs are measured in ways that can be legitimately compared. It may be necessary to transform measures from one metric to another for comparison and this should not present difficulty when the metrics are linked to an underlying interval scale and there is a known formula or procedure to

equate scales. This is clearly very straightforward in the case of metrics used to measure weight or temperature. In such cases the scales are sample-independent interval scales.

Greater difficulty arises where measurement is on a non-interval scale and/or where measurement is sample-dependent. These issues very commonly apply to test scores. Other types of data based on subjective reports are also affected but there are efforts to create sample-independent interval-scaled measures from such data. In education, TIMSS and PISA represent similar efforts but essential to this endeavour is the use of common items to measure common curricular constructs across contexts; and this is very problematic across societal, linguistic and cultural differences.

Tests which have no common items cannot be equated - they are different - and entirely relative to their particular items and their samples of test-takers. There is no underlying scale to which individual tests may be anchored. Tests designed specifically for comparison, such as PISA appear to be an obvious exception (although also based on the problematic Rasch models), but are very rarely used in research which makes its way into SRs.

Different tests and transformations of test-scores produce different distributions of scores (and, therefore, different standard deviations). These differences may be the result of true differences in the population, following a particular intervention, but they may also be the results of measurement error.

III.2.2 Other Educational Outcome Measures

Even apparently more objective measures - such as enrolment, attendance, drop out and repetition - have the same problem. We demonstrate this through the studies considered in our systematic review.

Student absenteeism/ attendance

In the case of absenteeism from school, for example, provided appropriate detailed data are available, measures may in principle be equated on a scale such as “annual percentage days absent” or similar, so that effect-sizes would be reported as percentage-point changes. In reality this is not always straightforward. In the Carr-Hill et al review, there were 6 studies which looked at student absenteeism or attendance (Barr et al., 2012; Blimpo and Evans, 2011; Di Gropello and Marshall, 2005; Jiminez and Sawada, 1999; Lassible et al. 2010; and

Sawada and Ragatz, 2005). But inspection (see Table 3) of the definitions they have used shows that there is no possibility of combining more than 2 of them:

- a. Barr et al. (2012) and Blimpo and Evans (2011) use absenteeism on day of visit across the school;
- b. Jimenez and Sawada (1999) and Sawada and Ragatz (2005) use days absent in previous month 3rd grade only;
- c. Lassibile et al. (2010) uses attendance during month prior to visit across the school
- d. Di Gropello and Marshall (2005) use a student reported ordinal measure.

In principle, (b) and (c) could be combined (absence and attendance in the previous month are complementary measures) except that (b) is for third grade and (c) is across the whole schools and we know that patterns of absence/ attendance are grade dependent (UIS/UNESCO, 2012).

Table 3: Student Absenteeism/ Attendance

Author(s)	Country/ Programme	Definition of Absenteeism/Attendance and Page reference
Barr et al., 2012	Uganda	Absenteeism on day of visit (p.19)
Blimpo & Evans, 2012	The Gambia	absenteeism (on day of classroom visit) (p.13)
Di Gropello& Marshall, 2005	Honduras	a student reported ordinal measure of absence over the year. (Table 9.23, p.352)
Jimenez & Sawada, 1999	El Salvador, EDUCO	N of days absent in past month, 3 rd Grade; page 425
Lassibille et al., 2010	Madagascar	All-school attendance during the month preceding school visit, according to director
Sawada & Ragatz, 2005	El Salvador, EDUCO	Days of absence during last four weeks, 3 rd grade only

Student Failure

There are 5 studies which look at student failure rates (Bando, 2010; Gertler et al, 2012; Murnane et al., 2012; Rodriguez et al., 2009; Skoufias and Schapiro, 2006). But none of them give a precise defini-

tion, in terms of which subjects are included in the assessment of a student's failure at the end of a year; and although it is probable that, in Latin America, these will include Spanish, Mathematics and Science, we do not know the relative weights given to each subject). Closer inspection (see Table 4) suggests that the only two that probably used equivalent definitions - because they were evaluating the same programme (PEC) in adjacent time periods - were Murnane et al. (2006) and Skoufias and Schapiro (2006), both of whom had similar overall failure rates. Bando (2010) was also evaluating PEC over the same time period but the value she has for failure rate is about four times that of the other authors, so her definition must have been different. Moreover, whilst Gertler et al. (2012) were also studying Mexico, it was several years later and another programme (AGE) and their failure rate was twice as high. Finally, Rodriguez et al. (2009) were studying in Columbia where primary school lasts five years rather than six years in Mexico, so that the basis for the calculation is different. Once again, there is no possibility of combining more than two of the studies:

Table 4: Student Failure (five studies)

Author(s)	Country or Programme	Text on how failure was measured
Bando (2010)	Mexico / PEC	"The second census is carried out at the end of the school year and includes information on failure and dropout rates."; probably across whole school. Failure rate about 20% in each year (Table B2)
Gertler et al 2008	Mexico / AGE	Mexico: School-level grade failure; possibly compatibles. Failure rate is 10%
Murnane et al., 2006	Mexico / PEC	1-(number of students who passed grade in school year t divided by the number who were enrolled at end of school year t). Overall rate 5%
Rodriguez, 2009	Columbia / PER	Data for indicators on efficiency from Ministry; probably across whole school (that is 5 school years rather than 6); declines from 14% to 8%
Skoufias & Shapiro (2006)	Mexico	1-(number of students who passed grade in school year t divided by the number who were enrolled at end of school year t). School level. Failure rate 5%

Studies on progression or continuation are even more vague (Table 4)

Table 4. Progression/ Continuation (three studies)

Author(s)	Country/ programme	Text on measurement
Barr et al., 2012	Uganda	Class grade (taking values of 3 for pupils enrolled in Primary 3, 4 for pupils in Primary 4, etc.) regressed on a set of treatment indicators and strata fixed effects
Beasley & Huillery, 2014	Niger	Single analysis (Table 15) of enrolment by grade but no suggestion that this a continuation from baseline.
Jimenez & Sawada (2003)	El Salvador	Jimenez and Sawada (2003) have results for progression (continuation).

IV. Conclusions

Studies need to be precise in their use of quantitative data: and most individual studies are careful in their description and use of their own data, although the care does tend to vary across health and social sciences, with studies authored in medical journals being more precise than those in economic/econometric/ social science journals. But the real problem arises when authors (try to) combine them: many of those reviews included in the Cochrane Collaboration Library do include only those studies using essentially identical outcomes measures and intervention procedures; whilst studies included in the Campbell Collaboration are more cavalier in combining studies where the equivalence of outcomes is less clear and here can be wide variability in the interventions, although there are some exceptions (e.g. Killias, ???); and in the particular studies we have reviewed here in the education area, no attention seems to have been paid at all to the fact that the actual outcomes measured and the interventions are different

It is all the more puzzling, because several of the authors of some of the studies reviewed here are PISA advocates (see Petrinis, 2013). It is amusing to see that authors of the Systematic Reviews tend to criticise PISA (for example, Boruch in Boe et al, 2002; Morgan in Pereyra, 201?) . Whatever the valid

criticisms of the use of IRT by PISA (Goldstein, 2004?), their technical teams move heaven and earth in terms of face and internal validity to make sure that all test items are measuring the same construct [They also move an ersatz heaven and earth via IRT to claim that all test items are on the same scale(ersatz because, despite their best efforts, IRT requires that too many assumptions are satisfied to correctly claim that there is a single scale).] But leaving the last point aside, PISA techies should be horrified at the cavalier way in which those conducting systematic reviews (implicitly) assume that scores on a Spanish 3rd grade test can be made 'equivalent' to English scores across the whole school, simply by using standardised mean differences (with no attempt to carry out a very complex version of IRT).

Systematic review of outcomes have to pay much more attention to whether the data they are using are measuring the same construct.

References

Bando, R. (2010). *The Effect of School Based Management on Parent Behavior and the Quality of Education in Mexico*. Unpublished PhD thesis.

Barr, A., Bategeka, L., Guloba, M., Kasirye, I., Mugisha, F., Serneels, P. & Zeitlin, A. (2012). *Management and motivation in Ugandan primary schools: an impact evaluation report*. PEP Working Paper. Nairobi: Partnership for Economic Policy.

Beasley, E. & Huillery, E. (2014). *Willing but Unable: Short-Term Experimental Evidence on Parent Empowerment and School Quality*. Unpublished manuscript. Available at:

<http://www.povertyactionlab.org/publication/willing-unable-short-term-experimental-evidence-parent-empowerment-and-school-quality>

Blimpo, M. & Evans, D.K. (2011). *School-Based Management and Educational Outcomes: Lessons from a Randomized Field Experiment*. Unpublished manuscript. Available at:

http://siteresources.worldbank.org/EDUCATION/Resources/Blimpo-Evans_WSD-2012-01-12.pdf.

* Carr-Hill R.A. (1996), *Welcome? To the Brave New World of Evidence Based Medicine, Social Science and Medicine* (1996)11: 1467-8

- Carr-Hill R. (2010) The poor may always be with us; but we don't know how many there are or where they are, Editorial, *Journal of Health Services Research and Policy*, 2010
- Carr-Hill, R.A., Schendel, R., Rolleston, C. et al. (2015) *The Effects of School-Based Decision Making on Educational Outcomes in Low and Middle Income Contexts: A Systematic Review* Campbell Systematic Reviews 9
- Di Gropello, E. & Marshall, J.H. (2005). 'Teacher effort and schooling outcomes in rural Honduras.' In: E. Vegas (ed), *Incentives to improve teaching*. Washington DC: World Bank, pages 307-358.
- Duflo, E., Dupas. P. & Kremer, M. (2012). *School Governance, Teacher Incentives, and Pupil-Teacher Ratios: Experimental Evidence from Kenyan Primary Schools*. NBER Working Paper No. 17939. Cambridge, MA: National Bureau of Economic Research.
- Evans, D.K. & Popova, A. (2015). *What Really Works to Improve Learning in Developing Countries? An Analysis of Divergent Findings in Systematic Reviews*. Policy Research Working Paper 7203. Washington, DC: World Bank.
- Gertler, P., Patrinos, H.A. & Rubio-Codina, M. (2012). 'Empowering parents to improve education: Evidence from rural Mexico.' *Journal of Development Economics*, 99(1):
- Goldstein, H. (2004) International comparisons of student attainment: some issues arising from the PISA study *Assessment in Education*
- Guerrero, G., Leon, J., Zapata, M., Sugimaru, C., & Cueto, S. (2012). *What works to improve teacher attendance in developing countries? A systematic review*. London: EPPICentre, Social Science Research Unit, Institute of Education, University of London. Available at: <http://eppi.ioe.ac.uk/cms/Default.aspx?tabid=3377>.
- Harvard Family Education Research Project. (2005) What is the Campbell Collaboration and how is it helping to identify "what works"? Volume XI, Number 2, Summer 2005 Issue Topic: Evaluation Methodology, 68-79.
- Jimenez, E. & Sawada, Y. (1999). 'Do Community-Managed Schools Work? An Evaluation of El Salvador's EDUCO Program.' *The World Bank Economic Review*, 13(3): 415-441.

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

Lassibille, G., Tan, J.P., Jesse, C. & Van Nguyen, T. (2010). 'Managing for Results in Primary Education in Madagascar: Evaluating the Impact of Selected Workflow Interventions.' *World Bank Economic Review*, 24(2): 303-329.

Laurant, M. Reeves, D., Hermens, R., Brasppnning, J., Grol, R. and Sibbald, B. (2005) Substitution of doctors by nurses in primary care, Cochrane Effective Practice and Organisation of Care Group

Murnane, R.J., Willett, J.B. & Cardenas, S. (2006). *Did the Participation of Schools in Programa Escuelas de Calidad (PEC) Influence Student Outcomes?* Unpublished manuscript.

Parker, C.E. (2005). 'Teacher incentives and student achievement in Nicaraguan autonomous schools.' In: E. Vegas (ed), *Incentives to improve teaching*. Washington DC: World Bank, pages 359-388.

Petrosino, A., Morgan, C., Fronius, T.A., Tanner-Smith, E.E., & Boruch, R.F. (2012). *Interventions in developing nations for improving primary and secondary school enrollment of children: A systematic review*. Campbell Systematic Reviews 2012: 19. Available at: <http://www.campbellcollaboration.org/lib/project/123/> . .

Petrosino, A. (2013) PISA Results: Which Countries Improved Most? 12/03/2013

Pereyra, M.G., Kotthoff, H-G. and Cowen, R. (2011) PISA Under Examination: Changing Knowledge, Changing Tests, and Changing Schools

McKenzie, D. (2015) *Notes from the AEA's: Present bias 20 years on + Should we give up on S.D.s for Effect Size?* Development Impact blog; January 13. Washington, DC: World Bank. Available at: <https://blogs.worldbank.org/impac-tevaluations/notes-aeas-present-bias-20-years-should-we-give-sds-effect-size>.

Pritchett, L. & Sandefur, J. (2013). *Context Matters for Size: Why External Validity Claims and Development Practice Don't Mix*. Working Paper 336. Washington, DC: Center for Global Development.

Rodriguez, C., Sanchez, F. & Armenta, A. (2009). 'Do Interventions at School Level Improve Educational Outcomes? Evidence from a Rural Program in Colombia.' *World Development*, 38(3): 415-428.

Rolleston, C., Z. James, L. Pasquier-Doumer and Tran Ngo Thi Minh Tam (2013) 'Making Progress: Report of the Young Lives School Survey in Vietnam'. Young Lives Working Paper 100 Oxford: University of Oxford

Santibañez, L., Abreu-Lastra, R. & O'Donoghue, J. (2014). 'School based management effects: Resources or governance change? Evidence from Mexico.' *Economics of Education Review*, 39: 97-109

Sawada, Y. & Ragatz, A.B. (2005). 'Decentralization of education, teacher effort, and educational outcomes.' In: E. Vegas (ed), *Incentives to improve teaching*. Washington DC: World Bank, pages 255-306.

Singh, A. (2015). *How standard is a standard deviation? A cautionary note on using SDs to compare across impact evaluations in education*. Development Impact blog; January 13. Washington, DC: World Bank. Available at: <http://blogs.worldbank.org/impac-tevaluations/how-standard-standard-deviation-cautionary-note-using-sds-compare-across-impact-evaluations>.

Skoufias, E. & Shapiro, J. (2006). *Evaluating the Impact of Mexico's Quality Schools Program: The Pitfalls of Using Nonexperimental Data*. Washington, DC: World Bank (World Bank Policy Research Working Paper, Impact Evaluation Series No. 4036).

Snilstveit, B. (2016) [Interventions for improving learning outcomes and access to education in low- and middle-income countries: a systematic review](#)., International Institute for Impact Evaluation

US Department of Justice(2012)

Villettaz, P., Killias, M. and Zoder, I. (2006) The Effects of Custodial vs. Non-Custodial Sentences on Re-Offending: A Systematic Review of the State of Knowledge Campbell Systematic Reviews

Westhorp, G., Walker, B. and Rogers, P. (2014). *Under what circumstances does enhancing community accountability and empowerment improve education outcomes, particularly for the poor? A realist synthesis*. London: EPPI-Centre, Social Science Research Unit, Institute of Education. Available at:

<http://r4d.dfid.gov.uk/pdf/outputs/SystematicReviews/Community-accountability-2014-Westhorp-report.pdf>.

White, H. (2009). 'Theory-Based Impact Evaluation: Principles and Practice.' *Journal of Development Effectiveness*, 1(3): 271-284.

World Bank. (2007). *What Do We Know About School-Based Management?* Washington, DC: World Bank.