Application of Bayesian Posterior Probabilistic Inference in Educational Trials

Germaine Uwimpuhwe[1], Akansha Singh[1], Steve Higgins[2], Adetayo Kasim[1,3*]

[1]Department of Anthropology, Durham University, Durham, UK

[2]School of Education, Durham University, Durham, UK

[3]Durham Research Methods Centre, Durham University, Durham, UK

[*]*Corresponding author: a.s.kasim@durham.ac.uk*

Abstract

Educational researchers advocate the use of an effect size and its confidence interval to assess the effectiveness of interventions instead of relying on a p-value, which has been blamed for lack of reproducibility of research findings and the misuse of statistics. The aim of this study is to provide a framework, which can provide direct evidence of whether an intervention works for the study participants in an educational trial as the first step before generalising evidence to the wider population. A hierarchical Bayesian model was applied to ten cluster and multisite educational trials funded by the Education Endowment Foundation in England, to estimate the effect size and associated credible intervals. The use of posterior probability is proposed as an alternative to p-values as a simple and easily interpretable metric of whether an intervention worked or not. The probability of at least one month's progression or any other appropriate threshold is proposed to use in education outcomes instead of using a threshold of zero to determine a positive impact. The results show that the probability of at least one month's progress ranges from 0.09 for one trial, GraphoGame Rime, to 0.94 for another, the Improving Numeracy and Literacy trial.

Keywords: Randomised Control Trial, Multisite Trial, Cluster randomised trial, Effect size, significant threshold, Bayesian probability, educational evaluation

Introduction

The aim of the Education Endowment Foundation (EEF) is to improve educational attainment, especially of disadvantaged pupils, in schools across England through different interventions and teaching approaches. Similar to the other disciplines, educational stakeholders make decisions about an intervention based on the statistics obtained from either a Bayesian or a frequentist approach. In the latter approach, this is often based on a p-value, which is mostly misinterpreted as 'a probability that the null hypothesis is (not) true'. According to Lesaffre and Lawson (2012), this probability can be formally called the posterior probability (i.e. $P(H_0|data)$) that can be correctly obtained through a Bayesian approach. Many statisticians have advocated replacing the p-value with a Bayesian approach as the latter approach describes how to think about probability as the plausibility of an outcome, rather than as the potential frequency of that outcome (Nuzzo, 2014). A p-value has wide sample-to-sample variability unless the statistical power is very high. It is also not able to reliably indicate the strength of evidence against the null hypothesis (Halsey *et al*., 2015). Additionally, the replication success of any study finding is negatively correlated with the p-value of the original study (Camerer *et al*., 2018).

Furthermore, the current practice of using p-values is problematic and provides oversimplified evidence based on an arbitrary threshold (Cumming, 2008; Goodman, 2019; Trafimow & Marks, 2015; Wasserstein, Schirm, & Lazar, 2019; Ying & Belitskaya-Levy, 2015). The point estimate of the treatment difference and its confidence interval (CI) can be used instead to quantify the impact of an intervention and to provide the magnitude of an intervention's effectiveness in educational trials (Valentine & Cooper, 2003). However, these statistics are not comparable between trials because of their dependence on the units, measuring methods, and scales used in the evaluation of each intervention (Ying & Belitskaya-Levy, 2015). In this case, a unitless statistic, typically the effect size (ES) is used. Hedges (2007, 2008) elaborated the methods, which can be used to estimate ESs with their standard errors and CI in the most commonly designed educational trials. However, a CI carries a similar message to a p-value, albeit with the advantage of providing a range of values as evidence of the impact of an intervention. Given the criticisms of p-values and significance tests, a CI should not be considered as a measure of significance (McShane *et al*., 2019; Wasserstein *et al*., 2019). Further, the frequentist interpretation of a CI as a percentage (mainly 95%) of probability that true value falls within confidence limits actually corresponds to the interpretation of a Bayesian credible interval (BCI) (Cumming, 2008; Lesaffre & Lawson, 2012; Thompson, 2002), but does not actually represent this. In most of the cases, especially non-statisticians intuitively apply a Bayesian interpretation to a frequentist approach.

It therefore seems natural to consider a Bayesian framework for the evaluation of educational trials. The correct BCI estimates can easily be obtained as each unknown quantity is treated as a random parameter with its prior distribution. The posterior distribution of the unknown quantity is also straightforward as the information from the data is combined with the prior information. One can obtain a BCI using the posterior distribution, as the uncertainty of each unknown quantity is explicitly indicated by the spread of the posterior distribution (Kruschke & Liddell, 2018). In general, the fundamentals of Bayesian inference can be summarised through Bayes' theorem, mathematically written as $P(\theta|data) \propto P(data|\theta) * P(\theta)$, where $P(\theta|data)$, $P(\theta)$ and $P(data|\theta)$ refers to the posterior, prior and the sampling distribution (or data distribution), respectively (Gelman *et al*., 2013; Lesaffre & Lawson, 2012). Priors can be developed using a range of information sources including previous information and expert opinion. Although the quantification of previous knowledge is possible in Bayesian statistics, it is also arguably its most controversial aspect as the choice of appropriate priors is inherently subjective (König & Schoot, 2018). However, uninformative or vague distributions can also be used to allow inferences to be driven largely by the data.

This study proposes the use of Bayesian posterior probability as a complementary metric for evidence in educational trials. Bayesian inference requires a posterior distribution that can be summarised by calculating the probability that an ES exceeds a certain threshold. A higher posterior probability indicates greater effectiveness of the intervention. This probability is induced purely from the data, which makes this method consistent across trials, as suggested by several researchers including Thompson (2002; 2007), who strongly advocated that the conclusion of a study should be based on what data tells about the magnitude of effects, not based on a dichotomous reject or not reject decision. A posterior probability is a more understandable way to inform policy-makers and educational stakeholders about the effectiveness of an intervention. This method has already been applied in several other disciplines (Cummings *et al*., 2003; Friston & Penny, 2003). For example, Bayesian methods are applied in population genetics, genomics, and human genetics which have allowed complex models to be studied and biologically relevant parameters to be estimated, as well as allowing prior information to be efficiently incorporated (Beaumont & Rannala, 2004). Hahn (2014) has given examples from the reasoning and argumentation literature whereby Bayesian methods have demonstrably increased the level of behavioural prediction relative to that previously available in the relevant domain of cognition research. Kruschke and Liddell (2018) argue that Bayesian methods achieve the goals of the New Statistics (estimation based on ESs, CIs, and meta-analysis) better than frequentist methods and can be helpfully used in randomized controlled trials. However, there is limited literature on the application of Bayesian methods (König & Schoot, 2018) and posterior probability in educational trials. This study estimates effect size using a hierarchical Bayesian model and proposes posterior probability as a measure of evidence, with values closer to 1 providing increasing evidence in favour of an intervention.

Method

Case Studies

The EEF is an independent charity that aims to raise the attainment of disadvantaged children in primary and secondary schools in England. In this light, different projects have been conducted to evaluate a range of interventions directly or indirectly involving pupils to improve their educational attainment. Each project is independently evaluated and the data collated in an archive. This study comprises an analysis of ten EEF projects, selected from the available

EEF archive data according to the study design and the implementation quality. Multisite trials (MST), where there are control and intervention pupils in the same school and cluster randomised trials (CRT), where each school contains only intervention or control pupils in educational trials with reading or mathematics as an outcome with a high degree of inference validity (i.e. an EEF padlock security ≥ 3) were selected, as shown in Table 1. The security ratings of EEF's educational trials vary from low (padlock = 0) to the best type of evidence that could be expected from a study (padlock = 5) (EEF, 2019). Note that CRT and MST are fundamentally different in terms of model specification. Whilst it is sufficient to specify only schools as random effects for cluster randomised trials (CRT), it is also important to specify school-by-intervention interactions as random effects for multisite trials (MST). A brief description of each selected project is provided below.

'Table 1 here'

Five CRT projects where schools were randomly allocated to either intervention or control (Xiao, Higgins, & Kasim, 2017) were selected. Improving Numeracy and Literacy (EEF Project 41) aimed to improve both numeracy and literacy abilities. It was implemented through two separate programmes of teacher training and accompanying teaching materials focusing on mathematical reasoning and morphological activities, as well as computer games. The programme was evaluated using the mathematics attainment of pupils in Year 2 (6-7 years old) (Worth *et al*., 2015). Embedding Formative Assessment (EEF Project 110) was a whole-school professional development programme designed to improve pupils' attainment through feedback. Schools received detailed resource packs to run monthly workshops and to implement specific strategies in lessons to pupils in all year groups. General Certificate of Secondary Education (GCSE) Attainment 8 maths and English scores were used to assess the impact of the intervention on pupils who were in Year 10 (aged 14-15) at the start of the trial (Speckesser *et al*., 2018). GCSE Attainment 8 measures a student's average grade across eight subjects at age 16. 1stClass@Number (EEF Project 122) was designed by Edge Hill University.

It covered five basic mathematics topics in 30 half-hour lessons to help pupils aged 6-7 years who were struggling with the mathematics curriculum of Year 2 so that they could continue to learn successfully in class after the end of the intervention. A quantitative reasoning test, which focuses on number knowledge and mathematical problem solving, was used to evaluate the intervention (Nunes *et al*., 2018). Tutor Trust Primary (EEF Project 126) was led by a Manchester-based charity that aims to provide affordable small group and one-to-one tuition, in order to improve the mathematics attainment of pupils in Year 6 (aged 10-11) who were working below age expected levels in mathematics, as identified by their class teachers (Torgerson *et al*., 2018). Catch Up® Literacy (project 133) is a structured one-to-one intervention that aims to improve the reading ability of readers struggling in Years 4 or 5 (8-10 years old). The intervention is book-based and comprises two 15-minute sessions each week for approximately 6 to 12 months depending on individual need (Rutt *et al*., 2019).

Five MST studies were also selected where randomisation was undertaken at the pupil or class level in each school (Xiao *et al*., 2017). Catch Up® Numeracy (EEF Project 9) was a one to one intervention for pupils in upper primary schools (aged 7 – 11) who were struggling with numeracy. It consisted of two 15-minute sessions per week, for 30 weeks delivered by teaching assistants (TAs) (Rutt, Easton, & Stacey, 2014). The Summer Active Reading Programme (EEF Project 17) aimed to improve reading skills, particularly comprehension, by encouraging children to read and enjoy reading at the transition from primary school to secondary school. The study involved pupils in the north of England who were identified as unlikely to achieve at least Level 4 in English by the end of Key Stage 2 (11 years old). Booktrust implemented the programme by offering book packs to pupils and volunteers recruited by Booktrust to support a range of activities, including one to one reading, at the summer events (Maxwell *et al*., 2014). The Vocabulary Enrichment Full Programme (EEF Project 22) aimed to improve the reading abilities of pupils in Year 7 and was delivered by school teachers. It combined a phonics programme, teaching new words as well as encouraging pupils to use these words in speaking and writing through the Vocabulary Enrichment Intervention Programme (VEIP). It provided extra support for young people who were late in literacy development to move from level 3 to level 4 in English (Styles *et al*., 2014). Texting Parents (EEF Project 67) was a school-level intervention designed to improve pupil outcomes by engaging parents in their children's learning through text messages (Miller *et al*., 2017). Lastly, the GraphoGame Rime (EEF Project 109) was a computer game originally developed by a Finnish University to analyse performance and constantly adjust the difficulty of the game to match learner's ability. The English version of GraphoGame Rime was developed at the University of Cambridge. The intervention aimed to improve the reading ability of pupils in Year 3 (7-8 years old) having low literacy skills, as measured by the phonics screening check taken at the end of Year 1 (5-6 years old) (Worth *et al*., 2018). An independent research team appointed by EEF, who submitted the data from the evaluation to EEF for archiving and further research, evaluated all of the projects.

Statistical Method

A Gaussian hierarchical or multilevel model (MLM) with continuous outcomes was applied to model the relationship between post-test scores and the intervention with adjustment for other important covariates. Since the outcome from each project was continuous and pupils were clustered in schools (Lesaffre & Lawson (2012); Verbeke & Molenberghs (2009)), 2-level models were considered. The first level of the multilevel models were pupils and the second level were schools. Pupils are nested within schools for CRT, and pupils and interventions both are nested within schools for MST.

One of the EEF's objectives is to compare the ES estimates across a series of studies around England. However, the comparability of ES estimates is not always straightforward for studies that have adjusted for different covariates. Xiao *et al*. (2017) argued that it is inappropriate to compare such varying ESs. Similar thoughts were also suggested by Hedges (2008) and Nakagawa and Cuthill (2007). Following this line of argument, EEF (2015) advised using an ANCOVA model with post-test as the dependent variable and pre-test and treatment indicator as covariates. The model considered in this study incorporated only the aforementioned covariates; school indicator was added as a random variable to account for the clustering of pupils within the same school in CRT studies (Xiao, Kasim, & Higgins, 2016). For MST studies, the random effects for school and school-by-intervention interactions were added to the model (Feaster, Mikulich-Gilbertson, & Brincks, 2011). The multilevel model specified in this study is given by:

$$\mathrm{Post_{ij}} = \beta_0 + \beta_1 \mathrm{Pret_{ij}} + \beta_2 \mathrm{T_{ij}} + \boldsymbol{Z_{ij}^T} \mathrm{b_i} + \varepsilon_{\mathrm{ij}}, \quad (1)$$

Where $\mathbf{Z_{ij}^T b_i} = \mathrm{b_i}$ for CRT and $\mathbf{Z_{ij}^T b_i} = \mathrm{b1_i} + \mathrm{b2_i} * T_{ij}$ for MST

$Post_{ij}$ is the post-test scores of j^{th} pupil from i^{th} school, $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2)$ is the vector of regression coefficients, that are intercept, effect of pre-test, and treatment effect respectively. Deviation of school i from average intercept is denoted by $b1_i$ for MST (or b_i for CRT) and $b2_i$ is the deviation of school i from average treatment effect β_2. Lastly $\varepsilon_{\mathrm{ij}}$ indicate the idiosyncratic error terms. The model assumptions are specified in equation 2, 3 and 4.

$$\text{Residual:} \quad \varepsilon_{ij} \sim N(0, \sigma^2) \quad (2)$$

$$\text{CRT:} \quad \mathbf{b_i} \sim N(0, G), G = {\sigma_b}^2 \quad (3)$$

$$\text{MST:} \quad (\mathbf{b1_i}, \mathbf{b2_i}) \sim \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \mathbf{G}\right], \text{with } \mathbf{G} = \begin{pmatrix} \sigma_{b1}^2 & \sigma_{b1,b2} \\ \sigma_{b1,b2} & \sigma_{b2}^2 \end{pmatrix} \quad (4)$$

Further, for MST studies, it was assumed that there is no relationship between school and school-by-intervention $\sigma_{b1,b2} = 0$. More information about the model specifications are provided in Appendix A1. According to Hedges (2007) and Xiao *et al.* (2016), ES for the frequentist model was calculated from the model parameters defined in equation 1, 2, 3 and 4.

Within the Bayesian approach, each parameter of the model defined in equation 1 was assigned a prior. Thus, the Bayesian MLM combines a hierarchical two level model with the third level that contains the priors for model parameter as shown in equation 5, 6 and 7:

Level 1: $Post_{ij}|\boldsymbol{\beta},\mathbf{b_i},\sigma^2 \sim N(x_{ij}^T\boldsymbol{\beta} + z_{ij}^T\mathbf{b_i},\sigma^2)$ for $j = 1,\ldots,m_i; i = 1,\ldots,n$ (5)

Level 2: $\mathbf{b_i}|\mathbf{G} \sim N(0,\mathbf{G})$ for $i = 1\ldots n$ (6)

Level 3: $\sigma_i{}^2 \sim p(\sigma_i{}^2),\quad \boldsymbol{\beta} \sim p(\boldsymbol{\beta}),\quad and \quad \mathbf{G} \sim p(\mathbf{G})$ (7)

The joint posterior distribution for the Bayesian MLM is a product of the likelihood ($\prod_{i=1}^{n}\prod_{j=1}^{mi} p\left(Post_{ij}\middle|\mathbf{b_i},\sigma^2,\boldsymbol{\beta}\right)$) and the prior distributions ($\prod_{i=1}^{n} p\,(\mathbf{b_i}|\mathbf{G})p(\boldsymbol{\beta})p(\mathbf{G})p(\sigma^2)$) given by:

$$p(\,\boldsymbol{\beta},\mathbf{G},\sigma^2,b_1,\ldots b_n \mid y_1,\ldots y_n\,) \propto \prod_{i=1}^{n}\prod_{j=1}^{mi} p\left(Post_{ij}\middle|\mathbf{b_i},\sigma^2,\boldsymbol{\beta}\right)\prod_{i=1}^{n} p\,(\mathbf{b_i}|\mathbf{G})p(\boldsymbol{\beta})p(\mathbf{G})p(\sigma^2) \quad (8)$$

Note that the likelihood is determined by the data. However, prior distributions for all the unknown parameters must be specified. Although it is advisable to use informative priors where possible, there is no single agreed set of prior specifications for unknown parameters. Consequently, the use of non-informative or vague priors is recommended for a Bayesian evaluation of educational trials to ensure the conclusion is largely determined by the data instead of the researchers' prior knowledge. Vague Gaussian priors ($N(0, 10^6)$ were specified independently for each of the regression parameters, and such a prior has been used elsewhere in other disciplines (Barrado, Coart, & Burzykowski, 2019; Wang, Zhang, McArdle & Salthouse, 2008). Please do note that the similar results can be obtained with $N(0, 10^3)$ or $N(0, 10^2)$. Whilst independent inverse gamma priors ($IG(0.0001, 0.0001)$) according to Congdon (2014) were specified for each of the variance parameters specified in equation 8.

Analytical determination of the posterior distribution in Bayesian models is often not feasible in practice. Hence, Markov Chain Monte Carlo (MCMC) a computer-driven sampling method (Van Ravenzwaaij, Cassey, & Brown, 2018) is used in this study to determine information about the posterior distributions. Different initial values per chain (in this study 3 chains were used) are needed to form the first iteration for the Markov Chain. Each new iteration depends on the previous iteration values; this process continues until sampling from stationary distribution. At this stage, all chains are sampling from the same distribution, even if their starting points are very different, and this is one of convergence indications. The number of iterations necessary to obtain convergence depends on the analysis at hand, the more you increase this number the greater the chance to sample from the target distribution (Raftery & Lewis, 1995). In this study, 200,000 iterations were considered for each chain and the first half of the iterations were discarded as the burn-in part, as the posterior distribution should not depend on the initial values,. The burn-in part is the number of iterations ignored since the beginning of an MCMC run so that the posterior distribution can be independent of the initial values. Further, the posterior distribution should be made of independent iterations, so thinning was used. The purpose of thinning is to reduce autocorrelation between iterations. We used thinning = 10 by keeping only every tenth iteration to build the posterior distribution. After checking model convergence, we believe that the remaining iterations are guaranteed to be a sample from the target or posterior distribution. The posterior distribution for each parameter was made up of 10,000 iterations from each chain. Depaoli, Clifton and Cobb (2016) provide more details about initial values, burn-in and thin specification in a Bayesian model.

All Bayesian results were reported after checking model convergence (Figure A1 and Figure A2) according to Lesaffre and Lawson (2012). Figure A1 and A2 in the appendix provides the trace and density plots from MCMC for convergence check in each trial. Another way to monitor the chain convergence is through the Rhat convergence diagnostic, which compares the between- and within-chain estimates for model parameters. If chains have mixed well, the between- and within-chain estimates agree, Rhat values for model parameters are very close to 1 (Brooks & Gelman, 1998). The Rhat estimate for all parameters of Bayesian model used for each trial in this study is 1, which suggests that the Bayesian model has converged well. Note that formal Geweke test can also be used to check convergence (Lesaffre & Lawson, 2012). Further, a sensitivity analysis was conducted to study the impact of using a different prior specification on the posterior probability estimates. The impact of using a different prior distribution for variance parameters was assessed on posterior probability estimates.

Non-identifiability was not an issue for the Bayesian hierarchical model used in this study, since our model converged. Although, it is worth noting that hierarchical models are often slow to converge for a number of reasons including identifiability issues, particularly when using vague priors (Gelfand & Sahu, 1999) and the likelihood is not sufficient to provide a unique estimate of the model parameters. Studies in the future can follow these steps mentioned below to resolve such an identifiability issue. The common practice is to use sum-to-zero or corner constraints (Ntzoufras, 2011, Congdon, 2019). For instance, the corner constraint is often used in classical analysis, by setting unidentified parameters to 0 to obtain unique solutions. Furthermore, non-identifiability can be improved by specifying suitable informative priors or by applying parameter constraints in a Bayesian model (Congdon, 2019). In this light, Gelfand and Sahu (1999) pointed out that placing a proper point mass prior on the unidentified parameters to make the posterior distribution proper, amount to constraining non-identifiable parameters to 0. Alternatively, efficient parameterisation such as hierarchical centring can be helpful in achieving Bayesian model convergence (Gelfand, Sahu, & Carlin, 1995).

Calculating Posterior Probabilities

For the Bayesian model, the ES estimate and its credible interval were obtained directly from the posterior distributions of the parameters. The mathematical expression used to calculate ES at each iteration is given by:

$$ES|\sigma^2, \mathbf{G}, \boldsymbol{\beta}, \mathbf{b}, y = \frac{\beta_2}{\sqrt{\sigma_T^2}} \quad (9)$$

Where $\sigma_T{}^2 = \sigma_b{}^2 + \sigma^2$ for CRT and $\sigma_T{}^2 = \sigma_{b1}{}^2 + \sigma_{b2}{}^2 + \sigma^2$ for MST.

Finally, the posterior probability that the intervention improves the outcome beyond a specific threshold can be used to evaluate the effectiveness of an intervention. For example, probability that $ES \geq 0.1\, SD$. Note that 0.1 SD (standard deviation) can be interpreted as equivalent to at least one month's progress (Higgins *et al*. 2016) in educational attainment. A threshold of 0.1 SD is often the minimum expected impact for an educational intervention. However, the posterior probability estimates were also reported for different threshold values ranging from 0.0 to 1.0. The formula used to obtain the posterior probability (Ntzoufras, 2011) is similar to the posterior predictive probability (Gelman *et al*., 2013; Meng *et al*., 1994), where instead of comparing the observed and replicated data, the comparison occurred between ESs from a model and the pre-specified threshold. Since the posterior probability is defined as the probability that the estimated ES is greater than or equal to a specified threshold given the data and the model (Lesaffre & Lawson, 2012; Yang & Rannala, 2005), it can be mathematically summarised as:

$$P(ES \geq \phi \mid ES, \sigma^2, \mathbf{G}, \boldsymbol{\beta}, \mathbf{b}, y) = \frac{\sum_{i=1}^{K} I\,(ES^{(i)} \geq \phi)}{K} \tag{10}$$

where K is the length of Markov Chain Monte Carlo (MCMC), after excluding the burn-in part, together with those excluded due to the thinning process (K=30000 iterations, 10000 from each chain). Note that $\phi = 0$ will provide evidence that the intervention has a positive effect, whilst $\phi = 0.1$ will provide evidence that the intervention improved educational outcome by at least one month's progress in accordance with EEF's conversion scale (Higgins *et al*. 2016).

All the frequentist analysis was done in R software using 'lme4 package'. Alternatively, the frequentist and Bayesian ES estimate and its CI (or BCI) can also be obtained directly using the 'eefAnalytics' R package (Kasim *et al*., 2017). 'R2jags' an R package, which interfaces with WinBUGS software, was used to obtain Bayesian ES estimates as well as posterior probability estimates. The WinBUGS programme used in the analysis is provided in the appendix A2.

Although our proposed method is based on MCMC, similar results can also be obtained with Stan, which uses the no-U-turn sampler based on Hamiltonian Monte Carlo (HMC). Readers need to be aware that HMC generally explores the posterior parameter space faster and more efficiently than BUGS and JAGS (Hoffman & Gelman, 2014), especially for hierarchical models (Betancourt & Girolami, 2015). For example, Stan Development Team (2017) pointed out that the analysis that BUGS requires 100,000 iterations to converge, in Stan, only 1,000 iterations might be enough. Stan codes equivalent to the WinBUGS programme used for this analysis are provided in the appendix A3.

Results

The distribution of pupils' and schools' participation are summarised in Table 2. In contrast to the school participation in each trial, pupils were not equally distributed in the control and intervention groups. The number of pupils varies from 182 in Project 17 to 25393 in Project 110 and the number of schools varies from 12 in trial 22 to 141 in trial 133. Each school had pupils who received intervention and others who continued with business as usual. Note that for CRT, all pupils from the same school either received the same intervention or continued with business as usual.

'Table 2 here'

Table 3 compares the ES estimates obtained from frequentist and Bayesian methods and Table 4 presents the posterior probabilities. All estimates presented in this table were obtained as specified in the methods section. Since non-informative priors were used in the Bayesian analysis, the point estimates of the ES were similar to the frequentist analysis. It turns out that the estimated CIs from frequentist and BCIs from Bayesian method were similar and a minimal difference was observed for each of the projects (see also (Xiao *et al*., 2016), figure 3). This finding is consistently independent of the magnitude of the ICC. However, this cannot be generalised to all studies but can be said for the CRT and MST studies considered in this study. Beta coefficients for the interventions from the frequentist and Bayesian methods is provided in the appendix Table A1, which were used to estimate ES.

'Table 3 here'

Policymakers and non-academics do not always easily understand using CI or credible intervals. Since a p-value is problematic and somewhat prone to misinterpretation, this study proposes using posterior probability as a metric of confidence in the estimated impact of an intervention. Specifically, it recommends evaluating each intervention on the likelihood that its effect is at least 0.1 standard deviation, which corresponds to one month's progress (Higgins *et al.*, 2016), using EEF's conversion scale. Among the CRT projects, Project 41, with an estimated ES of 0.30 and a credible interval of 0.04 to 0.55 (Table 3), has a posterior probability of 0.94 (Table 4) that the impact of the intervention is at least one month's progress (>= 0.1 SD). Project 110 with an ES of 0.10 (-0.05 0.25) has a posterior probability of 0.51 that the ES is at least 0.1 SD. Project 122 with an effect size of 0.17 (-0.06, 0.40) has 0.72 posterior probability, Project 126 with an ES of 0.21 (-0.04, 0.45) has a posterior probability of 0.80 and project 133 with an ES of 0.02 (-0.19, 0.22) has a 0.22 probability that the intervention improved the attainment outcome by at least one month's progress. It is clear from these results that the larger the ES the greater the probability of at least one month's progress. In this context, the posterior probability is more informative than a p-value and provides a helpful level of confidence to support each result.

Among the MST projects, Project 9 with an ES of 0.28 (0.02, 0.53) has a posterior probability of 0.90 that it improves outcome by at least one month's progress. Project 17 with an ES of 0.13 (-0.15, 0.42) has a 0.60 probability, Project 22 with an ES of 0.07 (-0.18, 0.31) has 0.42 posterior probability, Project 67 with an ES of 0.11 (-0.02, 0.24) has a 0.56 posterior probability, and Project 109 with an ES of -0.06 (-0.29, 0.17) has a posterior probability of only 0.09 that it improves outcome by at least one month's progress. Similar to CRTs, the larger the

ES, the greater the posterior probability of at least one month's progress in the outcome. It is important to note that the posterior probability is conditioned on the current data and it provides confidence for the internal validity of the evaluation effect. This can be interpreted as a focus on 'what worked' instead of 'what works' (Higgins, 2018).

Although this study proposes to focus on the probability that an intervention improves an educational outcome by at least one month's progress, the posterior probability that an intervention has a positive impact at all was reported as well i.e. $P_0(ES > 0)$. Table 4 presents the posterior probabilities for a grid of thresholds ranging from 0 to 1 for all CRT and MST projects. The posterior probability that an intervention has a positive effect on the participants was consistently above 0.90 for the trials with positive ES, except for Project 133 and Project 22. The probabilities, unlike p-values, can be interpreted as evidence of the effectiveness of the interventions. Trials with negative ESs had a posterior probability of less than 0.50, which is to be expected. It is advised against just testing for positive effects or using $P_0(ES > 0)$ because it is rarely the case that one would expect an intervention to have zero impact prior to a trial. As, it is expected that educational interventions improve outcomes for children and young people. Therefore, this study proposes to use $P_{0.1}(ES \geq 0.1)$ or any other appropriate threshold above zero. The posterior probability for an intervention to improve educational outcomes by at least one month's progress ($P_{0.1}(ES \geq 0.1)$) varies from 0.09 for project 109 (Table 4) to 0.94 for project 41 respectively (Table 4) across the ten evaluation analysed in this study. There is a clear relationship between posterior probability and the estimated magnitude of ES.

'Table 4 here'

In addition to the posterior probability estimates, the histogram of the posterior distribution of ES is also included in Figure 1 and Figure 2. These figures provide useful information about how to obtain posterior probability. For instance, for project 41, $P_{0.1}(ES \geq 0.1) = 1 - (1 + 8 + 59 + 350 + 1570)/30000$. Here 30000 in denominator is the number of iterations used to estimate posterior parameters.

Impact of prior on posterior probability

As mentioned in the introductory section, the posterior distribution is obtained from the prior distribution and the likelihood (data), which is one of the strengths of the Bayesian method. A prior reflects knowledge or understanding of the parameters regardless of actual data. However, informative priors are often not available, as in this study. In this case, AVCI, (2017) recommended the use of non-informative priors, as the use of an incorrect 'informative' prior may wrongly influence inferences and subsequent decisions (Morita, Thall, & Müller, 2010).

In this section, the aim was to investigate the effect of different priors on the posterior probabilities in education trials. First, their impact was evaluated on the posterior distribution of ES, which is the active ingredient in the estimation of the posterior probability. To do this, different hypothetical priors for the simulated dataset were considered. Assuming a pre-test variable follows a normal distribution with a mean of 28.16 and 5.52 as variance, two treatment groups each with 30 schools and 20 pupils per school, the post-test outcome was obtained as

$$y_{is} = -1.89 - 2.52 * pretest + 1.85 * t + b_s + \varepsilon_{is}.$$

Where $\varepsilon_{is} \sim N(0,5.92^2)$, $b_s \sim N(0,1.77^2)$.

Applying the ANCOVA model specified in equation 1, a likelihood distribution of ES was obtained with a mean of 0.40 and a variance of 0.0032 ($ES \sim N(0.40,0.0032)$).

It was assumed that no ES prior information was available by using a non-informative Gaussian prior $ES \sim N(0, 10^6)$, which is also used in the main analysis of this paper. In addition, we have also used $ES \sim N(0, 10^2)$ and $ES \sim N(0, 10^3)$ as non-informative priors These priors are very different from the likelihood of the ES but they do not have the power to shift the posterior distribution from the likelihood in the same way as shown in the first row of Figure 3 (see panel a-c). On the other hand, when prior information is available, the posterior distribution is effectively the compromise between the prior and the likelihood. In statistics, it is almost impossible to have full information about a parameter under study, so variance or precision is used to express the extent of unknown information. In this light, the term less informative, informative, and very informative priors was used in terms of the magnitude of variances that are respectively greater, equal, and less than that of likelihood. Further, we also shifted the mean to 0.25 for more clarity on the effect of priors on the likelihood. For the less informative prior, it was assumed that $ES \sim N(0.25{,}0.0032 * 10)$, the likelihood has greater impact on posterior distribution as observed in Figure 3 panel d. For the informative prior, it was assumed that $ES \sim N(0.25{,}0.0032)$, so both the likelihood and the prior has the same impact on posterior distribution (Figure 3 panel e). Finally, the very informative prior with the smallest variance (=0.00032/5) and mean 0.25 has a greater impact on posterior distribution than likelihood (Figure 3 panel f).

'Figure 1 here'

Knowing that the ES estimate, is the main parameter of interest for estimating posterior probability, which is sensitive to the choice of the prior, the posterior probability estimate is therefore also sensitive to the choice of prior. Yang and Rannala (2005) had similar thoughts in estimating the posterior probability of phylogeny. Table 5 shows how the choice of the prior and use of a different distribution for variance parameters and different threshold values in a MST trial modified the posterior probability estimates. It is worth knowing that when a prior is very informative relative to the likelihood, it has a high impact on posterior probabilities. In other words, the more informative the prior is, the less the variation in the posterior probability estimates. As the data plays only a little role in shifting the posterior from the prior distribution (the prior remains fixed as the iteration changes). It was assumed in this study that in an MST trial, the covariance parameter for the random intercept and slope component is independent, i.e., $\sigma_{b1,b2} = 0$. However, if one assumes that $\sigma_{b1,b2} \neq 0$, then in that case, Inverse Wishart distribution will be the natural choice for variance and covariance parameters (Congdon, 2006), in this study $IW\left(\begin{bmatrix}1 & 0\\0 & 1\end{bmatrix}, 2\right)$ was used. The first two columns of Table 5 provides the posterior probability estimates assuming prespecified Inverse Wishart and gamma distribution for the variance and covariance parameters. It is apparent that estimates of posterior probability do not vary much between the Inverse Wishart and gamma priors and some marginal differences were only observed for the threshold values between 0.3-0.5.

‘Table 5 here’

Discussion and Conclusion

The main mission of the Education Endowment Foundation (EEF) is to improve educational attainment, especially the attainment of disadvantaged pupils in the England. With this objective in mind, the EEF has commissioned over 150 evaluations with distinct educational interventions since 2011. The effectiveness of an intervention is typically assessed through the ES (standardised mean difference) with its associated confidence (or credible) interval. This study have demonstrated how Bayesian posterior probability can be used to provide evidence of the effectiveness of an intervention. The proposed application of posterior probability may be more easily communicated to the policymakers and education stakeholders.

‘Figure 2 here’

The Bayesian framework provides a natural way to specify models, estimate parameters, and draw inferences even in cases where classical statistical methods fail (Swaminathan, Rogers, & Horner, 2014). To reduce educational researchers’ dependence on p-values, a CI for the ES is often suggested as an alternative. However, a CI is based on the same frequentist assumptions as a p-value threshold, which means that they are also prone to the same fallacies and misinterpretations. Further, CI presume that the effect under consideration exists in a wider population and their use implies that every problem of inference is a problem of parameter estimation rather than hypothesis testing (Wagenmakers *et al*., 2018). With the Bayesian approach, obtaining the analogous ‘confidence’ interval for an ES (i.e. its credible interval) is straightforward, as each parameter in Bayesian model follows a distribution, including the ES.

‘Figure 3 here’

Although this study shows that the frequentist CI and BCI usually do not differ, this finding cannot be generalised to studies not considered in this paper, especially when the priors used are not relatively flat (non-informative) or models are not regular. Overall, Bayesian credible intervals around an ES are a more reliable estimate of impact as outlined in the argument above. Posterior probabilities estimated from a Bayesian analysis can provide estimates of the probability of an intervention’s effect being above a specific threshold or even the probability that the intervention’s impact might lie between a range of specific values. These answers cannot be obtained from the traditional frequentist framework (Wagenmakers, Morey, & Lee, 2016) but could be useful in decision making by education stakeholders and policymakers.

A comparison of the posterior probability given a similar threshold for different interventions can, therefore, be a useful way to identify which intervention is more effective. This is evident from the findings presented in this analysis. The posterior probabilities that the ESs are above 0.10 for EEF Project 41 is 0.94 and for Project 9 is 0.90. These are much higher than the posterior probabilities for other projects with a similar threshold. This suggests that the interventions in Project 41 and 9 are rather more effective than the other interventions in this study at this specific threshold. They might therefore be a ’better bet’ for other schools to try (Higgins, 2018).

According to the different factors involved (including the intervention, the targeted pupils, and the outcome of interest), researchers can specify the value of threshold differently according to Hill *et al*. (2008). There is a long debate about the use of specific thresholds to quantify the impact of an intervention in education, such as a small, medium, or large effect (Cohen, 1988; Hedges & Hedberg, 2007; Lipsey *et al.*, 2012). However, Glass, Smith, and McGaw (1981) suggest that for education a small effect of 0.1, can be considered as important, particularly if it is cheap to implement or reliable (Higgins, 2018). In this study, all the possible ES thresholds for effective interventions ranging from 0.0 to 1.0 were accommodated. Estimates of posterior probabilities for such a wide range of thresholds might better empower the educational stakeholders and policymakers to understand the change in the effectiveness of an intervention for a specific threshold. Bayesian posterior probabilities can also be used to provide odds for P[ES>0.1]/P[ES<0], which is relevant for decision making along with P[ES>0.1]. These odds can also be estimated for different threshold values like 0.2 or 0.3. However, the estimation of such odds was beyond the scope of this study but can be explored in the future.

This would help with cost/benefit analyses and might provide realistic goals for policy changes. Based on this analysis, and subject to further exploration, it is recommended using 0.1 to evidence the effectiveness of an educational intervention, if a threshold must be chosen.

Relying just on an ES to determine the practical significance of interventions in education can be problematic, as is using p-values to determine statistical significance (Pogrow, 2019). There is a need for educational researchers to move towards a simpler measure of practical benefit that can estimate the likely benefit of an intervention, based on how effective it has been in a specific evaluation. This study is making an important contribution in this direction by proposing a Bayesian approach to evaluate an intervention's effectiveness and by providing a range of probability estimates given the observed data. Since the estimates for posterior probability and Bayesian credible intervals are the estimates given the observed data, the results can safely be said to focus on internal validity and what's worked in a specific trial.

However, the choice of prior in Bayesian analysis required careful consideration. This study has shown the dependence of the posterior probability on the choice of priors in simulated data. Therefore, too informative prior will have a strong influence on the posterior, except when mixed with high-powered data in terms of sample size (Lemoine, 2019; Ley, Reinert, & Swan, 2017). If the researchers are confident about the accuracy of their knowledge of prior, then informative priors can be integrated with the empirical data to estimate posterior probabilities. This study recommends using vague prior for Bayesian evaluation of education trials so that any conclusion about the effectiveness of an intervention is largely driven by the data instead of the researcher's subjective prior knowledge.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

AVCI, E. (2017). Using informative prior from metaanalysis in Bayesian approach. *Journal of Data Science*, *15*(4).

Barrado, L. G., Coart, E., & Burzykowski, T. (2019). A Bayesian Framework Allowing Incorporation of Retrospective Information in Prospective Diagnostic Biomarker-Validation Designs. Statistics in Biopharmaceutical Research, 11(3), 311-323.

Beaumont, M. A., & Rannala, B. (2004). The Bayesian revolution in genetics. *Nature Reviews Genetics*, *5*(4), 251-261.

Betancourt, M., & Girolami, M. (2015). Hamiltonian Monte Carlo for hierarchical models. *Current trends in Bayesian methodology with applications, 79(30), 2-4.*

Brooks, S. P., & Gelman, A. (1998). General methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, 7, 434–455.

Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... & Altmejd, A. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour, 2*(9), 637–644.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed). *New Jersey: Laurence Erlbaum Associates, Publishers, Hillsdale*.

Congdon, P. (2006). Bayesian statistical modelling (2nd ed). John Wiley & Sons.

Congdon, P. (2014). Applied Bayesian modelling (Vol. 595). John Wiley & Sons.

Congdon, P. D. (2019). Bayesian Hierarchical Models: With Applications Using R (2nd). *CRC Press.*

Cumming, G. (2008). Replication and p intervals: p values predict the future only vaguely, but confidence intervals do much better. *Perspectives on Psychological Science*, *3*(4), 286–300.

Cummings, M. P., Handley, S. A., Myers, D. S., Reed, D. L., Rokas, A., & Winka, K. (2003). Comparing bootstrap and posterior probability values in the four-taxon case. *Systematic Biology*, *52*(4), 477–487.

Depaoli, S., Clifton, J. P., & Cobb, P. R. (2016). Just another Gibbs sampler (JAGS) flexible software for MCMC implementation. *Journal of Educational and Behavioral Statistics*, 41(6), 628–649.

EEF. (2015). *Policy analysis on EEF evaluations*. London: Education Endowment Foundation.

EEF. (2019). *Classification of the security of findings from EEF evaluations*. London: Education Endowment Foundation.

Feaster, D. J., Mikulich-Gilbertson, S., & Brincks, A. M. (2011). Modeling site effects in the design and analysis of multi-site trials. *The American Journal of Drug and Alcohol Abuse*, *37*(5), 383–391.

Friston, K., & Penny, W. (2003). Posterior probability maps and spms. *Neuroimage*, *19*(3), 1240–1249.

Gelfand, A. E., Sahu, S. K., & Carlin, B. P. (1995). Efficient parametrisations for normal linear mixed models. *Biometrika*, *82*(3), 479–488.

Gelfand, A. E., & Sahu, S. K. (1999). Identifiability, improper priors, and Gibbs sampling for generalized linear models. *Journal of the American Statistical Association*, *94*(445), 247–253.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian data analysis* (3rd ed.). Chapman and Hall/CRC.

Glass, G. V., Smith, M. L., & McGaw, B. (1981). Meta-analysis in social research. Sage Publications, Incorporated.

Goodman, S. N. (2019). Why is getting rid of p-values so hard? musings on science and statistics. *The American Statistician*, *73*(sup1), 26–30.

Hahn, U. (2014). The Bayesian boom: good thing or bad?. *Frontiers in Psychology*, *5*, 765.

Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature Methods*, *12*(3), 179.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, 32(4), 341–370.

Hedges, L. V. (2008). What are effect sizes and why do we need them? *Child Development Perspectives*, *2*(3), 167–171.

Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group randomized trials in education. *Educational Evaluation and Policy Analysis*, *29*(1), 60–87.

Higgins, S. (2018). *Improving learning: Meta-analysis of intervention research in education*. Cambridge: Cambridge University Press.

Higgins, S., Katsipataki, M., Villanueva-Aguilera, A., Coleman, R., Henderson, P., Major, L., ... Mason, D. (2016). *The sutton trust-education endowment foundation teaching and learning toolkit.* London: Education Endowment Foundation.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives*, *2*(3), 172–177.

Hoffman, M. D., & Gelman, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research, 15(1), 1593-1623.*

Kasim, A., Xiao, Z., Higgins, S., & Troyer, E. D. (2017). eefanalytics: Analysing education trials [Computer software manual]. (R package version 1.0.6).

König, C., & van de Schoot, R. (2018). Bayesian statistics in educational research: A look at the current state of affairs. *Educational Review*, *70*(4), 486–509.

Kruschke, J. K., & Liddell, T. M. (2018). The Bayesian New Statistics: Hypothesis testing, estimation, meta-analysis, and planning from a Bayesian perspective. *Psychonomic Bulletin and Review,* 25,178–206.

Lemoine, N. P. (2019). Moving beyond noninformative priors: why and how to choose weakly informative priors in Bayesian analyses. *Oikos*, *128(7), 912-928.*

Lesaffre, E., & Lawson, A. B. (2012). *Bayesian biostatistics*. John Wiley & Sons.

Ley, C., Reinert, G., & Swan, Y. (2017). Distances between nested densities and a measure of the impact of the prior in Bayesian statistics. *The Annals of Applied Probability*, *27*(1), 216-241.

Lipsey, M. W., Puzio, K., Yun, C., Hebert, M. A., Steinka-Fry, K., Cole, M. W., ... Busick, M. D. (2012). Translating the statistical representation of the effects of education interventions into more readily interpretable forms. *National Center for Special Education Research*.

Maxwell, B., Connolly, P., Demack, S., O'Hare, L., Stevens, A., Clague, L., & Stiell, B. (2014). *Summer active reading programme: evaluation report and executive summary*. London: Education Endowment Foundation.

McShane, B., Gal, D., Gelman, A., Robert, C., & Tackett, J. (2019). Abandon statistical significance. *The American Statistician*, *73*(sup1), 235–245.

Meng, X.-L., *et al*. (1994). Posterior predictive *p*-values. *The Annals of Statistics*, *22*(3),

1142–1160.

Miller, S., Davison, J., Yohanis, J., Sloan, S., Gildea, A., & Thurston, A. (2017). *Texting parents: Evaluation report and executive summary.* London: Education Endowment Foundation.

Morita, S., Thall, P. F., & Müller, P. (2010). Evaluating the impact of prior assumptions in Bayesian biostatistics. *Statistics in Biosciences*, *2*(1), 1-17.

Nakagawa, S., & Cuthill, I. C. (2007). Effect size, confidence interval and statistical significance: a practical guide for biologists. *Biological Reviews*, *82*(4), 591–605.

Ntzoufras, I. (2011). *Bayesian modeling using winbugs* (Vol. 698). John Wiley & Sons.

Nunes, T., Barros, R., Evangelou, M., Strand, S., Mathers, S., & Sanders-Ellis, D. (2018). *1stclass@ number evaluation report and executive summary*. London: Education Endowment Foundation.

Nuzzo, R. (2014). Statistical errors: P values, the 'gold standard' of statistical validity, are not as reliable as many scientists assume. *Nature*, *506*(7487), 150–153.

Pogrow, S. (2019). How effect size (practical significance) misleads clinical practice: The case for switching to practical benefit to assess applied research findings. *The American Statistician*, *73* (sup1), 223–234.

Raftery, A. E., & Lewis, S. M. (1995). The number of iterations, convergence diagnostics and generic Metropolis algorithms. *Practical Markov Chain Monte Carlo*, 7(98), 763-773.

Rutt, S., Easton, C., & Stacey, O. (2014). *Catch up numeracy: Evaluation report and executive summary*. London: Education Endowment Foundation.

Rutt, S., Roy, P., Buchanan, E., Rennie, C., Martin, K., & Fiona, W. (2019). *Catch up literacy (re-grant): Evaluation report and executive summary.* London: Education Endowment Foundation.

Speckesser, S., Runge, J., Foliano, F., Bursnall, M., Hudson-Sharp, N., Rolfe, H., & Anders, J. (2018). *Embedding formative assessment: evaluation report and executive summary*. London: Education Endowment Foundation.

Stan Development Team (2017) Modeling Language User’s Guide and Reference Manual, Version 2.17.0. https://mc-stan.org/users/documentation/

Styles, B., Stevens, E., Bradshaw, S., & Clarkson, R. (2014). *Vocabulary enrichment intervention programme: Evaluation report and executive summary*. London: Education Endowment Foundation.

Swaminathan, H., Rogers, H. J., & Horner, R. H. (2014). An effect size measure and Bayesian analysis of single-case designs. *Journal of School Psychology*, *52*(2), 213–230.

Thompson, B. (2002). What future quantitative social science research could look like: Confidence intervals for effect sizes. *Educational Researcher*, *31*(3), 25–32.

Thompson, B. (2007). Effect sizes, confidence intervals, and confidence intervals for effect sizes. *Psychology in the Schools*, *44*(5), 423–432.

Torgerson, C., Bell, K., Coleman, E., Elliott, L., Fairhurst, C., Gascoine, L., ... Torgerson, D. (2018). *Tutor trust: affordable primary tuition. evaluation report and executive summary*. London: Education Endowment Foundation.

Trafimow, D., & Marks, M. (2015). Editorial. *Basic and Applied Social. Psychology, 37(1)*, 1–2.

Valentine, J. C., & Cooper, H. (2003). Effect size substantive interpretation guidelines: Issues in the interpretation of effect sizes. *Washington, DC: What Works Clearinghouse*, 1–7.

Van Ravenzwaaij, D., Cassey, P., & Brown, S. D. (2018). A simple introduction to Markov Chain Monte–Carlo sampling. Psychonomic bulletin & review, 25(1), 143-154.

Verbeke, G., & Molenberghs, G. (2009). *Linear mixed models for longitudinal data*. Springer Science & Business Media.

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., ... others (2018). Bayesian inference for psychology. part i: Theoretical advantages and practical ramifications. *Psychonomic Bulletin & Review*, *25*(1), 35–57.

Wagenmakers, E.-J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science*, *25*(3), 169–176.

Wang, L., Zhang, Z., McArdle, J. J., & Salthouse, T. A. (2008). Investigating ceiling effects in longitudinal data analysis. Multivariate behavioral research, 43(3), 476-496.

Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond "p< 0.05". *The American Statistician*, *73*(sup1), 1–19.

Worth, J., Nelson, J., Harland, J., Bernardinelli, D., & Styles, B. (2018). *Graphogame rime: Evaluation report and executive summary.* London: Education Endowment Foundation.

Worth, J., Sizmur, J., Ager, R., & Styles, B. (2015). *Improving numeracy and literacy: Evaluation report and executive summary*. London: Education Endowment Foundation.

Xiao, Z., Higgins, S., & Kasim, A. (2017). An empirical unraveling of lord’s paradox. *The Journal of Experimental Education*, 1–16.

Xiao, Z., Kasim, A., & Higgins, S. (2016). Same difference? understanding variation in the estimation of effect sizes from educational trials. *International Journal of Educational Research*, *77*, 1–14.

Yang, Z., & Rannala, B. (2005). Branch-length prior influences Bayesian posterior probability of phylogeny. *Systematic Biology*, 54(3), 455-470.

Ying, L., & Belitskaya-Levy, I. (2015). The debate about p-values. *Shanghai Archives of Psychiatry*, *27*(6), 381.

Table 1

Description of the projects used in this study.

Design	Full EEF title (project number)	Outcome	Padlock
CRT	Improving Numeracy and Literacy (41)	Maths	5
	Embeddive formative assessment (110)	Reading	5
	1stClass@Number (122)	Maths	4
	Tutor Trust Primary (126)	Maths	4
	Catch Up® Literacy (133)	Reading	4
MST	Catch Up® Numeracy (9)	Maths	3
	Summer Active Reading (17)	Reading	3
	Vocabulary Enrichment (22)	Reading	4
	Texting Parents (67)	Maths	3
	Graphorime (109)	Reading	5

Table 2:

Number of pupils and schools participated in each project.

		Control		**Intervention**		**Overall**	
	Project	Pupils	Schools	Pupils	Schools	Pupils	Schools
CRT	**41**	848	19	517	17	1365	36
	110	13035	70	12358	70	25393	140
	122	227	62	239	67	466	129
	126	634	52	567	50	1201	102
	133	505	72	501	69	1006	141
MST	**9**	108	54	108	54	216	54
	17	89	42	93	41	182	48
	22	288	12	282	12	570	12
	67	5977	28	5613	29	11590	29
	109	177	14	185	14	362	14

Table 3:

Comparison of effect size estimates from frequentist and Bayesian methods.

		Frequentist			**Bayesian**			
	Trial	Estimate	95% LB	95% UB	Estimate	95% LB	95% UB	ICC
CRT	**41**	0.30	0.04	0.56	0.30	0.04	0.55	0.13
	110	0.11	-0.04	0.25	0.10	-0.05	0.25	0.20
	122	0.17	-0.06	0.39	0.17	-0.06	0.40	0.20
	123	0.21	-0.03	0.44	0.21	-0.04	0.45	0.30
	133	0.02	-0.19	0.23	0.02	-0.19	0.22	0.28
MST	**9**	0.28	0.03	0.52	0.28	0.02	0.53	0.14
	17	0.14	-0.14	0.41	0.13	-0.15	0.42	0.11
	22	0.07	-0.17	0.32	0.07	-0.18	0.31	0.05
	67	0.11	-0.02	0.24	0.11	-0.02	0.24	0.07
	109	-0.06	-0.30	0.17	-0.06	-0.29	0.17	0.05

Table 4:

Posterior probability estimates for different threshold values in both cluster randomised and multisite trials projects.

Cut-off	**CRT projects**					**MST projects**				
(φ)	**41**	**110**	**122**	**126**	**133**	**9**	**17**	**22**	**67**	**109**
0.0	0.98	0.92	0.92	0.95	0.57	0.98	0.82	0.74	0.94	0.31
0.1	0.94	0.51	0.72	0.80	0.22	0.90	0.60	0.42	0.56	0.09
0.2	0.77	0.11	0.39	0.53	0.04	0.72	0.32	0.14	0.08	0.02
0.3	0.49	0.01	0.13	0.22	0.00	0.43	0.12	0.03	0.00	0.00
0.4	0.22	0.00	0.02	0.06	0.00	0.18	0.03	0.00	0.00	0.00
0.5	0.07	0.00	0.00	0.01	0.00	0.04	0.01	0.00	0.00	0.00
0.6	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.7	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.8	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
0.9	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
1.0	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 5:

Posterior Probability estimates from simulated data with different priors, and posterior probability estimates from a MST model with vague prior and assuming that correlation between both random effect parameters exist.

Cut-off	**Vague**	**Vague**	**Less**	**Informative**	**Very**
(φ)	**Wishart***	**Gamma****	**informative**		**informative**
0.0	1.00	1.00	1.00	1.00	1.00
0.1	1.00	1.00	1.00	1.00	1.00
0.2	0.98	0.99	0.98	0.97	1.00
0.3	0.82	0.87	0.81	0.45	0.05
0.4	0.39	0.50	0.36	0.01	0.00
0.5	0.07	0.13	0.05	0.00	0.00
0.6	0.00	0.01	0.00	0.00	0.00
0.7	0.00	0.00	0.00	0.00	0.00
0.8	0.00	0.00	0.00	0.00	0.00
0.9	0.00	0.00	0.00	0.00	0.00
1.0	0.00	0.00	0.00	0.00	0.00

*Wishart: generalization to multiple dimensions of the gamma distribution and Off diagonal elements of variance covariance matrix are not 0.

**Gamma: random intercept and slope are independent.

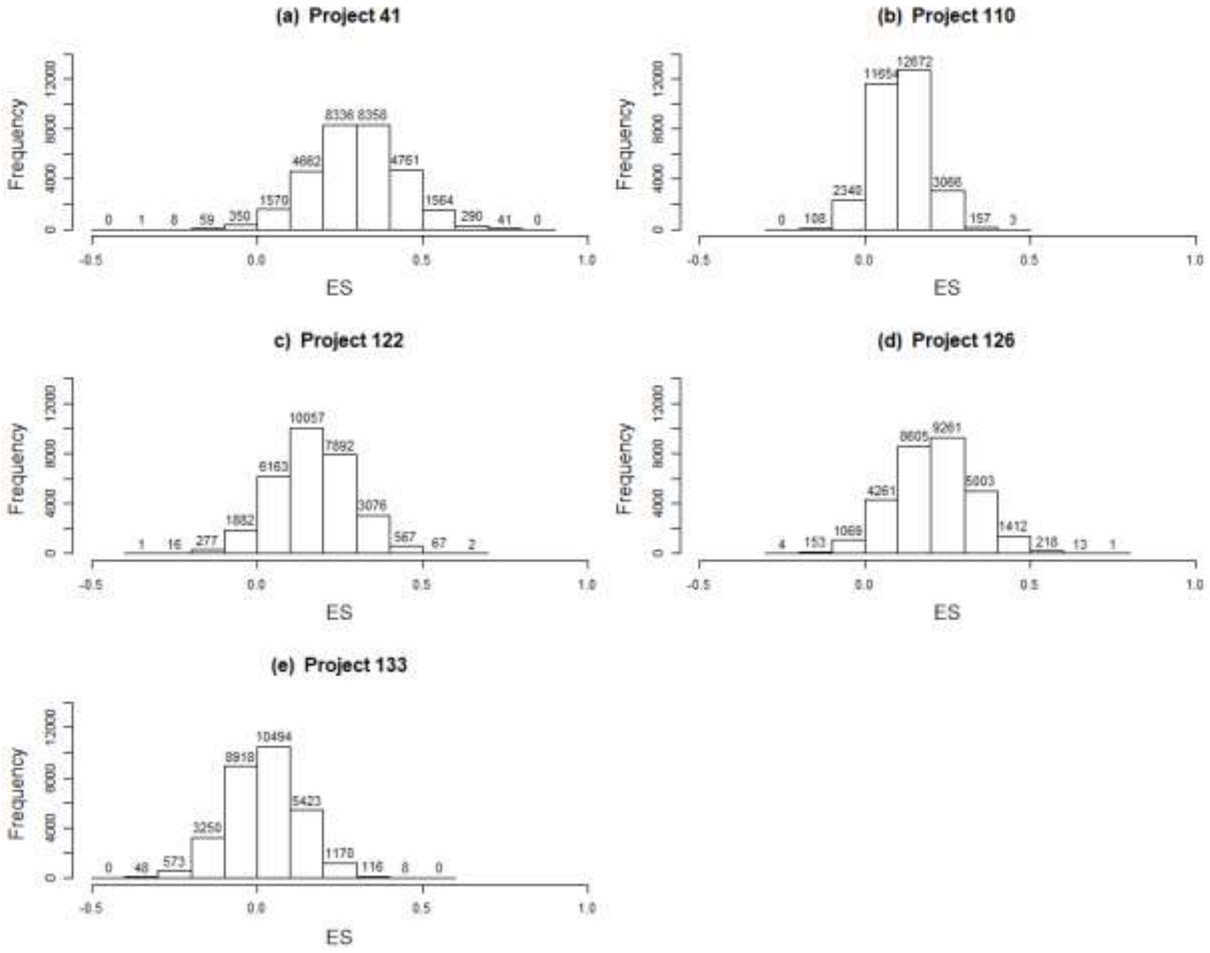


Figure 1: Plot of posterior distribution of effect size (ES) from CRT studies based on 30000 iterations.

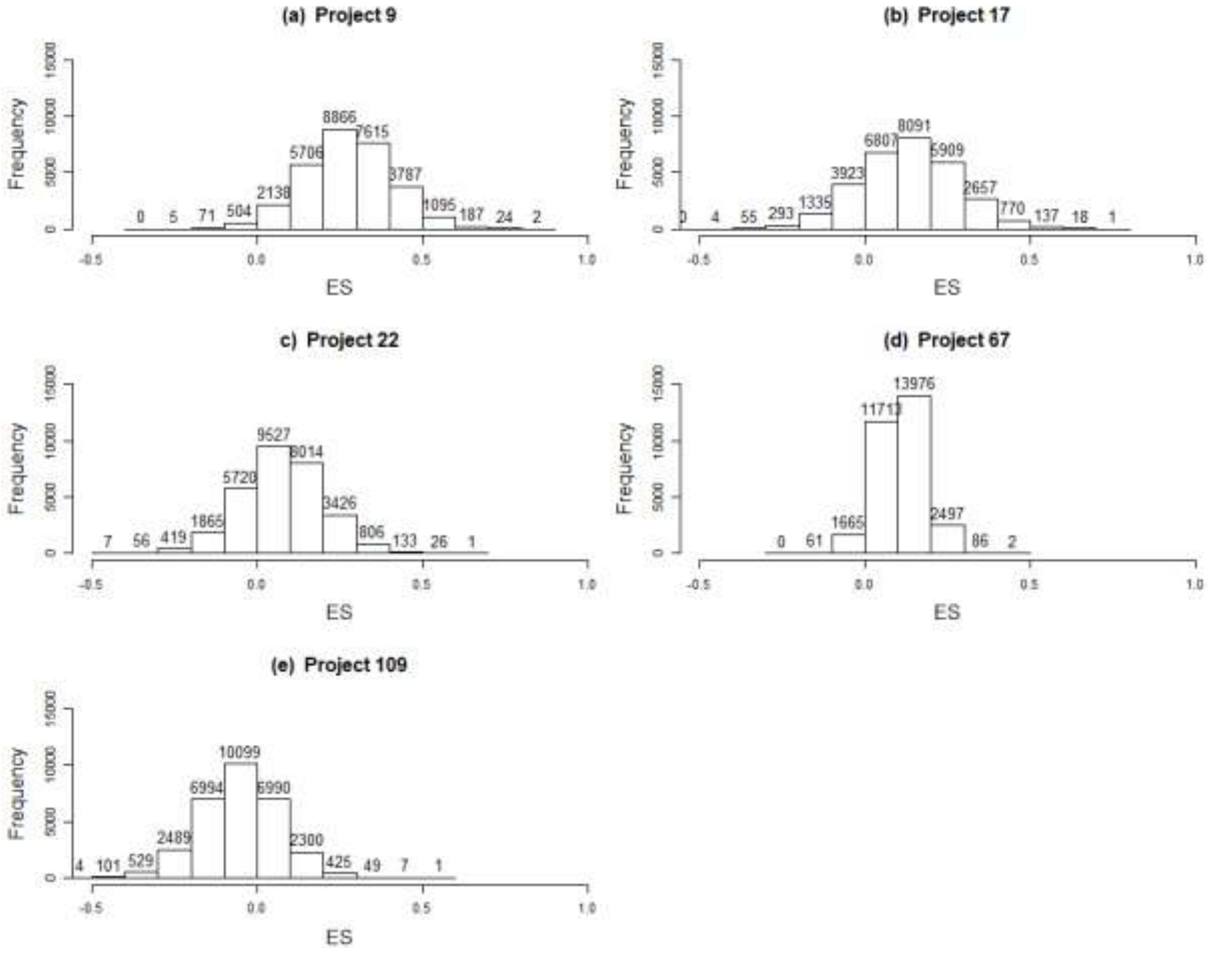


Figure 2: Plot of posterior distribution of effect size (ES) from MST studies based on 30000 iterations.

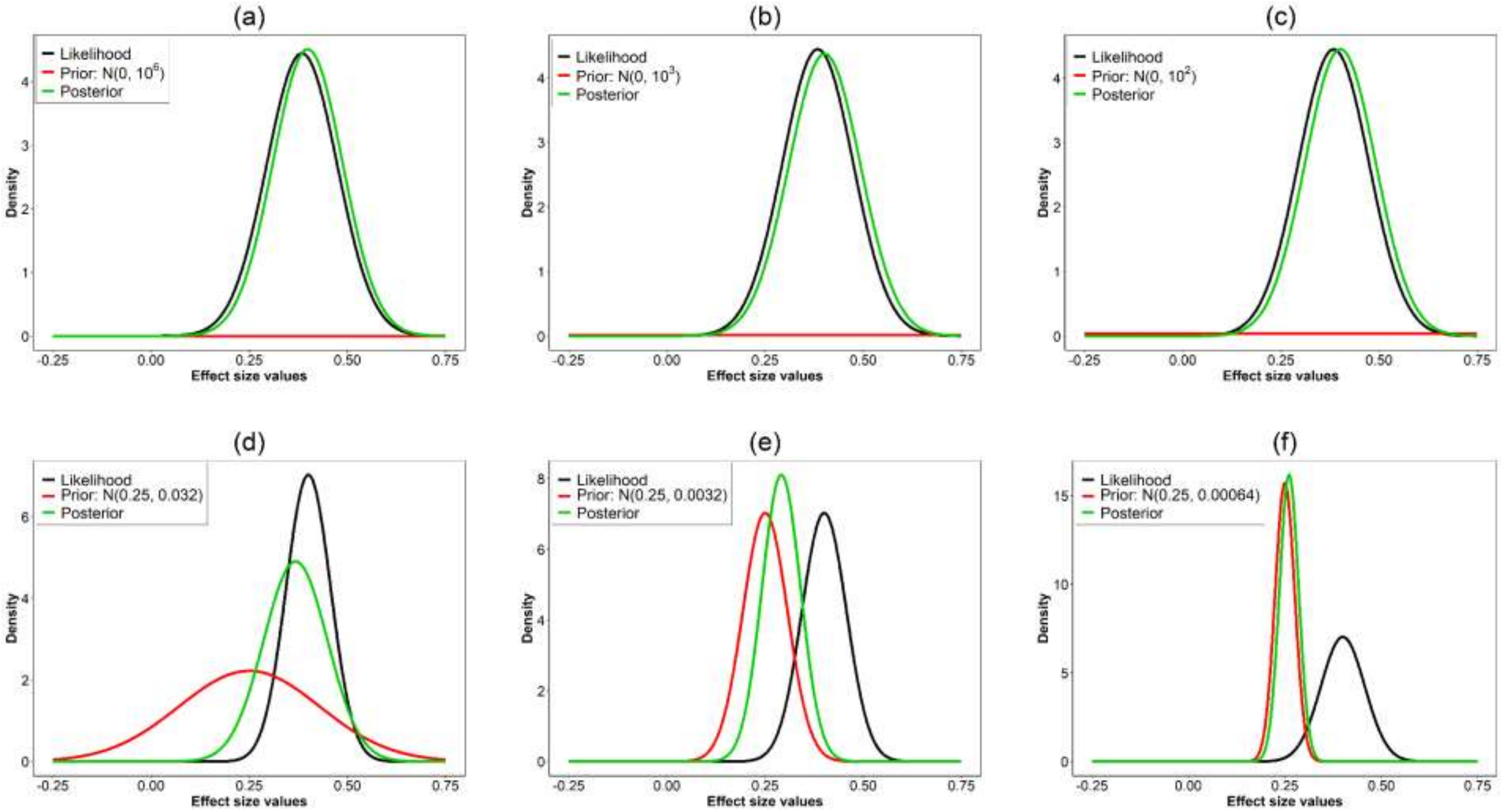


Figure 3: Prior, likelihood and posterior distributions of effect size. Panel a)-c) Non informative, d) less informative ES~N(0.25,0.032), e) Informative prior ES~N(0.25,0.0032) and f) very-informative prior ES~N(0.25, 0.00064).

List of Figures

Appendix

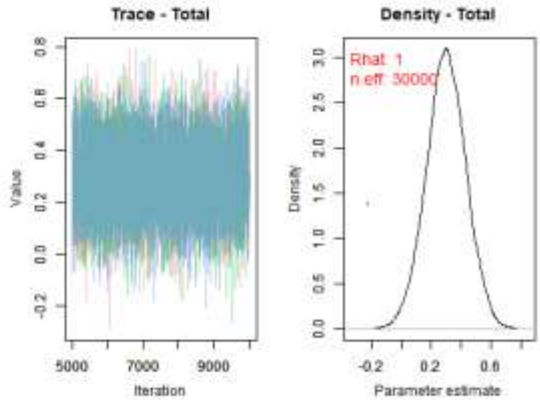
Table A1

Beta coefficients of intervention from archive frequentist and Bayesian analysis.

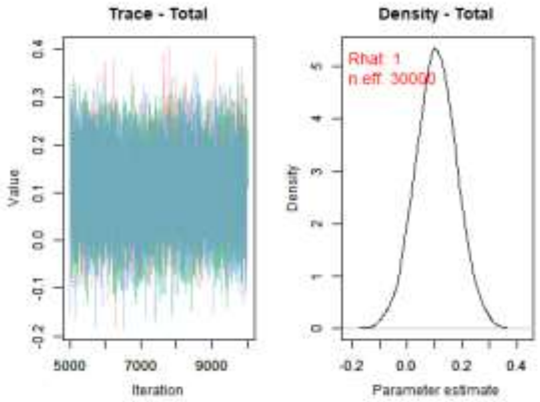
		Frequentist			**Bayesian**		
	Project	Estimate	95% LB	95% UB	Estimate	95% LB	95% UB
CRT	**41**	1.05	0.15	1.95	1.04	0.13	1.94
	110	0.10	-0.04	0.23	0.10	-0.04	0.24
	122	0.73	-0.25	1.71	0.74	-0.28	1.75
	126	1.04	-0.13	2.21	1.03	-0.18	2.21
	133	0.15	-1.56	1.85	0.14	-1.58	1.82
MST	**9**	2.92	0.34	5.50	2.91	0.22	5.55
	17	1.22	-1.25	3.70	1.22	-1.34	3.85
	22	0.38	-0.90	1.64	0.38	-0.94	1.63
	67	0.07	-0.01	0.14	0.07	-0.01	0.14
	109	-0.37	-1.93	1.08	-0.36	-1.85	1.11

Note: These are beta coefficients for intervention, which were used to calculate effect size and confidence/credible intervals.

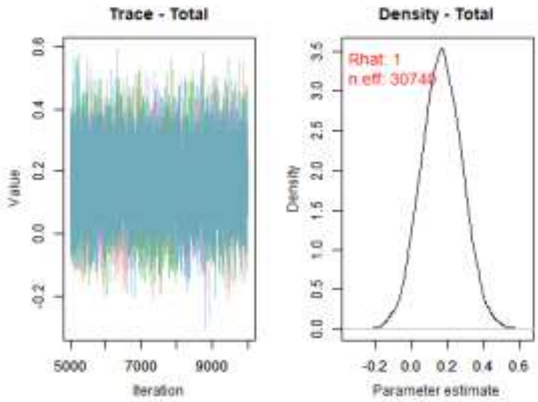
(a) Project 41



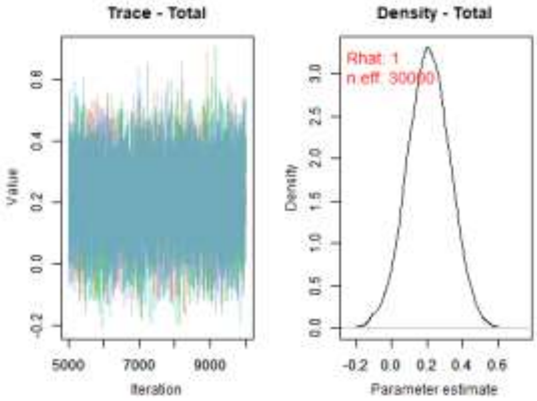
(b) project 110



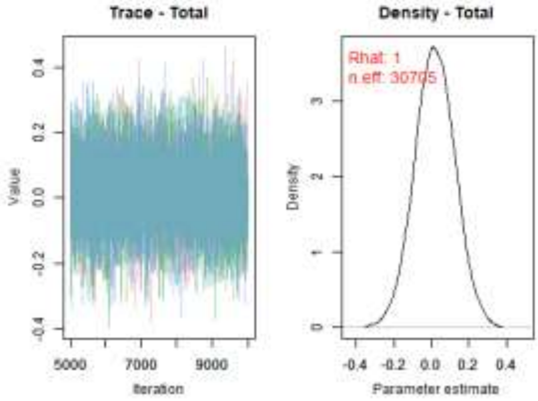
(c) Project 122

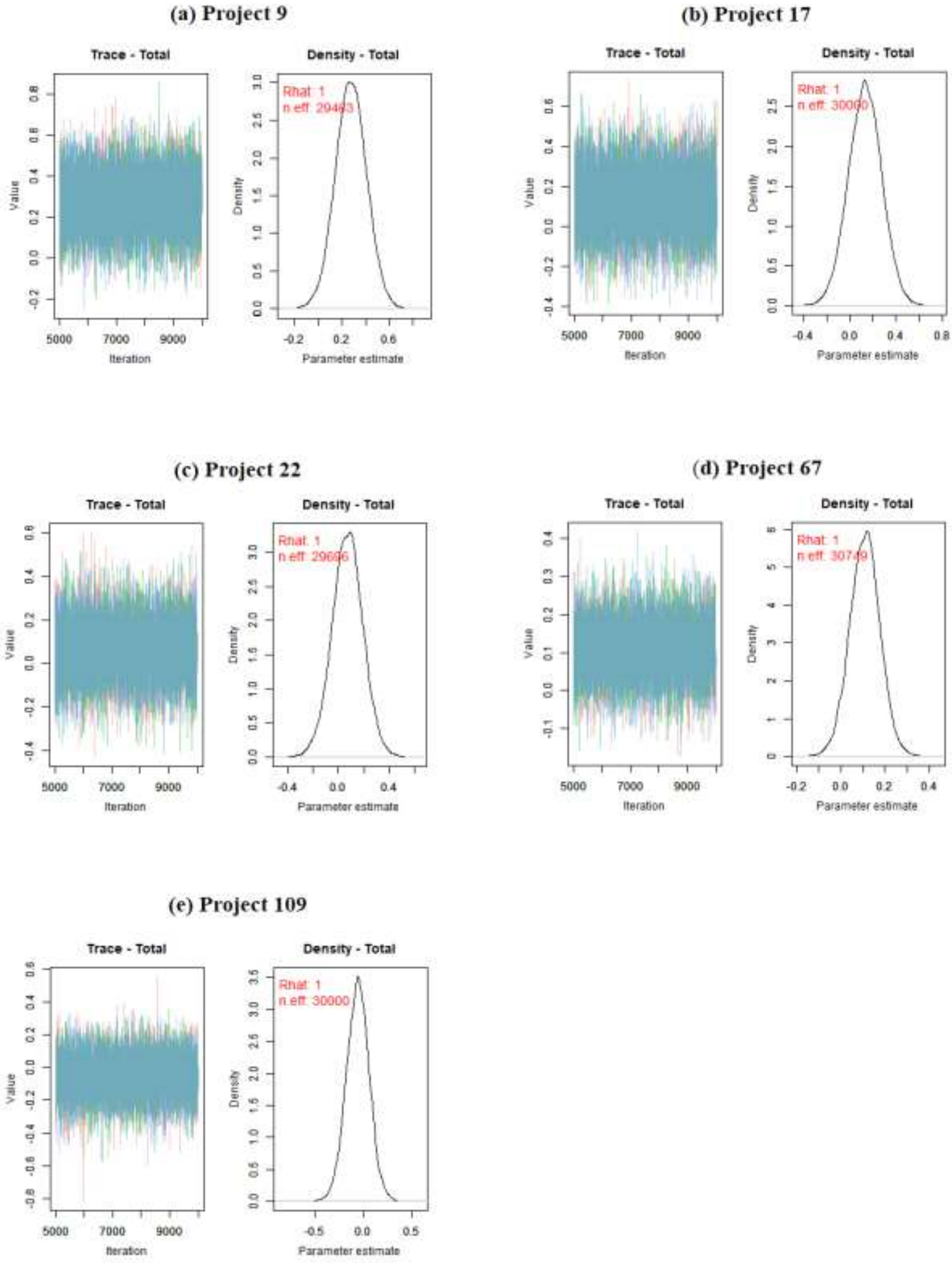


(d) Project 126



(e) Project 133





(a) Project 9

(b) Project 17

(c) Project 22

(d) Project 67

(e) Project 109

Appendix A1: Formulation of Multilevel model (MLM)

The ES estimation strategy starts from a multilevel model (MLM) where the school is defined at the highest level of the model (level 2) and the pupil within the school at the lowest level (level 1).

1. **CRT**

Schools are randomised to receive treatment or control

Level 1: $\mathrm{Post}_{ij} = \mathrm{mu}_i + \varepsilon_{ij}$,
Level 2: $\mathrm{mu}_i = \beta_0 + \beta_2 T_{ij} + b_i$,

Level 1 and 2 combined with important covariate (pre-test)

$$\mathrm{Post}_{ij} = \beta_0 + \boldsymbol{\beta_1 \mathrm{pret}_{ij}} + \beta_2 T_{ij} + b_i + \varepsilon_{ij}.$$

2. **MST**

Pupils within school are randomised to receive treatment or control

Level 1: $\mathrm{Post}_{ij} = \mathrm{mu}_{0i} + \mathrm{mu}_{1i} T_{ij} + \varepsilon_{ij}$
Level 2: $\mathrm{mu}_{0i} = \beta_0 + b_{1i}$,
$\mathrm{mu}_{1i} = \beta_2 + b_{2i}$,

Level 1 and 2 combined with important covariate (pre-test)

$$\mathrm{Post}_{ij} = \beta_0 + \boldsymbol{\beta_1 \mathrm{pret}_{ij}} + \beta_2 T_{ij} + b1_i + b_{2i} T_{ij} + \varepsilon_{ij}$$

Since according to the MST and CRT study designs, the pupil's attainments from the same school are assumed to be correlated while those from different schools are independent. So, we assumed a distribution(s) for the school deviation parameters $\mathbf{b_i}$ for CRT and $\mathbf{b1_i}$ for MST. In addition, we included the parameter that can help to estimate the school by treatment effects ($\mathbf{b2_i}$) only for MST studies. Since pupils within these schools were randomized to receive either treatment or control. Whereas for the CRT studies, it is not possible to estimate such parameter since the whole school is randomized to receive either treatment or control.

Appendix A2: WinBUGS programme for Bayesian multilevel model and posterior probability

1. CRT

```
model{
# 1. ANCOVA model
#------------------------
 for(i in 1:N){
   y[i] ~dnorm(mu[i],tau)
   mu[i] <- beta[1] + beta[2]*pret[i] + beta[3]*t[i] + b1[school[i]]
 }
 for(j in 1:M){
   b1[j]~dnorm(0.0,tau.b1)
 }
# 2. Priors
#------------------------
 tau.b1~dgamma(0.0001,0.0001)
 sigma.b1<-1/tau.b1
 tau~dgamma(0.0001,0.0001)
 sigma<-1/tau
 for(k in 1:p){beta[k]~dnorm(0.0,1.0E-06)}

# 3. ES, ICC and Total Variance
#---------------------------------------
 sigma.tt <-sigma + sigma.b1
 icc <- sigma.b1 * pow(sigma.tt ,-1)
 ES <- beta[3]/sqrt(sigma.tt)   #Effect Size

# 5. Check effectiveness of Intervention relative to particular Threshold
#----------------------------------------------------------------------------------------------
 for (c in 1:11) {
   TS.tt[c]<- step(Total-catof[c])  # catof=0, 0.1, …, 1
 }
}
```

2. **MST**

```
model{
# 1. ANCOVA model
#------------------------
  for(i in 1:N){
    y[i] ~dnorm(mu[i],tau)
    mu[i] <- beta[1]+beta[2]*pret[i]+beta[3]*t[i]+b1[school[i]]+b2[school[i],tt[i]]
  }
  for(j in 1:M){
    b1[j]~dnorm(0.0,tau.b1)
    #b2[j]~dnorm(0.0,tau.b2)
    for(k in 1:2){b2[j,k]~dnorm(0.0,tau.b2)}
  }
# 2. Priors
#-------------
  tau.b1~dgamma(0.0001,0.0001)
 sigma.b1<-1/tau.b1
  tau.b2~dgamma(0.0001,0.0001)
  sigma.b2<-1/tau.b2
  tau~dgamma(0.0001,0.0001)
  sigma<-1/tau
  for(k in 1:p){beta[k]~dnorm(0.0,1.0E-06)}

# 3. ES, ICC and Total Variance
#---------------------------------------
  sigma.tt <-sigma + sigma.b1 + sigma.b2
  icc <- (sigma.b1+sigma.b2) * pow(sigma.tt ,-1)
  ES <- beta[3]/sqrt(sigma.tt)  # 4. Effect Size

# 5. Check effectiveness of Intervention relative to particular Threshold
#------------------------------------------------------------------------------------------------
  for (c in 1:12) {
    TS.tt[c]<- 1-step(cutof[c]-Total) # catof=0, 0.1, …, 1
  }
}
```

Appendix A3: Stan with R

```
#*****************************************************************
#                                                  1. CRT
#***************************************************************


library(rstanarm)
  stan.mlm <- stan_lmer(post ~ pret + t + +(1|school),
                              adapt_delta=0.999,
                              data=MyData) # ,iter=4000: default
sims <- as.matrix(stan.mlm)
Sim_betas <- as.matrix(stan.mlm,pars="t") #treatment effect
Sim_resi <- as.matrix(stan.mlm,pars="sigma") #sigma(pupil): sqrt of residual
Sim_schl <- as.matrix(stan.mlm,pars="Sigma[school:(Intercept),(Intercept)]") #sigma(school)^2

# 3. ES, ICC and Total Variance
#--------------------------------------
sigma.tt <-  mean(Sim_resi^2+Sim_schl)
icc <- mean(Sim_schl) /sigma.tt
sim_ES <- Sim_betas/sqrt(Sim_resi^2+Sim_schl)
ES <- round(c("ES"=mean(sim_ES), quantile(sim_ES,probs=c(0.025,0.975))),2)

# 5. Check effectiveness of Intervention relative to particular Threshold
#-----------------------------------------------------------------------------------------------
Threshold <- 0:10/10
P0to1<- data.frame(sapply(Threshold, function(x) as.numeric(sim_ES>x)))
names(P0to1)<- paste0("P", Threshold)
Pprob<- round(sapply(P0to1, mean),2)
```

```
#****************************************************************
#                                        2. MST
#**************************************************************


library(rstanarm)
stan.mlm <- stan_lmer(post ~ pret + t  + (1|school/t),
            adapt_delta=0.999,
            data=MyData) # ,iter=4000: default
sims <- as.matrix(stan.mlm)
Sim_betas <- as.matrix(stan.mlm,pars="t") #treatment effect
Sim_resi <- as.matrix(stan.mlm,pars="sigma") #sigma(pupil): sqrt of residual
Sim_schl <- as.matrix(stan.mlm,pars="Sigma[school:(Intercept),(Intercept)]") #sigma(school)^2
Sim_Gschl<- as.matrix(stan.mlm,pars="Sigma[t:school:(Intercept),(Intercept)]") #sigma(treatment by school)^2


# 3. ES, ICC and Total Variance
#--------------------------------------
sigma.tt <-  mean(Sim_resi^2+Sim_schl +Sim_Gschl)
icc <- mean(Sim_schl+Sim_Gschl) /sigma.tt
sim_ES <- Sim_betas/sqrt(Sim_resi^2+Sim_schl)
ES <- round(c("ES"=mean(sim_ES), quantile(sim_ES,probs=c(0.025,0.975))),2)


# 5. Check effectiveness of Intervention relative to particular Threshold
#------------------------------------------------------------------------------------------
Threshold <- 0:10/10
P0to1<- data.frame(sapply(Threshold, function(x) as.numeric(sim_ES>x)))
names(P0to1)<- paste0("P", Threshold)
Pprob<- round(sapply(P0to1, mean),2)
```