International Journal of Educational Research 77 (2016) 1–14

Contents lists available at ScienceDirect

International Journal of Educational Research

journal homepage: www.elsevier.com/locate/ijedures

Same difference? Understanding variation in the estimation of effect sizes from educational trials

ZhiMin Xiao [a,*], Adetayo Kasim [b], Steve Higgins [a]

[a] *School of Education, Durham University, Durham, DH1 1TA, UK*
[b] *Wolfson Research Institute, Queen's Campus, Durham University, Stockton-on-Tees, TS17 6BH, UK*

ARTICLE INFO

Article history:
Received 21 November 2015
Received in revised form 22 January 2016
Accepted 1 February 2016
Available online 15 March 2016

Keywords:
Evaluation
Effect size
RCT
Multilevel modelling
Bayesian MLM

ABSTRACT

By applying four analytic models with comparable outcomes and covariates to a dataset of 20 outcomes from 17 educational trials, we found results closely matching in well-powered studies without serious implementation problems. The interventions and evaluations were all funded by the Education Endowment Foundation and independently evaluated. We demonstrated that when an analysis takes little account of research design, or where there were difficulties with implementation and data collection, point estimates of effect differ and estimates of precision vary. This adds to the challenge of understanding the comparative impact of interventions and deciding which are worth scaling up.

1. Introduction

The Education Endowment Foundation (EEF) is an independent grant-making charity, which aims to address the challenge of disadvantage in educational achievement associated with family income and to help children from all backgrounds achieve academically. Established in 2011 with a £125 million endowment from the Department for Education, the EEF is dedicated to raising the educational attainment of disadvantaged children in primary and secondary schools in England using research and evidence in three ways. This is first by identifying and funding promising educational innovations that address the needs of children facing disadvantage; second by evaluating these innovations to extend the evidence on what is educationally effective and what can be made to work at scale; and third by encouraging schools, governments, charities, and others to apply evidence and adopt innovations found to be successful.

This paper focuses on the second of these approaches and presents a repeated analysis of evaluation results from 17 educational trials (see Table 1) which all reported findings publicly in 2014–15. All EEF projects are independently evaluated by a number of evaluation teams which are from universities and independent research organisations. The data from these projects are deposited in an archive which will become a rich repository of findings from EEF interventions (over 100 have been commissioned so far involving over 650,000 pupils). One goal is to track the longer term impact of interventions as results from national tests become available where this is possible.

[*] Corresponding author.
E-mail address: zhimin.xiao@durham.ac.uk (Z. Xiao).

http://dx.doi.org/10.1016/j.ijer.2016.02.001

Table 1
Project information. The numbers 1-38 are EEF project numbers. We abbreviate full EEF titles to labels that mark each of the 20 outcomes for this study. The references to the 17 evaluation reports can also be used to identify evaluation teams.

Project	Archive label	Full EEF title	Evaluation report
1	ffe, ffm	Future Foundations	Gorard, Siddiqui, and See (2014)
2	sor	Switch-on Reading	Gorard, See, and Siddiqui (2014)
3	gfw	Grammar for Writing	Torgerson, Torgerson, Mitchell, et al. (2014)
4	rfr	Rhythm for Reading	Styles, Clarkson, and Fowler (2014b)
9	catchn, catcht	Catch Up Numeracy	Rutt (2014)
10	cbks+, cbks	Chatterbooks	Styles, Clarkson, and Fowler (2014a)
13	rp	Rapid Phonics	King and Kasim (2015)
14	ar	Accelerated Reader	Gorard, Siddiqui, and See (2015a)
15	bp	Butterfly Phonics	Merrell and Kasim (2015)
16	iwq	Improving Writing Quality	Torgerson, Torgerson, Ainsworth, et al. (2014)
17	sar	Summer Active Reading	Maxwell et al. (2014a)
18	text	TextNow	Maxwell et al. (2014b)
21	uos	Units of Sound	Sheard, Chambers, and Elliott (2015)
22	ve	Vocabulary Enrichment	Styles, Stevens, Bradshaw, and Clarkson (2014)
31	fs	Fresh Start	Gorard, Siddiqui, and See (2015b)
32	tfl	Talk for Literacy	Styles and Bradshaw (2015)
38	mms	Mathematics Mastery Secondary	Jerrim et al. (2015)

1.1. Rationale for the archive analysis

Andrew Gelman described statistics as "the science of defaults" (in Lin et al., 2014, p. 293), by which, he meant applied statisticians usually choose (and recommend) their default or preferred methods to solve problems in a wide range of settings, although these may not always be optimal in answering project-specific questions. In the EEF reports that are made publicly available, there are patterns of design and analysis associated with specific evaluation teams. As shown in the tables that follow, evaluators sometimes applied the same approach to different projects, even when the research designs and the quality of the data for causal inference varied. This arises, as Gelman noted, because there are competing philosophies, assumptions, and approaches to statistical analysis and inference, which makes consensus on *the* best approach difficult to achieve. This paper explores the differences these choices make in terms of the outcomes from different methods of analysis for each trial.

Archive analysis differs from replication studies in that the former does not require the collection of new data from the same population. Instead, it re-uses the data from original trials to reproduce the original results and/or answer new research questions. In this paper, our goal is mainly to answer a new question: how do effect size estimates and their uncertainties vary under different model and design specifications? Unlike meta-analyses, which usually rely on summary statistics extracted from secondary sources that do not always report research in consistent and transparent ways to synthesise evidence, this archive analysis re-evaluates the evidence already found from EEF trials. In other words, it investigates how sensitive the findings are to design and model specifications, using full datasets from the aforementioned evaluation projects. It also aims to explain what causes any variation in impact and to support any subsequent comparison of impact between the studies examined.

The educational interventions included in this analysis all set out to improve educational attainment for school-age pupils and mainly targeted literacy and/or mathematics outcomes, with some focusing on phonics, vocabulary, grammar or other aspects of literacy, some through summer school interventions, others in schools as pedagogical interventions, such as those based on developing mastery or promoting learning through talk or thinking strategies. The samples varied in size from 178 to 5830 pupils, with numbers of schools (clusters) involved varying from three to 54 (see Table 2). Full details of the interventions and evaluations can be found in the individual evaluation reports which are listed in the references.

1.2. Effect size and *p*-value

A key concept in this paper is that of effect size, which, according to Borenstein (2009), is an index used to quantify the magnitude of relationship between two variables or the difference between two groups (p. 222). In theory, effect sizes from different studies, regardless of the design, should measure, approximately at least, the same relationship and be comparable. Like *p*-values, effect sizes are scale free (Hedges, 2008, p. 168). The two are certainly related to each other, but they are not the same – a significant *p*-value could be a function of a large effect or a small effect in a study with a large sample size, likewise, a big *p*-value could reflect a small effect or a large effect in a small study (Borenstein, 2009, p. 223). Effect size estimates are based on the samples studied, and the uncertainties surrounding those point estimates give us a range of possible effect sizes for the corresponding populations. While the calculation of effect size is a mathematical process, its interpretation involves judgement, and it is of little practical value to say an effect is large or small without comparing it with others in a specific context (Hedges, 2008, p. 170).