Pre-print version of:
Singh, A., Uwimpuhwe, G., Li, M., Einbeck, J., Higgins, S., Kasim, A. (in press) Multisite educational trials: estimating the effect size and its confidence intervals *International Journal of Research & Method in Education*
Please check the published version before citing.

Multisite educational trials: estimating the effect size and its confidence intervals

Akansha Singh[1, 4], Germaine Uwimpuhwe[1, 4], Mengchu Li[2], Jochen Einbeck[2, 4], Steve Higgins[3], Adetayo Kasim[1, 5*]

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

[2]Department of Mathematical Sciences, Durham University, Durham, UK

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

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

[5]Institute for Data Science, Durham University, Durham, UK

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

Abstract

In education, multisite trials involve randomisation of pupils into intervention and comparison groups within schools. Most analytical models in multisite educational trials ignore that the impact of an intervention may be school dependent. This study investigates the impact of statistical models on the uncertainty associated with an effect size using comparable outcomes and covariates from ten multisite educational trials funded by the UK's Education Endowment Foundation. Ordinary least squares (OLS) models often assume that the pupil's outcomes within schools are independent, which is not always true. Multilevel models address this limitation by incorporating heterogeneity between schools to account for intra-school dependency. This inflates the confidence interval of an effect size obtained from the multilevel models than from an OLS model. For a multisite trial, the heterogeneity between schools also includes the differences in the expected impact of intervention between schools. Ignoring this additional school-by-intervention variation in a multisite trial could affect both its interpretation and conclusions. A robust approach to estimate the confidence intervals for effect size from multisite trials is by treating effect size as a parameter with its distribution. This paper is important for evaluating evidence from multisite trials by accounting for all sources of variability.

Keywords: multisite trial; education, multilevel model; effect size

Introduction

Multisite randomised controlled trials are routinely used in education to evaluate the benefit of an intervention on educational attainment. A multisite trial involves two or more sites with a common intervention and data collection protocol. An important characteristic of a multisite trial is randomisation of participants to intervention and comparison groups within sites (Feingold, 2015; Meinert, 2012). In education, multisite trials involve randomisation of pupils within schools. This design is different from cluster randomised trials, where schools instead of pupils are randomised to an intervention or comparison group (see Xiao, Kasim and Higgins 2016, pp. 4 for a fuller explanation of the different designs). It is however sometimes erroneously assumed that a multisite trial design removes heterogeneity between schools and that pupils from the same school can safely be considered as independent (Hill & Rowe, 1996; Hutchison & Styles, 2010). This assumption contradicts one of the main strengths of a multisite trial that it provides useful information on how the effect of an intervention differs between schools and the fact that impact of an intervention on children in a multisite trial could depend on their school (Bloom & Spybrook, 2017). The difference in the effect of an intervention between schools or the school-by-intervention interaction poses an additional challenge in analysing data from multisite trials, which is often overlooked. Ignoring school-by-intervention interaction can also offset the assumed benefit of a multisite trial compared with a cluster randomised trial because multisite trials enable a formal testing of the generalizability of an intervention over different site settings in which the intervention is implemented (Raudenbush & Liu, 2000). Similarly, Wampold and Serlin (2000) also suggest that ignoring intervention-by-site interaction can have serious consequences for testing the null hypothesis that the intervention is equally effective for each site as well as for estimating its variance. Understanding this variation in the impact of an educational intervention across schools has important implications for policy and practice (Weiss et al., 2017). A good statistical practice is to ensure that the choice of an analytical method is informed by the study design (Shadish, Cook, & Campbell, 2002). In the case of multisite trial, it is not sufficient just to assume that school-by-intervention interaction is zero, but a precise model should be specified to appropriately capture the variation in the effect of an intervention across schools, if such variation exists. Accounting for all sources of variation in a multisite educational trials can yield appropriate estimates of intervention effect and reduces the risk of false conclusions (Wampold & Serlin, 2000).

There are different analytical approaches to analyse multisite trial data. A common but incorrect approach is to treat the pupil level data as independent and ignore the clustering of pupils within a school and school-by-intervention interaction (Tracey, Chambers, Bywater, & Elliott, 2016). A second approach is to account for school effects in the model (Styles, Clarkson, & Fowler, 2014; Wolgemuth et al., 2013), but ignore school-by-intervention interactions. The third and more appropriate approach is to account for both school and school-by-intervention interactions in the model specification. Ignoring school effects can bias the estimation of the mean difference between intervention and comparison groups and its associated standard error (Feaster, Mikulich-Gilbertson, & Brincks, 2011). This occurs when children from the same school are more homogeneous than the children from another school. The other complication in analysing multisite trial data is to treat school and school-by-intervention as fixed or random effects (Clarke, Crawford, Steele, & Vignoles, 2015). Using a fixed effect model for multisite trials with school-by-intervention interactions can obscure the interpretation of results because the effect of an intervention cannot simply be evaluated based on the mean difference between the intervention and comparison groups (Feaster *et al*., 2011). In contrary to the common assumption that a random effects model is most appropriate when the aim is to draw inferences about the wider population, a random effects model offers a flexible framework to account for different sources of variation in a multisite trial (Bloom & Spybrook, 2017) even if there is no intention to generalise findings beyond the sample. Using a random effects model is also not free from disadvantage, particularly when the sample size is very small it can overestimate effects (Eager & Roy, 2017). Despite the challenges associated with the random effects model, it is the most appropriate model which can fully capture the different sources of variability in a multisite trial. We argue that treating schools as fixed effects in a multisite trial, but as random effects in a cluster randomised trial creates inconsistency that is difficult to reconcile. The random effects model in a multisite trial has two random effects components, one component for main random effects for schools and the second component for random effects for school-by-intervention interactions (Kasim, Xiao, Higgins, & Troyer, 2017; Raudenbush & Liu, 2000).

Another important issue in analysing multisite trials is how to calculate an effect size, a common metric or standardised mean difference for the impact of an intervention in education trials. The benefit of reporting the effect size in educational trials is to have a unitless metric that can also be

used in meta-analysis (Hattie, 2009; Higgins *et al*., 2005). The effect size ameliorates the discrepancies between measuring units, and enables comparisons of educational impact on outcomes with different units (Lee, 2016). However, effect size estimators are ratio estimators, whose variances must account for estimation errors in their numerators and denominators. But in practice, analysts often ignore the variation in the denominator terms, which are assumed to be known (Schochet & Chiang, 2011). There are important differences in the way effect sizes and there associated confidence intervals are calculated in multisite trials (Borenstein, 2009). The choice of the numerator (mean difference between intervention and comparison group) and denominator (standard deviation) of the effect size vary across different studies in education. Some studies suggest the use of a mean difference between the intervention and comparison group from a regression model as the numerator and the standard deviation of the raw post-test outcome as the denominator to estimate the effect size (Connolly, Biggart, Miller, O'Hare, & Thurston, 2017; Lipsey *et al*., 2012). Some studies use unconditional variance i.e. variance estimated from a multilevel model without covariates (Tymms, 2004; Wijekumar *et al*., 2014) to estimate the standard deviation. The use of an empty multilevel model without covariates seems redundant in the sense that the total variance from an empty model should be the same as the variance of the outcome data (Moerbeek, van Breukelen, & Berger, 2003). The main point for using unconditional variance is that it allows results to be generalised beyond the participants in the trial. This argument is weak because generalising results from a randomised control trial to non-participants depends on other important factors (Deaton & Cartwright, 2018). The alternative approach as proposed in this study is to use conditional variance and to use the standard deviation from the model where other covariates are also included as fixed effects. Conditional variance can best capture the effect of an intervention on study participants and can help in establishing internal validity. Although establishing the internal validity of an intervention in an education trial is necessary, it is not a sufficient condition for external validity.

Whatever approach is used for calculating an effect size, it is important always to report a measure of uncertainty (Houle, 2007). In education, this is commonly done by reporting the confidence interval of an effect size, though there is some controversy here due to lack of probabilistic sampling

of participants in most educational trials. Confidence intervals are often interpreted as the range of possibilities (Lee, 2016). However, there is no theoretically derived formula for calculating confidence intervals for an effect size in a multisite trial using a random effects model. In this paper, we propose an approach for quantifying uncertainty in effect size estimation in multisite trials based on a similar set of theorems used by Hedges (2007) to derive uncertainty for effect sizes in cluster randomised trials. Our approach uses conditional variance and treats effect size as a random variable with its own distribution.

Data

The Education Endowment Foundation (EEF) is an independent charity that aims to raise the attainment of disadvantaged children in primary and secondary schools in England. Different educational trials have been conducted by EEF to evaluate a range of interventions directly or indirectly involving pupils to improve their educational attainment. All these trials are independently evaluated by teams mainly from universities and independent research organisations, who then submit the individual pupil-level data from these evaluations to EEF for archiving and further research. This study uses individual level pupil data from ten multisite trials extracted from the EEF Archive, judged to be of high quality with 3 or higher 'padlocks' (Higgins *et al*., 2016) and involving at least ten schools. Brief descriptions of the trials are provided in Table 1. The number of schools varies from 10 to 54 schools per trial and the number of pupils ranges from 216 to 11,590, indicating the considerable variation in the number of schools and pupils between trials. Most of the multisite trials equally allocate children to the intervention and comparison groups, except for the three trials Parent Academy, Act, Sing, and Play and ReflectEd where the allocation of children within schools was unequal. As expected, a strong correlation between pre-test and post-test reading (or mathematics) scores was observed for the multisite trials in this study. More details about the trials can be found in their respective evaluation reports (references for the evaluation reports are provided in the last column of Table 1).

Methods

Most education trials are analysed using a two-level random effects model to capture the heterogeneity between schools and pupils. While this is appropriate for a cluster randomised trial, it is not adequate for analysing multisite trials for the reasons outlined above. Suppose Pre_{ij} and Y_{ij} are the pre-intervention and post-intervention scores, respectively, for pupil j in school i, and T_{ij} is an indicator variable taking the value 1 if pupil j in school i is in the intervention group and 0 if the pupil is in the comparison group. A random effect model for a multisite trial can be formulated as

$$Y_{ij} = \beta_0 + \beta_1 Pre_{ij} + \beta_2 T_{ij} + b_{i1} + b_{i2} T_{ij} + \varepsilon_{ij}, \quad (1)$$

where β_0 is the intercept, β_1 is the gradient between post and pre-test intervention score and β_2 is an adjusted mean difference between the intervention and comparison groups. The random effect b_{i1} is the random intercept quantifying baseline heterogeneity between schools and the random effect b_{i2} quantifies differential effects of the intervention across schools through school-by-intervention interactions. Lastly, ε_{ij} denotes the residual for pupil j in school i. To complete the model formulation, the random effects $\mathbf{b} \sim N_2(\mathbf{0}, \mathbf{\Sigma})$ and residuals $\varepsilon_{ij} \sim N(0, \sigma_W^2)$ are assumed as independent. The matrix $\mathbf{\Sigma}$ is structured as

$$\mathbf{\Sigma} = \begin{pmatrix} \sigma_B^2 & \tau \\ \tau & \sigma_E^2 \end{pmatrix}, \quad (2)$$

where σ_B^2 represents the baseline heterogeneity between schools, σ_E^2 represents the variation of the effect of the interventions across schools, and τ is the covariance between the baseline heterogeneity between school and the differential effect of the interventions. More information about multilevel random effects model formulation can be found in Verbeke and Molenberghs (2009).

It follows from equation (1) that the variance of the post-intervention score (Y_{ij}) would depend on whether a pupil is in the intervention or comparison group. Similarly, variance in the longitudinal analysis with random slope can be expressed as a function of time (Fitzmaurice, Laird, & Ware, 2012). For pupils in the intervention group, the variance of the score is

$$Var\left(b_{i1} + b_{i2} T_{ij} + \varepsilon_{ij}\right) = \sigma_B^2 + \sigma_E^2 + 2\tau + \sigma_W^2, \quad (3)$$

where σ_W^2 denotes the residual variance. If $\tau = 0$ it would mean that there is no association between school ability prior to the trial and how the effect of the intervention varies between schools. If $\sigma_E^2 = 0$, the effect of the intervention is same for every school in the trial. These extra sources of variability distinguish analysing cluster randomised trials from multisite trials. For pupils in the comparison group, the variance of their scores can be derived as

$$Var(b_{i1} + \varepsilon_{ij}) = \sigma_B^2 + \sigma_W^2 \quad (4)$$

In a more generalised form equation (1) can be formulated using matrix notation with a closed formula for deriving the covariance matrix and also shown in Verbeke and Molenberghs (2009) as:

$$\mathbf{Y}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{Z}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \quad (5)$$

where $\mathbf{Y}_i$ is a vector of post-test scores in school i, $\mathbf{X}_i$ is a fixed effect design matrix for pupils in school i, $\boldsymbol{\beta}$ is a vector of regression coefficients, $\boldsymbol{Z}_i$ is a random effect design matrix for random intercept and school-by-intervention interactions, $\mathbf{b}_i$ is a vector of random effects for schools and school-by-intervention interactions, $\boldsymbol{\varepsilon}_\mathbf{i}$ is a vector of residuals for pupils in school i. According to Verbeke and Molenberghs, (2009); Fitzmaurice, Laird, and Ware (2012), the covariance matrix can be obtained as:

$$\boldsymbol{\Sigma}_i = \mathbf{Z}_i\boldsymbol{\Sigma}\mathbf{Z}_i{}' + \sigma_W^2\mathbf{I}_{n_i}$$

with $\mathbf{I}_{n_i}$ denoting the $n_i \times n_i$ identity matrix, and n_i is the number of pupils in school i.

Suppose we assume that an equal number of pupils within schools were randomised to the intervention and comparisons groups and $n_i = 4$ for illustration purpose only. Then we have

$$\mathbf{Z}'_i = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 \end{pmatrix},$$

where the first row is for the random intercepts, and the second row is for the school-by-intervention interactions (which is 1 for pupils in the intervention groups within school and 0 for pupils from the same school who are randomised to the comparison group). With $\boldsymbol{\Sigma}$ as in equation (2), the matrix $\boldsymbol{\Sigma}_i$ can be structured further as

$$\mathbf{Z}_i\mathbf{\Sigma}\mathbf{Z}_i{}' + \sigma_W^2\mathbf{I}_{n_i} = \begin{pmatrix} \sigma_B^2 + \sigma_W^2 + \sigma_E^2 + 2\tau & \sigma_B^2 + \sigma_E^2 + 2\tau & \sigma_B^2 + \tau & \sigma_B^2 + \tau \\ \sigma_B^2 + \sigma_E^2 + 2\tau & \sigma_B^2 + \sigma_W^2 + \sigma_E^2 + 2\tau & \sigma_B^2 + \tau & \sigma_B^2 + \tau \\ \sigma_B^2 + \tau & \sigma_B^2 + \tau & \sigma_B^2 + \sigma_W^2 & \sigma_B^2 \\ \sigma_B^2 + \tau & \sigma_B^2 + \tau & \sigma_B^2 & \sigma_B^2 + \sigma_W^2 \end{pmatrix}.$$

By taking the trace of $\mathbf{\Sigma}_i$, the total variance σ_T^2 is defined as $\frac{tr(\mathbf{\Sigma}_i)}{n}$,which in this example will be $\sigma_B^2 + \sigma_W^2 + 0.5\sigma_E^2 + \tau$.

More generally for unbalanced data, the total variance in a multisite trial is defined as

$$\sigma_T^2 = \frac{N^T(\sigma_B^2+\sigma_W^2+\sigma_E^2+2\tau)+N^C(\sigma_B^2+\sigma_W^2)}{N} \quad (6)$$

where N is the average number of pupils in a school, and N^T (N^C) is the average number of pupils in the intervention (comparison) groups, respectively. This shows that the total variance in a multisite trial is not as straightforward to obtain as for a cluster randomised trial.

Effect size

The main aim of this paper is to present an alternative approach for calculating confidence intervals of the effect size in multisite trials. The common practice is to calculate confidence intervals for the effect size in a multisite trial using standardized confidence intervals of the regression coefficients. Effect size is usually estimated by the formula defined below:

$$ES = \frac{\overline{Y}_{i.}^T - \overline{Y}_{i.}^C}{SD} \quad (7)$$

The effect of the intervention $\overline{Y}_{i.}^T - \overline{Y}_{i.}^C$ is often replaced by the estimated regression coefficient of the intervention $\hat{\beta}_2$. The standard deviation can be defined as the square root of the unconditional variance from an empty model without covariates (Tymms, 2004) or the square root of the conditional variance. The standard deviation, whether unconditional or conditional, is commonly defined as $SD = \sqrt{\sigma_B^2 + \sigma_W^2}$, which correspond to the total variance from a random effect model or multilevel model with only school as random effects (Xiao, Kasim & Higgins 2016). As shown in equation (6), this does not fully capture total variability in a multisite site trial. From equation (6), effect size in a multisite trial can be calculated using total variance as

$$ES = \frac{\overline{Y}_{i.}^{T} - \overline{Y}_{i.}^{C}}{SD} = \frac{\widehat{\beta}_2}{\sqrt{\frac{N^T(\sigma_B^2+\sigma_W^2+\sigma_E^2+2\tau)+N^C(\sigma_B^2+\sigma_W^2)}{N}}}. \quad (8)$$

with the variance parameters and intervention effect being replaced by their corresponding estimates from a multilevel model with random intercept and random school-by-intervention interactions. The lower bound of the confidence interval is calculated as $\frac{L\beta_2}{SD}$ and the upper bound is calculated as $\frac{U\beta_2}{SD}$, where $L\beta_2$ and $U\beta_2$ are the 95% confidence limits of the regression parameters β_2 (Maxwell *et al*., 2014). The major drawback of this approach is that it ignores the uncertaity in estimating the standard deviation term and regards it as a constant. As a result, confidence intervals based on this approach are likely to be narrower than the true confidence intervals. A more appropriate approach is to treat the effect size as a random variable with its own distribution, as proposed for cluster randomised trials by Hedges (2007).

Estimating variance of effect size in multisite trials

The main challenge for deriving confidence intervals for an effect size is that its true distribution is unknown, but can be approximated as a noncentral t-distribution as proposed by Hedges (2007). For convenience, Theorem 1 for approximating the distribution of an effect size is formulated below, while proof of the theorem is included in the appendix.

Theorem 1: Let $Y \sim N(\beta, a\sigma^2)$ and S^2 be a quadratic form in normal variables that is independent of Y so that $E(S^2) = b\sigma^2$ and $Var(S^2) = 2c\sigma^4$, where a, b and c are constants. Then

$$D = \frac{Y\sqrt{b}}{S}$$

is a consistent estimate of effect size $ES = \frac{\beta}{\sigma}$ with approximate variance

$$a + \frac{c}{2b^2}(ES)^2.$$

Balanced data

Using Theorem 1 and the other theorems in Hedges (2007), it is possible to derive a formula to calculate the variance of an effect size in a multisite trial and consequently its associated confidence intervals. We first study the case of balanced data. Suppose there are M schools with $2n$ children (or pupils) in each school. Within each school, there are n children from the intervention group and an

equal number of children in the comparison group. The school-specific mean $\beta_0 + b_{i1}$ is estimated via $\overline{Y}_{i.}^{C}$, while the school-specific intervention effect $\beta_2 + b_{i2}$ is estimated by the mean difference within school i, $\overline{Y}_{i.}^{T}$ - $\overline{Y}_{i.}^{C}$. These estimates are obviously unbiased, with variances

$$Var\left(\overline{Y}_{i.}^{C}\right) = \mathrm{Var}\left(\frac{\sum_{j=1}^{n} Y_{ij}^{C}}{n}\right) = \sigma_B^2 + \frac{\sigma_W^2}{n} \quad (9)$$

$$Var(\overline{Y}_{i.}^{T}) = \sigma_B^2 + \sigma_E^2 + 2\tau + \frac{\sigma_W^2}{n} \quad (10)$$

Using variance estimate from equation (9), (10) and the covariance between treatment and control, variance of the mean difference in each site can be estimated as.

$$Var(\overline{Y}_{i.}^{T} - \overline{Y}_{i.}^{C}) = \sigma_E^2 + \frac{2\sigma_W^2}{n} \quad (11)$$

For balanced data, the fixed effects β_0 and β_2 are estimated by simply taking the average across the sites, i.e. $\hat{\beta}_0 = \overline{Y}_{..}^{C}$ and $\hat{\beta}_2 = \overline{Y}_{..}^{T} - \overline{Y}_{..}^{C}$ and the variance of these estimates can be obtained using equation (9) and (11):

$$Var(\hat{\beta}_0) = \frac{1}{M}(\sigma_B^2 + \frac{\sigma_W^2}{n}), \quad (12)$$

$$Var(\hat{\beta}_2) = \frac{1}{M}(\sigma_E^2 + \frac{2\sigma_W^2}{n}) \quad (13)$$

To estimate the variance of the effect size, we first define the following four mean squares considering all sources of variation in our multilevel nested design.

$$S_W^2 = \frac{\sum_{i=1}^{M}(\sum_{j=1}^{n}[(Y_{ij}^{T} - \overline{Y}_{i.}^{T})^2 + (Y_{ij}^{C} - \overline{Y}_{i.}^{C})^2])}{M(2n-2)}$$

$$S_B^2 = \frac{\sum_{i=1}^{M}(\overline{Y}_{i.}^{C} - \overline{Y}_{..}^{C})^2}{M-1}$$

$$S_T^2 = \frac{\sum_{i=1}^{M}(\overline{Y}_{i.}^{T} - \overline{Y}_{..}^{T})^2}{M-1}$$

$$S_E^2 = \frac{\sum_{i=1}^{M}((\overline{Y}_{i.}^{T} - \overline{Y}_{i.}^{C}) - (\overline{Y}_{..}^{T} - \overline{Y}_{..}^{C}))^2}{M-1}.$$

S_W^2 is within school sample variance and the next three mean squares are the sample variances of $\overline{Y}_{i.}^{C}$, $\overline{Y}_{i.}^{T}$ and $\overline{Y}_{i.}^{T} - \overline{Y}_{i.}^{C}$ respectively. Montgomery (2017) provides more information about defining sum of squares and analysis of variance for nested multilevel designs.

S_W^2 is an unbiased estimate of σ_W^2 (Hedges, 2007) with variance $\frac{\sigma_W^4}{(n-1)M}$. Mean and variance of the next three mean squares according to equation (9), (10), and (11) can be estimated using chi-square distribution assumptions (Snedecor & Cochran, 1989).

$$E(S_B^2) = \sigma_B^2 + \frac{\sigma_W^2}{n} \qquad Var(S_B^2) = \frac{2(\sigma_B^2 + \frac{\sigma_W^2}{n})^2}{M-1} \tag{14}$$

$$E(S_T^2) = \sigma_B^2 + \sigma_E^2 + 2\tau + \frac{\sigma_W^2}{n} \qquad Var(S_T^2) = \frac{2(\sigma_B^2 + \sigma_E^2 + 2\tau + \frac{\sigma_W^2}{n})^2}{M-1} \tag{15}$$

$$E(S_E^2) = \sigma_E^2 + \frac{2\sigma_W^2}{n}, \qquad Var(S_E^2) = \frac{2(\sigma_E^2 + \frac{2\sigma_W^2}{n})^2}{M-1} \tag{16}$$

We first solve the simplest task of estimating effect size based on within school variance $ES_{within} = \beta_2/\sigma_W$. Using the moments of S_W^2, we can estimate d_w as

$$\widehat{ES}_{within} = \frac{\hat{\beta}_2}{S_W}$$

To obtain the approximate variance of d_w, we also need values of a, b and c as shown in Theorem 1. In this case, apply theorem 1 by replacing $\sigma^2 = \sigma_W^2$ and $S^2 = S_W^2$.

a can be estimated as $\frac{Var(\hat{\beta}_2)}{\sigma_W^2}$ where $Var(\hat{\beta}_2)$ can be obtained from equation (13), this follows that $a = \frac{\sigma_E^2 + \frac{2\sigma_W^2}{n}}{M\sigma_W^2}$

Since, S_W^2 is an unbiased estimate of σ_W^2, this follows that $b = 1$. Because, the variance of S_W^2 is $\frac{\sigma_W^4}{(n-1)M}$, this follows that $c = \frac{1}{2M(n-1)}$

Together we apply Theorem 1 to estimate $Var(\widehat{ES}_{within})$

$$Var(\widehat{ES}_{within}) = \frac{1}{M}\left(\frac{\sigma_E^2}{\sigma_W^2} + \frac{2}{n}\right) + \frac{(ES_{within})^2}{4M(n-1)} \tag{17}$$

Note that in Hedges (2007), an external estimate of the intraclass correlation $\rho = \sigma_B^2/(\sigma_B^2 + \sigma_W^2)$ is required for the value of a to be regarded as a constant. In our case, the ratio σ_E^2/σ_W^2 plays a similar

role. In practice, the effect size in the variance formula needs to be replaced by its estimate $\widehat{ES}_{within}$ and internal estimates of σ_E^2 and σ_W^2 which can be obtained using the mean squares defined above.

To obtain the variance of the effect size based on total variance in a multisite trial data, we propose the following:

$$(\widehat{ES}_{total}) = \frac{\sigma_W}{\sigma_T}(\widehat{ES}_{within})$$

$$Var(\widehat{ES}_{total}) = \frac{1}{M\sigma_T^2}(\sigma_E^2 + \frac{2\sigma_W^2}{n}) + \frac{(ES_{total})^2}{4M(n-1)} \quad (18)$$

Unbalanced data

In most educational trials, the number of children per school is different and the assumption of equal numbers of children per school is therefore questionable, though it is often assumed in sample size calculations. Suppose that there are M schools and each school has $n_i = n_i^T + n_i^C$ children, where n_i^T and n_i^C are the number of children in the intervention and comparison groups within school i respectively. The school specific intervention effect estimate $\overline{Y}_{i.}^T - \overline{Y}_{i.}^C$ has variance

$$Var(\overline{Y}_{i.}^T - \overline{Y}_{i.}^C) = \sigma_E^2 + \frac{\sigma_W^2}{\tilde{n}_i}$$

where $\tilde{n}_i = \frac{n_i^T n_i^C}{n_i^T + n_i^C}$.

Unlike taking a simple average of $\overline{Y}_{i.}^T - \overline{Y}_{i.}^C$ across schools, the generalised least squares estimation for the overall intervention effect in unbalanced data is the precision weighted average of these mean differences within school (Raudenbush & Bryk, 2002) and is defined as

$$\hat{\beta}_2 = \left(\sum_{i=1}^M \frac{1}{\sigma_E^2 + \frac{\sigma_W^2}{\tilde{n}_i}} \right)^{-1} \sum_{i=1}^M \frac{\overline{Y}_{i.}^T - \overline{Y}_{i.}^C}{\sigma_E^2 + \frac{\sigma_W^2}{\tilde{n}_i}}$$

with variance

$$Var(\hat{\beta}_2) = \left(\sum_{i=1}^M \frac{1}{\sigma_E^2 + \frac{\sigma_W^2}{\tilde{n}_i}} \right)^{-1}, \quad (19)$$

The residual variance σ_W^2 can be estimated by:

$$S_W^2 = \frac{\sum_{i=1}^{M}(\sum_{j=1}^{n_i^T}(Y_{ij}^T - \overline{Y}_{i.}^T)^2 + \sum_{j=1}^{n_i^C}(Y_{ij}^C - \overline{Y}_{i.}^C)^2)^2}{\sum_{i=1}^{M}(n_i - 2)}$$

Similar to the case of balanced data, S_W^2 is an unbiased estimate of σ_W^2 (Hedges, 2007) with variance $\frac{2\sigma_W^4}{\sum_{i=1}^{M}(n_i-2)}$.

Using equation (19) and the mean and variance estimate of S_W^2 above, we can apply Theorem 1 by replacing $\sigma^2 = \sigma_W^2$ and $S^2 = S_W^2$ and obtain values of a, b and c as we did in the balanced data case.

$$a = (\sum_{i=1}^{M} \frac{\tilde{n}_i \sigma_W^2}{\tilde{n}_i \sigma_E^2 + \sigma_W^2})^{-1}$$

$$b = 1$$

$$c = \frac{1}{\sum_{i=1}^{M}(n_i - 2)}$$

Using theorem 1, a consistent estimate of ES_{within} $(= \frac{\hat{\beta}_2}{S_W})$ can be obtained with its approximate variance as

$$Var(\widehat{ES}_{within}) = \left(\sum_{i=1}^{M} \frac{\tilde{n}_i \sigma_W^2}{\tilde{n}_i \sigma_E^2 + \sigma_W^2}\right)^{-1} + \frac{(ES_{within})^2}{2\sum_{i=1}^{M}(n_i - 2)} \quad (20)$$

Similar to the case of balanced data, relating two effect size ES_{within} and ES_{total}, variance of the effect size can be obtained as:

$$Var(\widehat{ES}_{total}) = \left(\sum_{i=1}^{M} \frac{\tilde{n}_i \sigma_T^2}{\tilde{n}_i \sigma_E^2 + \sigma_W^2}\right)^{-1} + \frac{(ES_{total})^2}{2\sum_{i=1}^{M}(n_i - 2)} \quad (21)$$

Using equations (8) and (21), we propose to calculate effect size and its associated 95% confidence intervals in multisite trials as

$$\widehat{ES}_{total} \pm 1.96\sqrt{Var(\widehat{ES}_{total})} \quad (22)$$

Model Implementation

Our analysis aims to estimate the effect size and its confidence intervals using the methods proposed in this study for ten multisite trials and compare it with the previous statistical methods typically used to analyse multisite trials. Post-test maths/reading scores were the outcome variables and the covariates used in each model were pre-test maths/reading scores and intervention indicator. We used multilevel models with both random intercept and random school-by-intervention interactions as shown in equation 1 to obtain intervention effect and its variance estimates. These estimates were compared with a simple OLS model and multilevel model with only random intercepts for school. The effect size was estimated using two different approaches. First, dividing the adjusted mean difference obtained from a multilevel model by the square root of the total variance from the same model (conditional variance) or the empty model (unconditional variance). Confidence intervals were generated by directly dividing lower and upper confidence limits of the intervention effect by the square root of the total variance estimates from the same model (conditional variance) or the empty model (unconditional variance). The second approach for estimating effect size and their confidence intervals were based on our theoretical derivation. All these analyses were performed using *R* software. We implemented the OLS model using lm package and multilevel models using lme4 package in R.

Results

In this section, we provide empirical evidence about the impact of the different choices of models for analysing multisite trials data. We first investigate whether the difference in the effect of an intervention across schools can be assumed to be zero and whether this assertion is supported by the multisite educational trials analysed in this study. We further discuss the impact of using unconditional or conditional variance. Unconditional variance implies standardising the adjusted mean difference by an unadjusted variance. Lastly, we investigate the impact of assuming the same distribution for mean difference and effect size by treating the variance estimated from the same data as a constant. Although there is no standard distribution for the effect size, it is theoretically a random variable and approximating its distribution as proposed by Hedges (2007) is a more coherent approach than treating estimated variance from the same data as a constant.

Why account for school-by-intervention interactions?

First, it is important to understand that randomisation of children within schools does not imply that the data are independent. Table 2 presents the decomposition of the sources of variation in the analysed multisite trials. As expected most of the variation is due to the difference in the outcomes between the children in each trial. However, there are cases where the proportion of variability explained by baseline heterogeneity between schools or the proportion explained by the school-by-intervention interactions is more than 10%. It should be noted that the proportion explained by either school (P.σ_B^2) or school-by-intervention (P.σ_E^2) cannot be interpreted as the intracluster correlation coefficient because the variation in a multisite trial can also be due to the differences in the quality of delivery of the intervention between schools and not just the correlation between pairs of pupils from the same school.

As shown in equation (11), baseline heterogeneity between schools does not contribute to the standard error associated with the mean difference between intervention and the comparison groups because the interventions are delivered within schools. However, the variation resulting from the differential effect of an intervention between schools contributes to the standard error. This means that assuming that the data is independent is likely to result in a smaller standard error. Table 3 presents the estimated adjusted mean difference between intervention and the comparison groups by first assuming that the data are independent (OLS), second assuming that the data are not independent and accounting for baseline difference between schools (MLM_Sch) and lastly that the data are not independent and accounting for baseline differences between schools as well as the different effects of the interventions between schools (MLM_Sch_Int).

It is interesting to note that for the outcomes where the proportion of variability due to school-by-intervention interactions is negligible (Table 2), all the three models produced similar point estimates and confidence intervals (Table 3). This is apparent for the *catchn* maths outcome, where point estimate and confidence intervals for both MLM_sch and MLM_Sch_Int model shows no difference. For the outcomes where the proportion of variability due to school-by-intervention is tangible, the multilevel model that accounted for baseline difference between school and the different effect of the intervention between schools has much wider confidence intervals. This is the case for maths outcomes *rflEdm* and *textpm* and the reading outcomes *graphorime, rflEdr, sor* and *ve*. It is not

unexpected that for trials with substantial heterogeneity in baseline differences between schools, the results from OLS and a multilevel model with only school as random effects are comparable. This is because as shown in equation (11), the baseline differences between schools cancels out in calculating the variance of mean difference in a multisite trial. It is therefore important to re-emphaise that assuming multisite trial data as independent could be misleading when the effect of an intervention vary between schools. A multisite trial is more powerful than a cluster randomised trial when there is substantial variation in the effect of an intervention across schools.

Unconditional or conditional variance?

An important discussion for calculating the effect size is about using unconditional variance or conditional variance. The main argument in favour of using unconditional variance is based on the intention to generalise the results from a selective population in randomised control trial to a wider population (Millard *et al*., 2014). In an educational trial, the aim is usually to generalise these results beyond the schools and the children that participated in the trial (Schagen & Elliot, 2004). This intention relates to the external validity of findings from a trial, which is a greater issue than just using unconditional variance. To establish external validity, one must first establish that trial participants are representative of the intended population (Deaton & Cartwright, 2018). This is, however, difficult to justify in education, where participants are not selected on the basis of probabilistic sampling and the sample size is not large enough to be representative of the intended population. Using unconditional variance is dipping one foot in external validity and another foot in internal validity. This means it cannot be robustly concluded that an intervention worked for the participants involved and at the same time conclude that it would work for non-participants from the wider intended population.

In line with the power calculation where the correlation between pre and post scores is included to increase the power of a trial, using conditional variance with pre-intervention scores would reduce unwanted variation in the post-test scores. In terms of the effect size, unconditional variance is expected to be greater than conditional variance and will consequently be more likely to result in smaller effect size and narrower confidence intervals compared to using conditional variance. However, this discussion is most likely for the studies where the effect size and its confidence intervals are calculated as a standardised mean difference with standardised confidence intervals of

the mean difference (Maxwell *et al*., 2014). Table 4 present the results of the effect size calculated using both unconditional and conditional variance from multilevel models with school and school-by-intervention as random effects. Unconditional variance consistently results in narrower confidence intervals than conditional variance. Although point estimates are similar for most of the outcomes except *catchn*, *pr9* and *sor*.

Constant variance or random variable?

It is clear from the previous discussion that it is important to account for all sources of variability in multisite trials. It is therefore recommended that conditional variance is used in concordance with the power calculation for multisite trials to reduce unwarranted variation between pupils, particularly taking advantage of the correlation between pre and post-intervention outcomes. A remaining puzzle in effect size calculation is whether to treat effect size as a random variable or whether it is sufficient to divide the confidence intervals of the adjusted mean difference by the square root of the total variance. Figures 1 and 2 present the plots comparing the effect size and their 95% confidence intervals using three different methods for maths and reading outcomes. MLM_Sch shows estimates from a multilevel model with only school as a random effect and the effect size was estimated dividing the adjusted mean difference by square root of the total variance estimates from the same model. MLM_Sch_Int estimates were obtained using a multilevel model with school and school-by-intervention as random effects and confidence intervals were estimated by dividing the adjusted mean difference by the square root of total variance estimates from the same model. Our estimates were generated based on the formula derived in the method section of this paper (equation (8), (21) and (22)) and the distribution of the effect size is approximated by a noncentral t-distribution. The point to be noted for these figures is that the effect size estimates and their confidence intervals are estimated using conditional variance on the basis of our previous analysis and discussion in this paper.

Confidence intervals based on the multilevel model with only school as random effects are generally narrower than the confidence intervals estimated from the model where school-by-intervention are included as random effects. This concern was more pertinent for the studies where the impact of an intervention was significant without accounting for school-by-intervention interactions, but became non-significant with the inclusion of school-by-intervention interactions as random effects (Figures 1

and 2). This was the case for the reading outcome *rflEdr* and maths outcome *textpm*. However, confidence intervals based on dividing confidence interval for the mean difference by constant variance is consistently similar to the confidence intervals which are estimated using an approximate distribution of the effect size. This result is unexpected since treating variance as a constant should generally result in narrower confidence intervals than when the effect size is considered as a random variable with its own approximate distribution. This should not be concluded as the general case, but it can be said for the trial data analysed in this paper. This means that the most important factor for calculating effect size in a multisite trial is to capture all the sources of variability in the data as shown in Figure 1 and Figure 2.

Discussion and Conclusions

This paper provides theoretical and empirical evidence for understanding uncertainty in the effect size in multisite trials in education. We investigated the impact of assuming data as independent in multisite trials. The impact of using conditional or unconditional variance was also discussed, as well as the impact of treating variance as a constant or random variable. The results from the analysed trial data clearly show that it is not always the case that the effect of an intervention is the same across schools. This extra source of variability in multisite trial design means that the statistical model and calculation of the effect size should account appropriately for the school-by-intervention interactions, although selection of any statistical model must also depend on the objectives of the trial. Further, allowing variation in intervention across schools can increase the level of uncertainty in the effect size and random effects model generally requires a larger sample size (Feaster *et al*., 2011).

Though it can be argued that a fixed effect model can be used (Clarke *et al*., 2015) to adjust for school-by-intervention interaction, it is advisable to capture this as random effects in multilevel models because it does not compromise the simple interpretation of the mean difference and the resulting effect size. In a fixed effect model with interactions, it is problematic interpreting main effects when there are interactions (Andersson, Cuervo-Cazurra, & Nielsen, 2014). Further, modeling interaction terms in a fixed effect framework does not benefit from shrinkage as in the random effects model framework (Bell, Fairbrother, & Jones, 2019). Since the primary aim of randomised controlled trials

in education is to investigate whether an intervention is beneficial or not, on average, the school-by-intervention interaction is therefore a nuisance parameter. A previous study (Moerbeek *et al*., 2003) conducted in clinical settings also shows that an intervention effect and especially its standard error in multisite studies are generally incorrectly estimated by the traditional OLS or fixed effect methods and should not to be used as an alternative to multilevel models. However, a further argument is whether to use a random effect for schools only or to include school-by-intervention interactions as well in the same model. Results from this study clearly shows that ignoring variation in intervention across schools can affect estimation of the effect size and its confidence intervals.

This study also shows that using unadjusted variance or unconditional variance consistently resulted in narrower confidence intervals for the effect size even though the point estimates in most cases are similar. It is perhaps a debatable view that using unconditional variance is like hedging one's bets about both the internal and external validity of a trial. However, randomised controlled trials (RCTs) must be internally valid first (i.e., the design and conduct must eliminate the possibility of bias). Moreover, claims regarding the external validity of the results derived from a randomised control trial are difficult to justify (Cartwright, 2007) even though they are routinely overgeneralized. External validity depends on the representativeness of the trial participants to the intended population (Lesko *et al*., 2017), which is often a challenging task to interpret. In balancing internal and external validity, it makes the most sense to first attend to internal validity and then external validity. If a study lacks internal validity, there is no point in generalizing this result to rest of the population (Smelser & Baltes, 2001). Using conditional variance will not provide evidence of the impact of the intervention to another population, but can show how beneficial the intervention was for the study participants. This is an important consideration when interpreting the findings of a trial for both policy and practice audiences.

Lastly, treating variance as a constant in calculating confidence intervals for an effect size appeared to make very little difference from treating the variance as a random variable and approximating its distribution by a noncentral t-distribution. However, the total variance in a multisite trial data must be correctly estimated. The difference in the impact of an intervention across schools has important

consequences for research design and can influence the statistical precision of parameter estimates from a multisite study, and thus should be taken into account (Weiss *et al*., 2017). Using total variance from the multilevel model with only school as random effects resulted consistently in narrower confidence intervals. Although it is more convenient to calculate effect size using the standardised confidence interval of the mean difference and constant variance, this study recommends using a distributional approach as derived in this paper. This approach is theoretically more appropriate (Hedges, 2007) and will be robust in a situation where the distribution of an effect size and the mean difference are not the same. Both the numerator and denominator of an effect size are measured with error, and thus, both sources of error should be taken into account in the variance calculations (Schochet & Chiang, 2011) for reliable estimation of the effect size and its confidence intervals. Distributional approaches have been extensively studied in the context of simple or cluster randomised trial effect size estimates and power analysis (Hedges, 2007) and they are reasonably accurate unless sample sizes are quite small (Fan, 2001). This approach will also protect against the potential risk of false positive findings, though the risk of this happening is very low given the empirical findings reported in this paper. Future research will investigate with an extensive simulation study to better understand the risk of treating effect size as a constant value instead of an unknown parameter with its own distribution. The methods proposed in this study have been implemented in a statistical package "eefAnalytics" in R, which has been developed to support statisticians and researchers to perform sensitivity analysis for educational trials using different analytical approaches. This package is available in the Comprehensive R archive network (CRAN) (Uwimpuhwe *et al*., 2020).

This study provides a theoretical derivation of confidence intervals for an effect size in a multisite trial using noncentral t-distribution. However, confidence intervals for any statistical parameter can be estimated alternatively using bootstrapping (Wood, 2005; Kelley, 2005). Bootstrap methods do

not assume that the data are drawn from a parametric distribution, or that the shape of the distribution is known. There are also bootstrapping procedures to calculate an effect size and its confidence intervals when the distribution of the data is known. However, bootstrapping in education trials requires further research to fully understand the implication of bootstrapping at the level of school or pupils in cluster or multisite randomised trials (Gehlbach *et al*., 2016; Huang, 2018).

It can be argued that the effect size estimation is affected by several other factors not just the selection of appropriate variance based on study design. Recently, Cheung and Slavin (2016) stated that the effect size can be affected by methodological characteristics such as researcher-made measures, sample size, randomized versus quasi-experimental designs, and published/unpublished reports, and Simpson (2017) highlighted some other reasons such as researcher's choice of the comparison group and restricted range of participants. However, Simpson (2017) supported reporting the effect size for individual studies to support future power calculation and replication. Some of these discrepancies mentioned earlier can be addressed by selecting studies with similar study design, outcomes, and covariates (Xiao, Kasim & Higgins 2016) to facilitate comparison of the effect size between trials. In this study, an effect size was estimated for multisite trials only (similar study design) using the same set of covariates to enable comparison of the effect size between trials.

Overall, this study provides an appropriate theoretical and empirical model to precisely estimate the effect size and its confidence intervals in multisite trials. This will facilitate accurate measurement of uncertainty in the effect size estimation for multisite educational trials and support programmatic decisions by setting realistic expectations about the potential magnitude of intervention effect in multisite trials.

Funding

This research was funded by the grant from the Education Endowment Foundation, United Kingdom.

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Andersson, U., Cuervo-Cazurra, A., & Nielsen, B. B. (2014). From the editors: Explaining interaction effects within and across levels of analysis. *Journal of International Business Studies*, *45*(9), 1063–1071.

Bell, A., Fairbrother, M., & Jones, K. (2019). Fixed and random effects models: making an informed choice. *Quality & Quantity*, *53*(2), 1051–1074.

Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using covariates to improve precision for studies that randomize schools to evaluate educational interventions. *Educational Evaluation and Policy Analysis*, *29*(1), 30–59.

Bloom, H. S., & Spybrook, J. (2017). Assessing the precision of multisite trials for estimating the parameters of a cross-site population distribution of program effects. *Journal of Research on Educational Effectiveness*, *10*(4), 877–902.

Borenstein, M. (2009). Effect sizes for continuous data. In C. Harris, H. Larry V., & V. Jeffrey C. (Eds.), *The handbook of research synthesis and meta-analysis* (pp. 221–235). New York: Russell Sage Foundation.

Cartwright, N. (2007). Are RCTs the gold standard? *BioSocieties*, *2*(1), 11–20.

Cheung, A. C., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. *Educational Researcher*, 45(5), 283-292.

Clarke, P., Crawford, C., Steele, F., & Vignoles, A. (2015). Revisiting fixed-and random-effects models: some considerations for policy-relevant education research. *Education Economics*, *23*(3), 259–277.

Connolly, P., Biggart, A., Miller, S., O'Hare, L., & Thurston, A. (2017). *Using randomised controlled trials in education*. Sage.

Deaton, A., & Cartwright, N. (2018). Understanding and misunderstanding randomized controlled trials. *Social Science & Medicine*, *210*, 2–21.

Eager, C., & Roy, J. (2017). Mixed effects models are sometimes terrible. *arXiv:1701.04858*.

Fan, X. (2001). Statistical significance and effect size in education research: Two sides of a coin. *The Journal of Educational Research*, *94*(5), 275–282.

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.

Feingold, A. (2015). Confidence interval estimation for standardized effect sizes in multilevel and latent growth modeling. *Journal of Consulting and Clinical Psychology*, *83*(1), 157–168.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2012). Applied longitudinal analysis (Vol. 998). John Wiley & Sons.

Gehlbach, H., Brinkworth, M. E., King, A. M., Hsu, L. M., McIntyre, J., & Rogers, T. (2016). Creating birds of similar feathers: Leveraging similarity to improve teacher–student relationships and academic achievement. *Journal of Educational Psychology*, *108*(3), 342.

Gorard, S., See, B. H., & Siddiqui, N. (2014). *Switch-on reading: Evaluation report and executive summary.*

Gorard, S., Siddiqui, N., & See, B. H. (2016). An evaluation of fresh start as a catch-up intervention: a trial conducted by teachers. *Educational Studies*, *42*(1), 98–113.

Hattie, J. (2009). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. London: Routledge.

Haywood, S., Griggs, J., Lloyd, C., Morris, S., Kiss, Z., & Skipp, A. (2015). *Creative futures: Act, sing, play. evaluation report and executive summary.*

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

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, manual. London: Education Endowment Foundation.

Hill, P. W., & Rowe, K. J. (1996). Multilevel modelling in school effectiveness research. *School effectiveness and school improvement*, 7(1), 1-34.

Houle, T. T. (2007). Importance of effect sizes for the accumulation of knowledge. *Anesthesiology: The Journal of the American Society of Anesthesiologists*, *106*(3), 415–417.

Huang, F. L. (2018). Using cluster bootstrapping to analyze nested data with a few clusters. *Educational and psychological measurement*, *78*(2), 297-318.

Husain, F. et. al. (2016). *Parent academy: Evaluation report and executive summary.*

Hutchison, D., & Styles, B. (2010). A guide to running randomised controlled trials for educational researchers. Slough: NFER.

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

Kelley, K. (2005). The effects of nonnormal distributions on confidence intervals around the standardized mean difference: Bootstrap and parametric confidence intervals. *Educational and Psychological Measurement*, 65(1), 51-69.

Lee, D. K. (2016). Alternatives to p value: confidence interval and effect size. *Korean Journal of Anesthesiology*, *69*(6), 555–562.

Lesko, C. R., Buchanan, A. L., Westreich, D., Edwards, J. K., Hudgens, M. G., & Cole, S. R. (2017). Generalizing study results: a potential outcomes perspective. *Epidemiology (Cambridge, Mass.)*, *28*(4), 553–561.

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.* Washington, D.C: National Center for Special Education Research, Institute of Education Sciences, U.S. Department of Education, NCSER 2013-3000.

Lloyd, C., Edovald, T., Kiss, Z., Morris, S., Skipp, A., & Ahmed, H. (2015). *Paired reading: Evaluation report and executive summary.*

Maxwell, B., Conolly, P., Demack, S., O'HARE, L., Stevens, A., Clague, L., & Stiell, B. (2014). *Summer active reading programme: Evaluation report and executive summary.*

Meinert, C. L. (2012). *Clinical trials: design, conduct and analysis*. OUP USA.

Millard, J. D., Muhangi, L., Sewankambo, M., Ndibazza, J., Elliott, A. M., & Webb, E. L. (2014). Assessing the external validity of a randomized controlled trial of anthelminthics in mothers and their children in Entebbe, Uganda. *Trials*, *15*(1), 310.

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

Moerbeek, M., van Breukelen, G. J. P., & Berger, M. P. F. (2003). A comparison between traditional methods and multilevel regression for the analysis of multicenter intervention studies. *Journal of Clinical Epidemiology*, *56*(4), 341–350.

Montgomery, D. C. (2017). *Design and analysis of experiments.* John Wiley & Sons.

Motteram, G., Choudry, S., Kalambouka, A., Hutcheson, G., & Barton, H. (2016). *Reflected: Evaluation report and executive summary.*

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods*. Sage.

Raudenbush, S. W., & Liu, X. (2000). Statistical power and optimal design for multisite randomized trials. *Psychological Methods*, *5*(2), 199–213.

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

Schagen, I., & Elliot, K. (2004). *What does it mean? the use of effect sizes in educational research*. Slough, Berks: National Foundation for Educational Research.

Schochet, P. Z., & Chiang, H. S. (2011). Estimation and identification of the complier average causal effect parameter in education RCTs. *Journal of Educational and Behavioral Statistics*, *36*(3), 307–345.

Simpson, A. (2017). The misdirection of public policy: Comparing and combining standardised effect sizes. *Journal of Education Policy*, 32(4), 450-466.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Boston, MA, US: Houghton, Mifflin and Company.

Smelser, N. J., & Baltes, P. B. (2001). *International encyclopedia of the social & behavioral sciences*. Amsterdam ; New York : Elsevier.

Styles, B., Clarkson, R., & Fowler, K. (2014). *Rhythm for reading: Evaluation report and executive summary.*

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

Tracey, L., Chambers, B., Bywater, T., & Elliott, L. (2016). *Spokes: Evaluation report and executive summary.*

Tymms, P. (2004). Effect sizes in multilevel models. In I. Schagen & K. Elliot (Eds.), *But what does it mean? the use of effect sizes in educational research.* (pp. 55–66). Slough, Berks: National Foundation for Educational Research.

Uwimpuhwe G., Singh, A., Higgins, S., Xiao, Z., Troyer, E. D., & Kasim, A. (2020). eefAnalytics: Robust Analytical Methods for Evaluating Educational Interventions using Randomised Controlled Trials Designs [Computer software manual]. (R package version 1.0.8)

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

Wampold, B. E., & Serlin, R. C. (2000). The consequence of ignoring a nested factor on measures of effect size in analysis of variance*. Psychological Methods*, 5(4), 425.

Weiss, M. J., Bloom, H. S., Verbitsky-Savitz, N., Gupta, H., Vigil, A. E., & Cullinan, D. N. (2017). How much do the effects of education and training programs vary across sites? evidence from past multisite randomized trials. *Journal of Research on Educational Effectiveness*, *10*(4), 843–876.

Wijekumar, K., Meyer, B. J. F., Lei, P.-W., Lin, Y.-C., Johnson, L. A., Spielvogel,J. A., Cook, M. (2014). Multisite randomized controlled trial examining intelligent tutoring of structure strategy for fifth-grade readers. *Journal of Research on Educational Effectiveness*, *7*(4), 331–357.

Wolgemuth, J. R., Savage, R., Helmer, J., Harper, H., Lea, T., Abrami, P. C. et al. (2013). Abracadabra aids indigenous and non-indigenous early literacy in Australia: Evidence from a multisite randomized controlled trial. *Computers & Education*, *67*, 250–264.

Wood, M. (2005). Bootstrapped confidence intervals as an approach to statistical inference. *Organizational Research Methods*, 8(4), 454–470.

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

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.

Table 1

Descriptive statistics for multisite educational trials used in this study.

Full project title	outcome	n.sch	n.t	n.c	n	pp.corr	Report
maths							
Creative Futures: Act, Sing, Play	aspm	19	542	274	816	0.72	(Haywood et al., 2015)
Catch up Numeracy	catchn	54	108	108	216	0.59	(Rutt, Easton, & Stacey, 2014)
Parent Academy	paicm	16	509	803	1312	0.65	(Husain et al., 2016)
Parent Academy	pauicm	16	611	803	1414	0.65	(Husain et al., 2016)
ReflectEd	rflEdm	28	800	707	1507	0.67	(Motteram, Choudry, Kalambouka, Hutcheson, & Barton, 2016)
Texting Parents	textpm	29	5613	5977	11590	0.78	(Miller et al., 2017)
reading							
Fresh Start	fs	10	215	204	419	0.72	(Gorard, Siddiqui, & See, 2016)
GraphoGame Rime	graphorime	14	185	177	362	0.57	Worth, Nelson, Harland, Bernardinelli, & Styles, 2018)
Parent Academy	paicr	16	497	793	1290	0.65	(Husain et al., 2016)
Parent Academy	pauicr	16	605	793	1398	0.65	(Husain et al., 2016)
Paired Reading (Year 7)	pr7	10	627	682	1309	0.81	(Lloyd et al., 2015)
Paired Reading (Year 9)	pr9	10	620	656	1276	0.78	(Lloyd et al., 2015)
ReflectEd	rflEdr	28	719	622	1341	0.63	(Motteram et al., 2016)
Switch-on Reading	sor	19	155	153	308	0.65	(Gorard, See, & Siddiqui, 2014)
Vocabulary Enrichment Intervention Programme	ve	12	282	288	570	0.68	(Styles, Stevens, Bradshaw, & Clarkson, 2014)

Notes: n is the sample size, n.t and n.c are sample sizes for intervention and control groups, n.sch is the number of schools for each study, pp.corr denotes pre-test and post-test scores correlations.

Table 2

Decomposition of the total variance in each multisite educational trial obtained from the multilevel model for each outcome with school and school-by-intervention as random effects.

outcome	σ_B^2 (school variance)	σ_E^2 (school intervention interaction variance)	σ_W^2 (residual variance)	P.σ_B^2 (Proportion)	P. σ_E^2 (Proportion)	P. σ_W^2 (Proportion)
maths						
aspm	11.55	0.15	41.15	21.85	0.28	77.86
catchn	15.60	0.42	91.27	14.54	0.39	85.07
paicm	13.19	0.86	176.09	6.94	0.45	92.61
pauicm	12.01	0.47	165.94	6.73	0.26	93.01
rflEdm	129.21	59.88	166.69	36.32	16.83	46.85
textpm	0.03	0.04	0.34	7.32	9.76	82.93
reading						
fs	0.00	63.37	1558.71	0.00	3.91	96.09
graphorime	0.00	3.25	39.20	0.00	7.66	92.34
paicr	5.18	0.12	128.10	3.88	0.09	96.03
pauicr	5.37	0.32	118.34	4.33	0.26	95.41
pr7	21.23	18.79	995.74	2.05	1.81	96.14
pr9	117.07	0.15	1349.22	7.98	0.01	92.01
rflEdr	82.05	51.53	97.86	35.45	22.26	42.28
sor	0.47	7.83	47.80	0.84	13.96	85.20
ve	1.56	4.00	25.18	5.07	13.01	81.91

Notes: σ_B^2 is variance for school only, σ_E^2 is school and intervention interaction variance, σ_W^2 is residual variance and P.σ_B^2 , P.σ_E^2 , P.σ_W^2 shows proportion of variance. Multilevel models with school and school-by-intervention as random effects as mentioned in the equation 1 was used to obtain these estimates.

Table 3

Adjusted mean difference between intervention and comparison groups from ordinary least square model (OLS), multilevel model with only school as random effects (MLM_Sch) and multilevel model with school and school-by-intervention as random effects (MLM_Sch_Int). Note that all models include the pre-intervention scores and intervention groups as predictors.

outcome	OLS	MLM_Sch	MLM_Sch_Int
maths			
aspm	-0.11 (-1.14,0.93)	-0.07 (-1.01,0.86)	-0.10 (-1.05,0.85)
catchn	2.86 (0.06,5.66)	2.92 (0.35,5.50)	2.92 (0.35,5.50)
paicm	0.10 (-1.43,1.63)	-0.07 (-1.57,1.43)	-0.05 (-1.62,1.52)
pauicm	-0.56 (-1.96,0.84)	-0.50 (-1.88,0.87)	-0.58 (-2.00,0.83)
rflEdm	-0.13 (-1.72,1.46)	-0.17 (-1.54,1.21)	-0.22 (-3.39,2.96)
textpm	0.07 (0.04,0.09)	0.07 (0.04,0.09)	0.07 (-0.01,0.14)
reading			
fs	3.05 (-4.71,10.80)	2.96 (-4.65,10.57)	2.42 (-6.98,11.81)
graphorime	-0.25 (-1.56,1.07)	-0.30 (-1.59,0.99)	-0.49 (-2.17,1.19)
paicr	-0.28 (-1.57,1.01)	-0.20 (-1.48,1.09)	-0.22 (-1.52,1.08)
pauicr	0.00 (-1.17,1.18)	0.08 (-1.09,1.24)	0.06 (-1.14,1.26)
pr7	-1.73 (-5.18,1.73)	-1.56 (-5.01,1.88)	-1.44 (-5.89,3.01)
pr9	-3.73 (-7.88,0.41)	-4.34 (-8.40,-0.29)	-4.42 (-8.48,-0.36)
rflEdr	-1.24 (-2.52,0.04)	-1.45 (-2.60,-0.31)	-1.54 (-4.45,1.37)
sor	1.91 (0.33,3.49)	1.91 (0.33,3.49)	2.18 (0.17, 4.19)
ve	0.43 (-0.42,1.27)	0.42 (-0.42,1.26)	0.37 (-1.05,1.80)

Notes: 95 percent confidence interval is reported in the parentheses.

Table 4

Effect size estimation using unconditional and conditional variance from a multilevel model with school and school-by-intervention as random effects. Note that only pre-intervention scores and intervention groups are included in the conditional variance model.

outcome	Unconditional Variance	Conditional Variance
maths		
aspm	-0.01 (-0.10,0.08)	-0.01 (-0.15,0.12)
catchn	0.22 (0.03,0.42)	0.28 (0.03,0.53)
paicm	0.00 (-0.09,0.08)	0.00 (-0.12,0.11)
pauicm	-0.03 (-0.11,0.05)	-0.04 (-0.15,0.06)
rflEdm	-0.01 (-0.16,0.14)	-0.01 (-0.21,0.19)
textpm	0.07 (-0.01,0.15)	0.11 (-0.02,0.23)
reading		
fs	0.04 (-0.12,0.20)	0.06 (-0.18,0.30)
graphorime	-0.06 (-0.28,0.15)	-0.08 (-0.34,0.19)
paicr	-0.01 (-0.10,0.07)	-0.02 (-0.13,0.09)
pauicr	0.00 (-0.08,0.09)	0.01 (-0.10,0.11)
pr7	-0.03 (-0.11,0.05)	-0.05 (-0.18,0.09)
pr9	-0.07 (-0.14,-0.01)	-0.12 (-0.22,-0.01)
rflEdr	-0.10 (-0.28,0.09)	-0.13 (-0.37,0.11)
sor	0.23 (0.02,0.45)	0.31 (0.02,0.59)
ve	0.05 (-0.15,0.26)	0.07 (-0.20,0.35)

Notes: Estimates in the unconditional and conditional variance columns show effect sizes with 95 percent confidence interval reported in the parentheses.

Figure 1.

Comparison of the effect size and its confidence intervals for maths outcome using different models specifications.

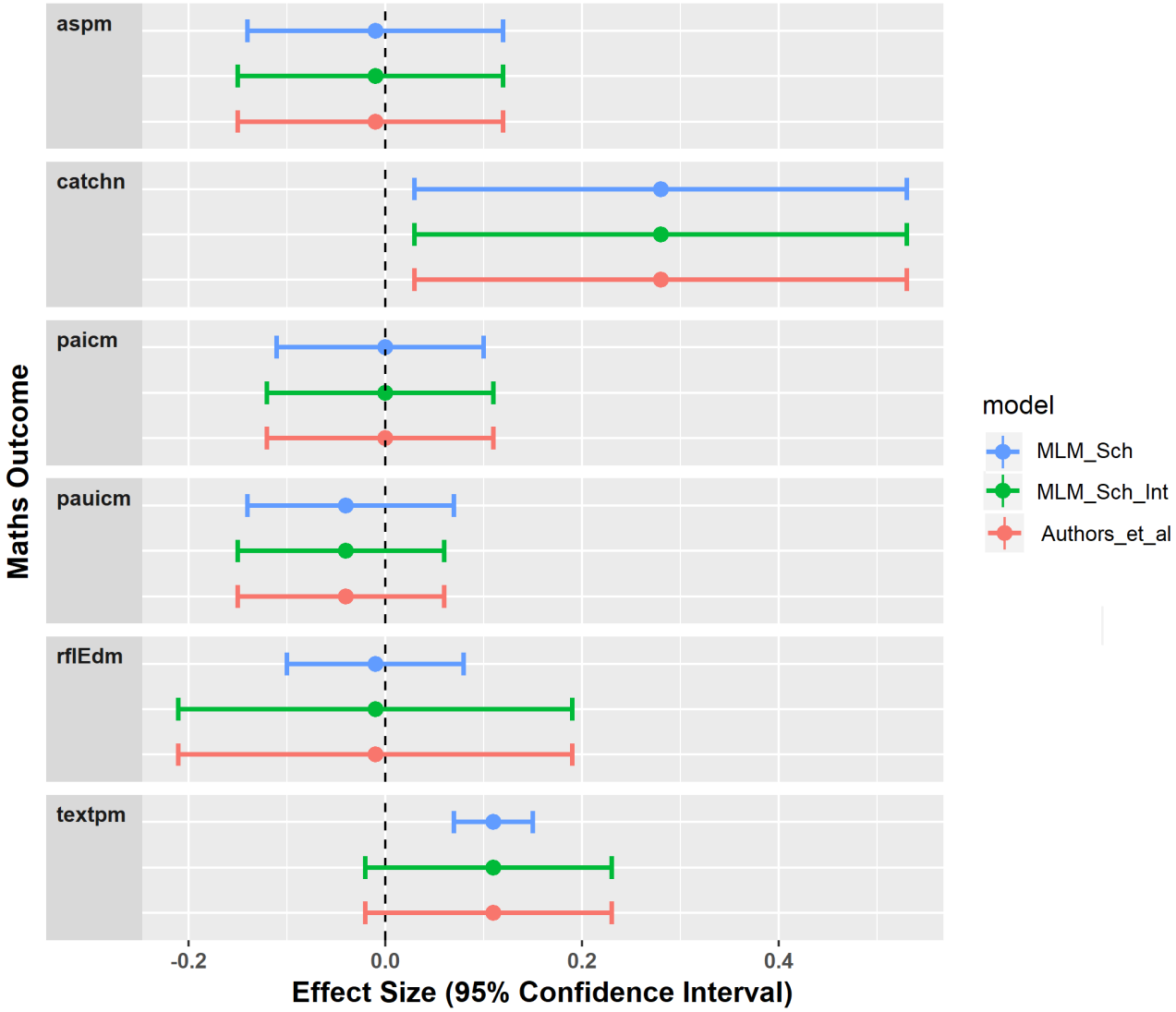
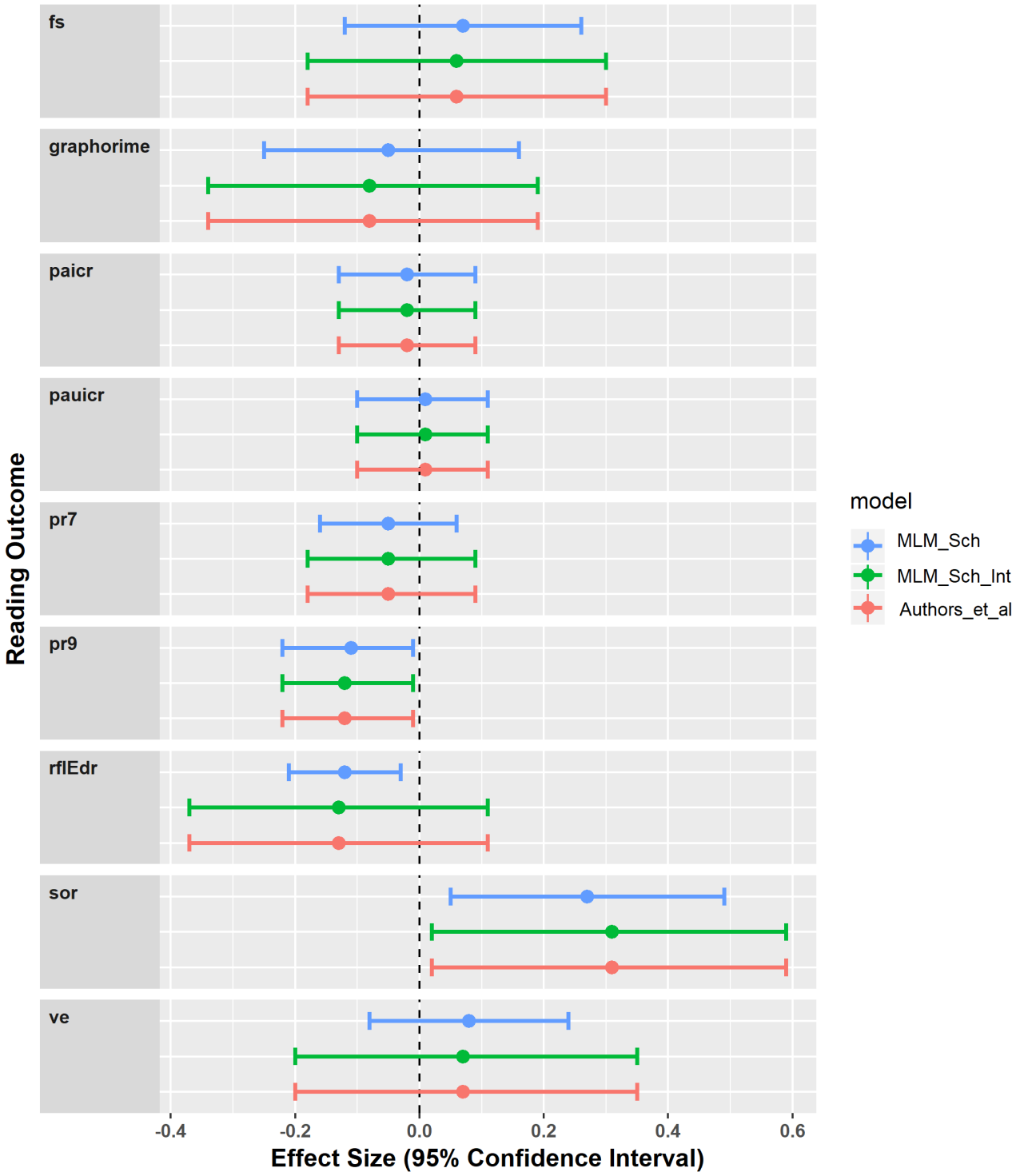


Figure 2.

Comparison of the effect size and its confidence intervals for reading outcomes using different model specifications.



Appendix

Theorem: Let $Y \sim N(\mu, \frac{a\sigma^2}{\widetilde{N}})$ and S^2 be a quadratic form in normal variables that is independent of Y so that $E(S^2) = b\sigma^2$ and $Var(S^2) = 2c\sigma^4$, where a, b and c are constants. Then

$$D = \frac{Y\sqrt{b}}{S}$$

is a consistent estimate of effect size $\delta = \frac{\beta}{\sigma}$ with approximate variance

$$a + \frac{c}{2b^2}(ES)^2$$

An approximately unbiased estimate of δ is given by $DJ(b^2/c)$ where

$$J(x) = 1 - \frac{3}{4x-1}.$$

Proof. Given the assumptions that

$$Y \sim N(\mu, \frac{a\sigma^2}{\widetilde{N}}),$$

$$E(S^2) = b\sigma^2, Var(S^2) = 2c\sigma^4,$$

we have that $\frac{Y-\mu}{\sqrt{\frac{a\sigma^2}{\widetilde{N}}}}$ is a standard normal random variable, and $\frac{S^2}{b\sigma^2}$ is distributed approximately as $\frac{\chi^2(h)}{h}$ where $h = \frac{b^2}{c}$. Therefore

$$T = \frac{Y\sqrt{\frac{\widetilde{N}}{a\sigma^2}}}{\sqrt{\frac{S^2}{b\sigma^2}}} = \frac{(Y-\mu+\mu)\sqrt{\frac{\widetilde{N}}{a\sigma^2}}}{\sqrt{\frac{S^2}{b\sigma^2}}} = \frac{(Y-\mu)\sqrt{\frac{\widetilde{N}}{a\sigma^2}}+\mu\sqrt{\frac{\widetilde{N}}{a\sigma^2}}}{\sqrt{\frac{S^2}{b\sigma^2}}}$$

becomes a noncentral t-distribution with degree of freedom $\frac{b^2}{c}$ and noncentrality parameter

$$\theta = \mu\sqrt{\frac{\widetilde{N}}{a\sigma^2}} = \frac{\mu}{\sigma}\sqrt{\frac{\widetilde{N}}{a}} = \delta\sqrt{\frac{\widetilde{N}}{a}}$$

Note that the equation for T can be simplified as $T = \frac{Y}{S}\sqrt{\frac{\widetilde{N}b}{a}}$. The expectation and variance of T can be derived via using the formulas above as

$$E(T) = \delta\sqrt{\frac{\widetilde{N}}{a}}f(h)$$

$$Var(T) = \frac{h(1+\theta^2)}{h-2} - \theta^2 f^2(h)$$

$$= \frac{h}{h-2}\left(1 + \delta^2 \frac{\widetilde{N}}{a}\right) - \delta^2 \frac{\widetilde{N}}{a} f^2(h)$$

Consequently, $D = T\sqrt{\frac{a}{\widetilde{N}}} = \frac{Y\sqrt{b}}{S}$ with the expectation being $E(D) = \delta f(h)$ and variance being $Var(D) = \frac{h}{h-2}\left(\frac{a}{\widetilde{N}} + \delta^2\right) - \delta^2 f^2(h)$, is a consistent estimate of the effect size δ since

$$E(D) = \delta f(h) \approx \frac{\delta}{J(h)} \to \delta \quad as \quad h \to \infty$$

$$Var(D) \to \frac{a}{\widetilde{N}} \; as \quad h \to \infty$$

$$\to 0 \quad as \quad \widetilde{N} \to \infty$$

If we use the large sample normal approximation to the noncentral t-distribution for T, $Var(D)$ can be approximated by

$$\frac{a}{\widetilde{N}}\left(1 + \frac{\theta^2}{2h}\right) = \frac{a}{\widetilde{N}} + \frac{c\delta^2}{2b^2}$$

Finally, an unbiased estimator for the effect size can be constructed as

$$\frac{D}{f(h)} \approx DJ(h) = DJ\left(\frac{b^2}{c}\right)$$

Moreover, the variance of the unbiased estimator is always smaller than that of D since

$$Var(DJ(h)) = J^2(h)Var(D) < Var(D)$$

The consistency argument and the construction of the unbiased estimator follow closely from Hedges (2007) in which the Hedges's g was proposed.