EEF Research Paper

No. 002
February 2018

Education Endowment Foundation

Standard Deviation as an outcome on interventions: a methodological investigation

Authors:

Peter Tymms
Adetayo Kasim

About the authors:

Peter Tymms (Durham University, School of Education)
Adetayo Kasim (Durham University, Department of Anthropology)

Contact details:

Peter Tymms
School of Education
Durham University
Durham, DH5 9RG
Email: p.b.tymms@dur.ac.uk

Executive Summary

This report considers the change in Standard Deviation (SD) which may result from an intervention. It explores a variety of metrics which may be used to report changes, the confidence intervals around those metrics, and the statistical models which might be used during analyses.

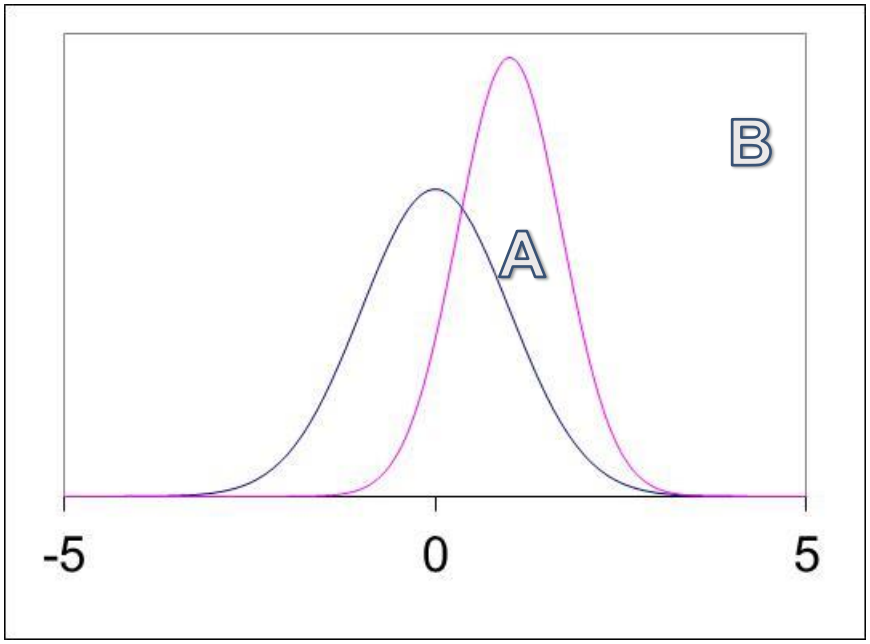
It recommends the simple recording of percentage change in SD in the first instance, together with 95% Confidence Intervals. For meta-analyses it recommends the use of the log of the ratio of variances which is a metric with the suitable properties.

The preferred statistical model is ANCOVA (Analysis of Covariance) and it is noted that this assumes equal variances at baseline. Despite randomisation, that assumption may not always hold and it is recommended that large sample sizes are employed together with high quality measures.

Background

Interventions commonly report Effect Sizes[1] as the summary outcome of the extent to which the intervention group benefit compared to the control group. However, a secondary impact might well be present and that is a shift in the variance (the square of the standard deviation) of the intervention group. Figure 1 below helps to make this clear.

Figure 1: The impact of an intervention of the mean and SD



The normal distribution of the control group scores is shown in blue (A). It has a mean of 0 and a SD of 1. The distribution of the intervention group scores (B) is also normal but it has a mean of 1 and a SD of 0.7. The intervention has had a large impact both on the mean and standard deviation, and consequently, the variance. It is the change in standard deviation of the intervention group which is the subject of this report.

[1] Commonly defined as $= \frac{\bar{X}_{experimental} - \bar{X}_{control}}{SD_{pooled}}$

Why is the topic important?

There are two reasons why it is useful to study a change in the group variance. The first is that it is important in its own right, as it might well be that the purpose of an intervention is to reduce inequality and a reduction in inequality can be measured by a reduction in the standard deviation. By contrast, it may be that some interventions actually increase the standard deviation and, on occasions, that may be appropriate. For example, a gifted programme might be intended to take those who show promise and to extend them; the standard deviation should increase if this is achieved.

The second reason for studying standard deviation changes is that it might help to give us an insight into why an intervention works on some occasions and not on others. This could be because it is more effective with some groups rather than others; this would result in a change in standard deviation and researchers could exploit this to help understand one of the several possible reasons for the differential effectiveness of an intervention. For example, if an intervention was most effective with the less able students but had a small impact on average students and no impact on the more able students, then the SD of the experimental groups would be lower than that of the control.

Previous work

We select three key features from the literature which are relevant to this paper:

1. Over 50 years ago Levene (1960) developed a test for changes in SD which is named in his honour. It is very widely used and appears as standard in, for example, the Statistics Package for the Social Science (SPSS) output when a t-test for the difference in means is requested.
2. If a difference is found in SDs between groups, it has been common to report the ratio of SDs. However, Hedges and Friedman (1993) report that the log of the ratio is preferable. They note, after Bartlett and Kendal (1946), that the normality of the distribution of sample variances from a normally distributed population is "greatly improved" by a log-transformation. They also note that the log-transformed variance ratio is a suitable metric for meta-analysis as it is normally distributed with a variance that depends only on sample size and the magnitude of the effect. A similar point is made by Katzman and Alliger (1992).
3. In 2013, Bloom and Brock set out a conceptual framework for looking at the impact of interventions including the impact on variance. Raudenbush and Bloom (2015a and 2015b) then looked at how variance changes during intervention. Their aim is to explore the differential impact of interventions on subgroups. They write:

   "*typically, these trials have only focused on average program impacts and program impacts for common socio-demo subgroups. That is about to change, however, as researchers, policymakers and practitioners are beginning to see the value of learning about and from variation in program impacts across individuals, across theoretically- and policy-relevant subgroups of individuals, and across program sites*"

Aims

It makes sense to record the change in standard deviation for every intervention which has been carried out for the EEF, if possible. This paper aims to develop recommendations to ensure that appropriate records can be kept. There are a number of issues to be looked at before a recommendation can be made and these are addressed below, before we suggest ways to quantify the impact on the spread of scores, as a result of the intervention, which is applicable to all well-designed interventions.

This will enable the EEF to record changes in variance in a standard way which can then be integrated to further analyses. One metric will allow meta-analyses to be carried out.

Methods

Analyses of EFF trials have, to date, focussed on assessing improvement in pupils' attainment between intervention and control groups assuming constant variance between the two groups. The implication of this approach is that all interventions are assumed to benefit both high and low performing pupils equally. This implies that inequality (variance) in pupils' attainments is constant within the groups; intervention and control, pre and post. The current approach therefore favours interventions that improves average performance over those that have tendency to reduce inequality in pupils' attainments as captured by the spread of the outcome data.

ANCOVA approach

Due to post baseline randomisation, we assume that pupils' attainment between the intervention and the control group is the same at baseline. Hence, it is sufficient to account for baseline performance in the comparison of average performance between the intervention and control groups. Let y_{ij} be the post intervention scores of pupil j from school i with $i = 1,2,\cdots,N_j$ and $j = 1,2,\cdots,M$. The intervention effect is usually evaluated using the following multilevel model,

$$y_{ij} = \beta_0 + \beta_1 * Pretest_{ij} + \beta_2 * Treatment_{ij} + b_i + \varepsilon_{ij} \qquad \textit{eqn 1}$$

where $b_i \sim N(0, \gamma^2)$ is the random effect specification for the differences between schools and $\varepsilon_{ij} \sim N(0, \sigma^2)$ captures the residual variance or differences between pupil's attainments. β_0 is the intercept, β_1 is the gradient between the pre- and post-test scores and β_2 is the average intervention effect.

To evaluate reduction in inequality (variance) within the ANCOVA approach, we propose modifying the above model specification so that the variance of random effects and residual variance differ between the intervention and the control groups.

$$y_{ij} = \beta_0 + \beta_1 * Pretest_{ij} + \beta_2 * Treatment_{ij} + b_i^C + b_i^T + \varepsilon_{ij}^C + \varepsilon_{ij}^T \qquad \textit{eqn 2}$$

$b_i^C \sim N(0, \gamma_C^2)$, $b_i^T \sim N(0, \gamma_T^2)$ are the random effects for the intervention and the control groups, respectively. $\varepsilon_{ij}^C \sim N(0, \sigma_C^2)$ and $\varepsilon_{ij}^T \sim N(0, \sigma_T^2)$ are the residuals for the pupils in the intervention and control groups, respectively. If the spread of the data in the intervention and the control groups are similar, then $\gamma_C^2 = \gamma_T^2 = \gamma^2$. Similarly, for the residual variance $\sigma_C^2 = \sigma_T^2 = \sigma^2$. We propose using the likelihood ratio test (Verbeke and Molenberghs 2000) to compare the two models to examine whether an education intervention reduces or increases inequality between pupils' performance. In addition, we propose to use the following metrics to facilitate interpretation and communication of results:

- F-Statistics: Adjusting the traditional F-statistic we define the ratio as SD_T/SD_C where SD_T is the standard deviation for the intervention group and $SD_C$is the standard deviation for the control group based on post test scores. Instead of a traditional F-test, we propose to bootstrap the data to generate confidence intervals. The estimate will be non-statistically significant if one is included within the 95% confidence interval.
- $\log(SD_T/SD_C)$: We propose to adjust the ratio of standard deviations by logarithmic transformation. On the log scale, statistical significance will be considered to have been achieved if the 95% confidence interval excludes zero. The log scale is more suitable for meta-analysis if the interest is to pool standard deviation across different studies.
- Reduction in standard deviation: To facilitate interpretation we propose to calculate reduction in standard deviation as 1-(SD_T/SD_C). Multiplying this by 100 will show by what percentage is

the standard deviation in the intervention group different from the standard deviation in the control group. Since this can be either positive or negative, statistical significance will be considered if the 95% confidence interval excludes zero.

Note that SD_C and SD_T are based on equation 2 and they are calculated as $SD_C = \sqrt{\gamma_C^2 + \sigma_C^2}$ and $SD_T = \sqrt{\gamma_T^2 + \sigma_T^2}$. In addition to 95% non-parametric confidence intervals, we also consider generating p-values using likelihood ratio test statistics by comparing eqn 2 and eqn 1. If the p-value is not significant, it means that the variation in the data is the same for the intervention and control groups. However, if the p-value is significant it means that residual variance and between-school heterogeneity differs between the two groups.

Difference-in-Difference approach

The ANCOVA approach is generally more powerful for randomised controlled trials for comparing average differences. However, we do not know for sure whether this also holds for comparing variances. As a sensitivity analysis, we propose a generalised model to test inequality (variance) in attainment by comparing the spread of the outcome data between the intervention and the control groups, and between the pre- and post- intervention periods. The model is formulated as:

$$y_{ijk} = \beta_0 + \beta_1 * P_{jk} + \beta_2 * Trt_{ij} + \beta_3 * P_{jk} * Trt_{ij} + b_i^C + b_i^T + \varepsilon_{ijk}^C + \varepsilon_{ijk}^T \qquad \textit{eqn 3,}$$

Where $\boldsymbol{P}$ indicates whether the outcome data for pupil j is at baseline (k=1) or post intervention (K=2), note that for convenience P is coded as 0/1. y_{ij1} is the baseline score and y_{ij2} is the post-test score for pupil j from school i. $b_i^C \sim N(0, \gamma_{k,C}^2)$ and $b_i^T \sim N(0, \gamma_{k,T}^2)$ are the random effects for the intervention and the control groups, respectively. $\varepsilon_{ij}^C \sim N(0, \sigma_{k,C}^2)$ and $\varepsilon_{ij}^T \sim N(0, \sigma_{k,T}^2)$ are the residual variances for the intervention and control groups, respectively. The reduction in standard deviation is calculated as $(Pre - Post)/Pre$, where $Pre = SD_{pre,C}/SD_{pre,T}$ and $SD_{post,C}$ and $SD_{pre,C}$ are $\sigma_{2,C}^2 + \gamma_{2,C}^2$ and $\sigma_{1,C}^2 + \gamma_{1,C}^2$, respectively. Similarly, $Post = SD_{post,C}/SD_{post,T}$. $SD_{post,T}$ and $SD_{pre,C}$ are defined as $\sigma_{2,T}^2 + \gamma_{2,T}^2$, $\sigma_{1,T}^2 + \gamma_{1,T}^2$, respectively. 95% non-parametric confidence intervals based on 500 bootstraps of the data will be generated. P-values will also be generated based on likelihood ratio tests comparing eqn 3 and eqn 1.

Analysis of the EEF database using the different metrics and a comparison of the results.

The data from the 16 EEF trials for which the baseline and outcome data used the same measure, were analysed using equations 2 and 3. The analyses were carried out using the raw data and the normalised data. The latter were generated by normalising the pre-intervention scores (intervention and control data together) using rank ordering using Blom's (1958) transformation. Then the post intervention scores were allocated scores based on a look up table with imputations from the initial normalisation.

Results from ANCOVA

Table 1 presents the results from analysis of covariance between intervention and control groups. Assuming equal variation in the two groups at baseline, 3 out of the 16 trials analysed show a significantly (p<.05) larger standard deviation for the intervention group compared to the control group based on the raw data and 95% non-parametric bootstrapped confidence intervals, and the p-values from likelihood ratio test statistics.

The 'Switch-on Reading efficacy trial' shows bigger variation in the intervention group than the control group. The variation in the intervention group is about 20% more than the variation in the control group.

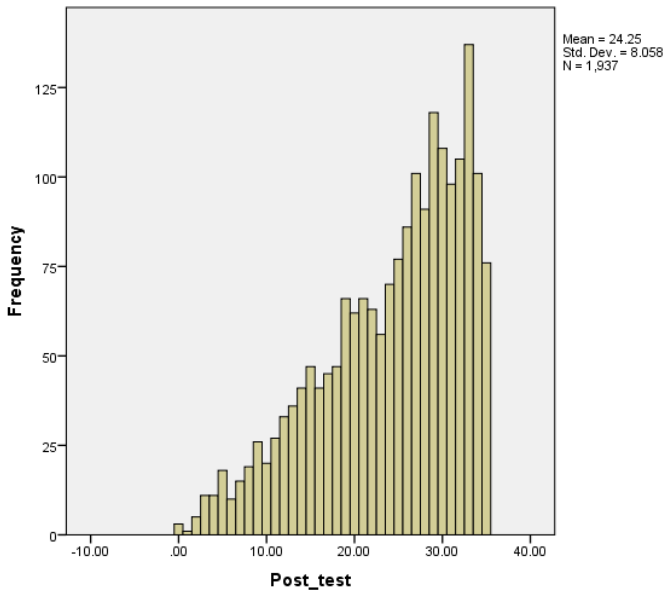
Although this trial had significant positive impact with effect size of 0.30 (0.03, 0.56), it also appears to widen the gap between pupils. This may be because the lower performing pupils are not gaining as much as higher performing pupils.

Another trial with significant variation between the intervention and control group is 'Catch-Up Numeracy© effectiveness trial'. The variation in mathematics (Time) in the intervention group is about 28% more than the variation in the control group.

The mathematics outcome from 'Improving Numeracy and Literacy efficacy trial' also showed a bigger variation in the intervention group than the control group. The variation in the intervention group is about 97% more than the variation in the control group. This trial also had positive impact on performance with effect size of 0.30 (0.02, 0.58).

We also analysed the normalised data to ensure that variation between the intervention and control groups is not distorted by skewed distributions. The results are consistent with the results from the analysis of the actual data without transformation although for one trial, the 'Improving Numeracy and Literacy efficacy trial', with English as the outcome, a significant increase in the spread of scores was recorded for the intervention group. The reason for this difference surely lies in the distribution of the post test scores which is far from normal – see Figure 2 below.

Figure 2: Distribution of English Post test scores for the 'improving numeracy and literacy efficacy trial'



In general, most of the trials appear to show bigger variation in the intervention group than the control group. In only 3 cases, the control group has a larger variance than the intervention group, but those differences are small and not statistically significant.

Table 1: ANCOVA analysis of difference in variation between intervention and control group using multilevel model

		Actual Data				Normalised Data			
Trial	Outcome	SD_T/SD_C	$Log(SD_T/SD_C)$	$1-(SD_T/SD_C)$	P-value	SD_T/SD_C	$Log(SD_T/SD_C)$	$1-(SD_T/SD_C)$	P-value
Rapid Phonics Efficacy Trial – p13	*NGRT 3B*	1.02 (0.75,1.37)	0.02 (-0.29, 0.32)	-0.02 (-0.37,0.25)	0.951	1.07 (0.84,1.34)	0.07 (-0.29,0.32)	-0.07 (-0.34,0.16)	0.779
Switch-On Reading Efficacy Trial – p2	NGRT B	1.20 (1.01,1.4)	0.18 (0.01,0.39)	-0.20 (-0.47, -0.01)	0.021	1.19 (0.99,1.43)	0.17 (0.01,0.39)	-0.19 (-0.43,0.01)	0.041
Word and World Reading Pilot Study-p26	*PiE*	0.87 (0.81,0.96)	-0.13 (-0.21, -0.04)	0.12 (0.04,0.19)	0.368	0.86 (0.80,0.95)	-0.14 (-0.21, -0.04)	0.13 (0.05,0.2)	0.301
Fresh Start Pilot Study-p31	*NGRT B*	1.13 (0.88,1.40)	0.12 (-0.12,0.34)	-0.13 (-0.4,0.12)	0.577	0.98 (0.82,1.2)	-0.01 (-0.12,0.34)	0.01 (-0.2,0.18)	0.192
Talk for Literacy Efficacy Trial-p32	*NGRT B*	1.16 (0.94,1.48)	0.15 (-0.06,0.39)	-0.16 (-0.48,0.06)	0.287	1.17 (0.95,1.49)	0.16 (-0.06,0.39)	-0.17 (-0.49,0.05)	0.387
Improving Numeracy and Literacy Efficacy Trial –p41	English (*PiM raw*)	2.39 (1.71,2.59)	0.87 (0.54,0.95)	-1.39 (-1.59, -0.71)	0.273	6.40 (2.45,5.61)	1.85 (0.54,0.95)	-5.40 (-4.61, -1.45)	0.047
Improving Numeracy and Literacy Efficacy Trial –p41	Maths (*PiM raw*)	1.97 (1.53,2.19)	0.68 (0.42,0.78)	-0.97 (-1.19, -0.53)	0.008	1.81 (1.43,2.02)	0.59 (0.42,0.78)	-0.81 (-1.02, -0.43)	0.017
SHINE in Secondaries Efficacy trial -p46	*PiE raw*	1.03 (0.92,1.19)	0.03 (-0.08,0.18)	-0.03 (-0.19,0.08)	0.259	0.89 (0.78,1.04)	-0.10 (-0.08,0.18)	0.10 (-0.04,0.22)	0.741
Response to Intervention Efficacy Trial-p5	*NGRT B*	1.13 (0.9,1.34)	0.12 (-0.11,0.29)	-0.13 (-0.34,0.1)	0.259	1.13 (0.94,1.29)	0.12 (-0.11,0.29)	-0.13 (-0.29,0.06)	0.317
Changing Mindsets Efficacy Trial –p7	Maths (*PiE*)	1.04 (0.83,1.40)	0.04 (-0.19,0.34)	-0.04 (-0.4,0.17)	0.905	1.12 (0.84,1.52)	0.11 (-0.19,0.34)	-0.12 (-0.52,0.16)	0.427
Changing Mindsets Efficacy Trial –p7	Reading (*PiE*)	1.15 (0.96,1.39)	0.14 (-0.04,0.33)	-0.15 (-0.39,0.04)	0.705	1.10 (0.93,1.32)	0.09 (-0.04,0.33)	-0.10 (-0.32,0.07)	0.387
Graduate Coaching Programme Efficacy Trial-p73	*PiE raw*	0.95 (0.77,1.12)	-0.051 (-0.26,0.11)	0.05 (-0.12,0.23)	0.387	0.96 (0.78,1.13)	-0.03 (-0.26,0.11)	0.03 (-0.13,0.22)	0.273
Changing Mindsets Efficacy Trial –p78	English (*PiE*)	1.07 (0.95,1.21)	0.068 (-0.06,0.19)	-0.07 (-0.21,0.05)	0.35	0.98 (0.87,1.1)	-0.01 (-0.06,0.19)	0.01 (-0.1,0.13)	0.449
Changing Mindsets Efficacy Trial –p78	Maths (*PiE*)	0.89 (0.76,1.05)	-0.115 (-0.28,0.05)	0.109 (-0.05,0.24)	0.055	1.01 (0.87,1.14)	0.01 (-0.28,0.05)	-0.01 (-0.14,0.13)	0.549
Catch Up Numeracy Efficacy Trial – p9	Maths (catch)	1.15 (0.96,1.39)	0.14 (-0.04,0.33)	-0.15 (-0.39,0.04)	0.705	1.101 (0.93,1.32)	0.09 (-0.04,0.33)	-0.10 (-0.32,0.07)	0.387
Catch Up Numeracy Efficacy Trial – p9	Maths (Time)	1.28 (1.06,1.54)	0.25 (0.06,0.43)	-0.28 (-0.54,-0.06)	0.007	1.20 (0.99,1.45)	0.18 (0.06,0.43)	-0.20 (-0.45,0.01)	0.01

Results from the Difference-in-Difference approach

Table 2 presents the results from Difference-in-Difference (D-i-D) analyses.

The results for the raw data show that 'Improving Numeracy and Literacy efficacy trials', with maths as the outcome and 'Catch-Up Numeracy© effectiveness trial' with maths (Time) as the outcome have bigger variation in the intervention group than the control group. These two results are in agreement with the ANCOVA analysis. However, variation between the intervention and control group for 'Switch-On Reading efficacy trial' is not statistically significant. Furthermore, the 'Changing Mindsets efficacy trial', with Reading as the outcome, the 'Fresh Start pilot study', and the 'Catch Up Numeracy efficacy trial' with Maths (catch) as the outcome, show significantly greater variation in the intervention group than the control group. The differences range from 3% to 11%.

Analysis of the normalised data show similar results although there were two differences. 'Improving Numeracy and Literacy efficacy trials', with English as the outcome, generated a significant result whereas, for the 'Changing Mindsets efficacy trial', the p value rose above 0.05.

As with the ANCOVA results, most of the trials analysed appear to have bigger variation in the intervention group than the control group.

Table 2: Difference-in-difference analysis of variation between intervention and the control groups using multilevel model

		Actual Data					Normliased Data				
Trial	Outcome	Pre (SDc /SDt)	Post (SDc /SDt)	Post/Pre	1-Post/Pre	P-value	Pre (SDc /SDt)	Post (SDc /SDt)	Post/Pre	1-Post/Pre	P-value
Rapid Phonics Efficacy Trial – p13	*NGRT 3B*	0.95 (0.78,1.19)	0.98 (0.77,1.26)	1.0 3(0.76,1.38)	-0.03 (-0.38,0.24)	0.064	1.01 (0.85,1.23)	0.97 (0.80,1.21)	0.96 (0.74,1.21)	0.04 (-0.21,0.26)	0.078
Switch-On Reading Efficacy Trial – p2	NGRT B	1.02 (0.84,1.24)	1.00 (0.86,1.16)	0.97 (0.81,1.16)	0.03 (-0.16,0.19)	0.267	1.00 (0.88,1.14)	1.05 (0.93,1.19)	1.05 (0.90,1.19)	-0.05 (-0.19,0.10)	0.861
Word and World Reading Pilot Study-p26	*PiE*	1.05 (0.97,1.14)	1.14 (1.07,1.22)	1.08 (1.00,1.17)	-0.08 (-0.17,0.00)	0.126	1.06 (0.98,1.15)	1.14 (1.07,1.21)	1.08 (1.00,1.18)	-0.08 (-0.18,0.00)	0.107
Fresh Start Pilot Study-p31	*NGRT B*	0.97 (0.82,1.16)	0.94 (0.79,1.12)	0.97 (0.81,1.15)	0.03 (-0.15,0.19)	0.001	1.00 (0.87,1.15)	0.97 (0.84,1.10)	0.97 (0.85,1.10)	0.03 (-0.10,0.15)	0.001
Talk for Literacy Efficacy Trial-p32	*NGRT B*	0.94 (0.78,1.12)	0.94 (0.76,1.16)	1.01 (0.78,1.26)	-0.01 (-0.26,0.22)	0.791	0.94 (0.78,1.12)	0.94 (0.77,1.14)	1.00 (0.80,1.25)	0.00 (-0.25,0.20)	0.844
Improving Numeracy and Literacy Efficacy Trial –p41	English (*PiM raw*)	1.01 (0.93,1.09)	0.85 (0.71,1.13)	0.84 (0.67,1.14)	0.16 (-0.14,0.33)	0.061	1.00 (0.93,1.11)	0.85 (0.62,1.14)	0.85 (0.59,1.13)	0.15 (-0.13,0.41)	0.007
Improving Numeracy and Literacy Efficacy Trial –p41	Maths (*PiM raw*)	1.15 (1.08,1.25)	0.76 (0.66,0.93)	0.67 (0.56,0.83)	0.33 (0.17,0.44)	0.001	1.15 (1.08,1.26)	0.8 4(0.70,1.01)	0.73 (0.60,0.90)	0.27 (0.10,0.40)	0.001
SHINE in Secondaries Efficacy trial -p46	*PiE raw*	0.82 (0.75,0.91)	1.01 (0.90,1.13)	1.23 (1.08,1.39)	-0.23 (-0.39, -0.08)	0.493	0.9 7(0.85,1.11)	1.14 (1.01,1.28)	1.17 (1.01,1.37)	-0.17 (-0.37,-0.01)	0.736
Response to Intervention Efficacy Trial-p5	*NGRT B*	1.10 (0.95,1.25)	1.13 (0.97,1.32)	1.03 (0.89,1.22)	-0.03 (-0.22,0.11)	0.663	1.08 (0.96,1.21)	1.11 (0.99,1.25)	1.03 (0.92,1.18)	-0.03 (-0.18,0.08)	0.736
Changing Mindsets Efficacy Trial –p7	Maths (*PiE*)	1.27 (0.98,1.66)	1.26 (1.02,1.55)	1.00 (0.81,1.20)	0.00 (-0.20,0.19)	0.343	1.26 (1.03,1.53)	1.16 (0.96,1.40)	0.92 (0.75,1.14)	0.08 (-0.14,0.25)	0.681
Changing Mindsets Efficacy Trial –p7	Reading (*PiE*)	0.91 (0.77,1.11)	0.81 (0.70,0.94)	0.89 (0.75,1.06)	0.11 (-0.06,0.25)	0.001	0.93 (0.82,1.09)	0.86 (0.75,1.00)	0.93 (0.76,1.09)	0.07 (-0.09,0.24)	0.001
Graduate Coaching Programme Efficacy Trial-p73	*PiE raw*	0.88 (0.77,1.01)	1.00 (0.87,1.18)	1.13 (0.98,1.33)	-0.13 (-0.33,0.02)	0.982	0.85 (0.73,0.99)	1.00 (0.86,1.18)	1.17 (1.00,1.39)	-0.17 (-0.39,0.00)	0.938
Changing Mindsets Efficacy Trial –p78	English (*PiE*)	1.05 (0.98,1.14)	1.00 (0.93,1.09)	0.95 (0.89,1.03)	0.05 (-0.03,0.11)	0.844	1.10 (1.00,1.21)	1.03 (0.94,1.15)	0.94 (0.85,1.05)	0.06 (-0.05,0.15)	0.319
Changing Mindsets Efficacy Trial –p78	Maths (*PiE*)	1.25 (1.10,1.45)	1.17 (1.07,1.31)	0.94 (0.84,1.04)	0.06 (-0.04,0.16)	0.001	1.08 (0.99,1.21)	1.06 (0.96,1.17)	0.98 (0.87,1.10)	0.02 (-0.10,0.13)	0.627
Catch Up Numeracy Efficacy Trial – p9	Maths (catch)	0.91 (0.77,1.11)	0.81 (0.70,0.94)	0.89 (0.75,1.06)	0.11 (-0.06,0.25)	0.001	0.93 (0.82,1.09)	0.86 (0.75,1.00)	0.93 (0.76,1.09)	0.07 (-0.09,0.24)	0.001
Catch Up Numeracy Efficacy Trial – p9	Maths (Time)	0.8 6(0.71,1.03)	0.81 (0.70,0.93)	0.94 (0.79,1.14)	0.06 (-0.14,0.21)	0.001	0.90 (0.78,1.03)	0.87 (0.77,1.00)	0.98 (0.84,1.16)	0.02 (-0.16,0.16)	0.001

Why did the ANOCOVA and Difference-in-Difference approaches generate apparently contradictory findings?

It might seem puzzling, at first sight, to find that the two analytical approaches sometimes produced different results. There are several reasons for this.

Unlike the ANCOVA analyses, the D-i-D analyses do not assume equal variation between the intervention and the control groups at baseline. Instead, the model estimates the total variation in the intervention and the control groups separately for pre-and post-test scores.

Further, D-i-D artificially doubles the sample size by treating baseline data as outcome which make it more prone to false positives. Also, significant p-value from the likelihood ratio test in D-i-D means could mean that the SD is different between treatment groups as well as between the post- and pre-intervention period. Note that most of the trials that were significant in the ANCOVA approach were also significant in the D-i-D approach.

Conclusions

Most EEF trials focus on improving performance of school children, particularly those that are performing below average. However, improving performance does not necessarily mean reducing inequality in performance between pupils. It is therefore important to not just compare average performance between intervention and control groups in a randomised trial, but also to compare variation between the two groups. A promising intervention, like 'Improving Numeracy and Literacy efficacy' with maths as the outcome, may improve performance, but it may also widen the gap in performance between pupils if low performing pupils are not gaining as much as high performing pupils. Another possible explanation for widening the inequality gap between pupils may be how eligible

children are defined. Whilst it may be desirable to give the intervention to as many children as possible, if the purpose is to increase equality, then it is advisable to only give it to low performing pupils to reduce the gap between them and their counterparts.

However, our present understanding of variance change is not sufficiently advanced to set up trials which are specifically designed to impact on variance. We are unclear about the mechanisms which may cause the differences and are unsure about the necessary sample sizes for designing such investigations. This has three implications: firstly, it will be important to keep abreast of the international literature and, in particular, to see the results of the William T Grant funded work which was mentioned earlier. Secondly, as the EEF database grows an investigation of the results of completed trials using a meta-analytical approach would be useful. Finally, it would be helpful to conduct some simulation work to get a handle on the sample sizes need to detect shifts in SD of varying degrees.

Recommendations

We have applied two modelling approaches to compare variation between the intervention and the control group, but we would recommend an ANCOVA approach because of its simplicity and the fact that it is the same model mostly used for comparing average difference between groups. We expect variation at baseline to be comparable between intervention and control groups if randomisation is done properly and the sample size is sufficiently large. The second modelling approach, the Difference-in-Difference approach did not assume equal variances at baseline and it produced some contrasting findings from ANCOVA.

When commissioning and designing projects we suggest that, at the moment, given our present state of knowledge, the focus should be on the mean impact rather than the shift in SD. However, when recording the results, we recommend using the percentage change in the standard deviation of the intervention group when compared to the control group. For meta-analyses it makes sense to use the log of the variance ratio. We also recommend reporting 95% non-parametric confidence intervals.

Acknowledgments

We express our thanks to ZhiMin Xiao for his help in extracting the data and also to Lee Copping for his help with the manuscript.

References

Bartlett, M. S., & Kendall, D. G. (1946). The statistical analysis of variance-heterogeneity and the logarithmic transformation. *Supplement to the Journal of the Royal Statistical Society*, *8*(1), 128-138.

Blom, G. (1958) *Statistical estimates and transformed beta variables*. New York: John Wiley and Sons

Levene, Howard (1960). "Robust tests for equality of variances". In Ingram Olkin, Harold Hotelling, et alia. *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*. Stanford University Press. pp. 278–292.

Hedges, L. V., & Friedman, L. (1993). Gender differences in variability in intellectual abilities: A reanalysis of Feingold's results. *Review of Educational Research*, *63*(1), 94-105.

Katzman, S., & Alliger, G. M. (1992). Averaging untransformed variance ratios can be misleading: A comment on Feingold. *Review of Educational Research*, *62*(4), 427-428.

Raudenbush, S.R. and Bloom H.S. (2015a) *Learning **About** and **From** Variation in Program Impacts Using Multisite Trials* MDRC Working Papers on Research Methodology.

Raudenbush, S.R. and Bloom H.S. (2015b) *It's No Longer All about the Mean: Using Multi-site Trials to Learn About and From Impact Variation* William T. Grant post

Verbeke G, Molenberths G 2000 Linear mixed models for longitudinal data. New York: Springer-Verlag 32.

Weiss, M.J., Bloom, H.S. and Brock, T. (2013) *A Conceptual Framework for Studying the Sources of Variation in Program Effects* MDRC Working Papers on Research Methodology.