advantage in league table accountability measures, in response to the second bullet point above.

Each of these factors merit consideration here, to set the scene for how this evaluation evolved.

3.1. The increased use of effect sizes

The use of effect sizes has moved steadily from the routine use for meta-analysis (combining and comparing estimates from different studies) to gradually becoming a simple way of quantifying the difference between two groups [8]; [5]. It is a way of measuring the *extent* of the difference between two groups and allows an evaluation of impact to move beyond the simplistic 'Does it work?' to the more valuable insight of 'How *well* does it work across a *range* of contexts?'. As [5] point out: For these reasons, effect size is the most important tool in reporting and interpreting effectiveness, particularly when drawing comparisons about relative effectiveness of different approaches. The technical aspects relating to the use of effect sizes in this evaluation appear in the subsequent sections relating to 'Plan' and 'Action'.

3.2. The changes in school accountability measures in England

Recent changes in school accountability measures first came to prominence in the DfE performance tables for 2016. Rather than a focus on attainment, the key pupil outcome measure has become progress/value-added. Specifically, in primary schools this takes the form of:

- Average progress in mathematics
- Average progress in reading
- Average progress in writing
  [9]

Progress scores for primary schools are centred around 0, with most schools in the range of -5 to +5. A score of 0 means pupils in this school on average do about as well at KS2 (the end of the primary phase – age 11) as those with similar prior attainment nationally. A positive score means pupils in this school do better at KS2 than those with similar prior attainment nationally. A negative score means pupils in this school on average do worse at KS2 than those with similar prior attainment nationally.

That said, when we look at the more complicated progress measure used for secondary schools, 'Progress 8' (P8), it is the primary school progress measures of mathematics and reading that come to dominate, as these measures form the baseline from which progress in secondary schools is judged [10]. Progress 8 aims to capture the progress of a pupil from the end of primary school to the end of secondary school. It is a type of value-added measure, where pupils' results are compared to the actual achievements of pupils with the same prior attainment (average of **mathematics** and **reading** scores at the end of KS2). P8 is based on a calculation of pupils' performance across eight qualifications. These qualifications are:

- A double weighted **mathematics** element that will contain the point score of the pupil's English Baccalaureate (EBacc) mathematics qualification.
- An English element based on the highest point score in a pupil's EBacc **English language** or **English literature** qualification. This will be double weighted provided a pupil has taken both qualifications.
- An element which includes the three highest point scores from any of the **EBacc** qualifications in science subjects, computer science, history, geography, and languages. The qualifications can count in any combination and there is no requirement to take qualifications in each of the 'pillars' of the EBacc.
- The remaining element contains the three highest point scores in any three **other** subjects, including English language or literature (if not counted in the English slot), further GCSE qualifications (including EBacc subjects) or any other technical awards from the DfE approved list.

If a pupil has not taken the maximum number of qualifications that count in each group they will receive a point score of zero where a slot is empty. No legacy GCSEs (A*-G), International GCSEs or Level 1/Level 2 certificates in these subjects will count in performance tables once new GCSEs (9-1) in that subject are introduced. A score of zero means pupils in this school on average do about as well at key stage 4 (KS4) (end of the secondary education phase – age 16) as other pupils across England who got similar results at the end of KS2. A score above zero means pupils made more progress, on average, than pupils across England who got similar results at the end of KS2. A score below zero means pupils made less progress, on average, than pupils across England who got similar results at the end of KS2.

Given the scope and comprehensive nature of P8, it is clearly a more encompassing measure than the legacy measure of attainment, 5A*-C including English and mathematics.

3.3. 'Gaming the system'

In a letter to inspectors, [11], Ofsted's national director of education (UK), draws the attention of inspectors to the importance of following up unusual examination entry patterns, what is frequently termed as 'gaming the system'. Unfortunately, when an accountability measure is introduced (P8 for example) there is a risk that 'game theory' emerges [12]. [11] raises particular concerns in relation to what is viewed as increasingly common practices, namely:

- Schools which enter large numbers of pupils for qualifications that are not core subjects or do not reflect a school's specialisms – often subjects of a technical or vocational nature not suited to the majority of pupils.
- Double entry in qualifications that overlap in content. For example, statistics and free-standing mathematics qualifications; GCSE English and IGCSE[4] English as a second language qualification for pupils who have English as a first language.
- Schools which enter pupils for GCSE in English literature, without teaching the latter properly. Pupils sit the exam purely to ensure that the language result is counted doubly towards P8.
- Moving underperforming pupils into alternative provision so that they will not bring down results – a practice known as 'off-rolling'.

It appears that the motive for some schools is to boost their league table position rather than act in the best interests of the children. Clearly, there is a need to preserve authentic education for all our children. This authenticity lies at the heart of the values of 'Thinking Schools'. The 'setting' or context, outlined in this section, provided the springboard for the initial 'plan'. It is this 'plan' that features next.

4. Plan

Given the prominent use of effect sizes to judge the impact of educational innovations, the author decided to use this metric in order to retain a degree of consistency so that the results of the evaluation could be compared to research outcomes already available, particularly the findings of [4] and [5]. Thus, the original calculation of effect size followed that of Hattie [4] and is illustrated below.

English state schools were chosen, as opposed to non-state schools (or independent schools), to retain consistency in terms of statutory curriculum expectations. Due to the changing nature of accountability measures in England, it was decided that this metric be applied to progress scores in mathematics, reading and writing for primary schools, and progress 8 (P8) scores for secondary schools. Primary reading scores were chosen alongside mathematics progress scores as a main focus due to these outcomes being used as the baseline from which P8 is calculated – again the idea was to retain consistency and also the ability to analyse how the primary measures might influence P8 going forward. On that basis, the original plan was to conduct three separate effect size calculations to provide four measures of impact as follows:

- Impact on mathematics progress in primary schools (key focus)
- Impact on reading progress in primary schools (key focus)
- Impact on writing progress in primary schools (to provide a supplementary measure for comparative purposes)
- Impact on progress 8 (P8) scores in secondary schools (key focus)

In addition to using the common metric of effect size for this summative evaluation, the plan was to provide, through this evaluation, guidance to schools as to how they might use effect sizes formatively in an on-going way in order to judge their own impact.

[4] IGCSE – International General Certificate of Secondary Education. A series of exams taken at the end of secondary phase of education in international schools and English schools as an alternative to other secondary phase examinations.

$$\text{Effect size} = \frac{\text{[Mean of experimental group]} - \text{[Mean of control group]}}{\text{Average spread (standard deviation, or sd)}}$$

(The experimental group being accredited state 'Thinking Schools' in England and the control group being other schools in England.)

The idea here was to provide schools with a robust mechanism by which they could advance pupils' learning and in doing so undertake authentic school improvement without having to resort to *gaming the system.*

5. Action

Although the essence of the original plan was actually applied, three refinements were made. The first two refinements related to the actual effect size metric. Although using the pooled standard deviation (SD) to calculate the effect size generally provides a better estimate than the control group SD, [8] points to the issue of bias in that it generally gives a value slightly larger than the true population value. [13] give a formula which provides an approximate correction to this bias which was subsequently applied to the effect size formula used in the evaluation. This correction had the added benefit of allowing the calculated effect sizes to be compared to The Sutton Trust research [5] on an 'even', 'like for like' basis, as this research also corrected for this bias. In addition, due to the increased rigour allowed by this correction, high impact results could be viewed with increased status as results not corrected for this bias would tend to yield higher scores. The second refinement, also relating to the effect size metric, came in response to [8]'s recommendation that effect sizes should be calculated and reported with confidence intervals. The confidence interval for effect size is a measure of the significance of the effect size, taking into account the spread of the data and also the number of observations. A confidence interval that doesn't include zero indicates that there is a significant difference between the two groups. Another way of looking at this issue is that statistical significance does *not* tell you the key feature: the size of the effect. To overcome this, effect sizes are reported together with an estimate of their likely 'margin for error' or 'confidence interval'. In keeping with common research protocols [5], a 95% confidence interval was calculated and reported alongside the effect size. A 95% confidence interval also features in England's Department for Education (DfE) performance tables for both primary and secondary school progress data. Confidence intervals are presented as two numbers – the lower and upper limits within which we are 95% confident the true score may lie. This is a loose interpretation, but useful as a rough guide. The strictly correct interpretation of a confidence interval is based on the hypothetical notion of considering the results that would be generated if the study were repeated many times. If a study was repeated infinitely often, and each time a 95% confidence interval calculated, then 95% of these intervals would contain the true effect.

The third refinement to the original plan related to increasing the original 'pool' of schools to include schools which had 'registered' with TM as embarking on the Thinking Schools journey. Thus, two separate illustrations of impact were selected, namely:

- Accredited Thinking Schools
- All registered Thinking Schools (including accredited Thinking Schools)

The idea behind this extra dimension was two-fold. Firstly, it allowed the impact of taking a whole school approach to cognitive education to be evaluated more widely, by including schools which were well on the journey (accredited), together with schools who had committed to developing their practice in this way but who had yet to fully embed the innovation. Secondly, by adding this dimension, the 'progress' and 'gain' of moving to full accreditation could also be judged by comparing all registered Thinking Schools with those which had successfully become accredited. In short, it would provide a 'proxy' measure of the impact of pursuing accreditation.

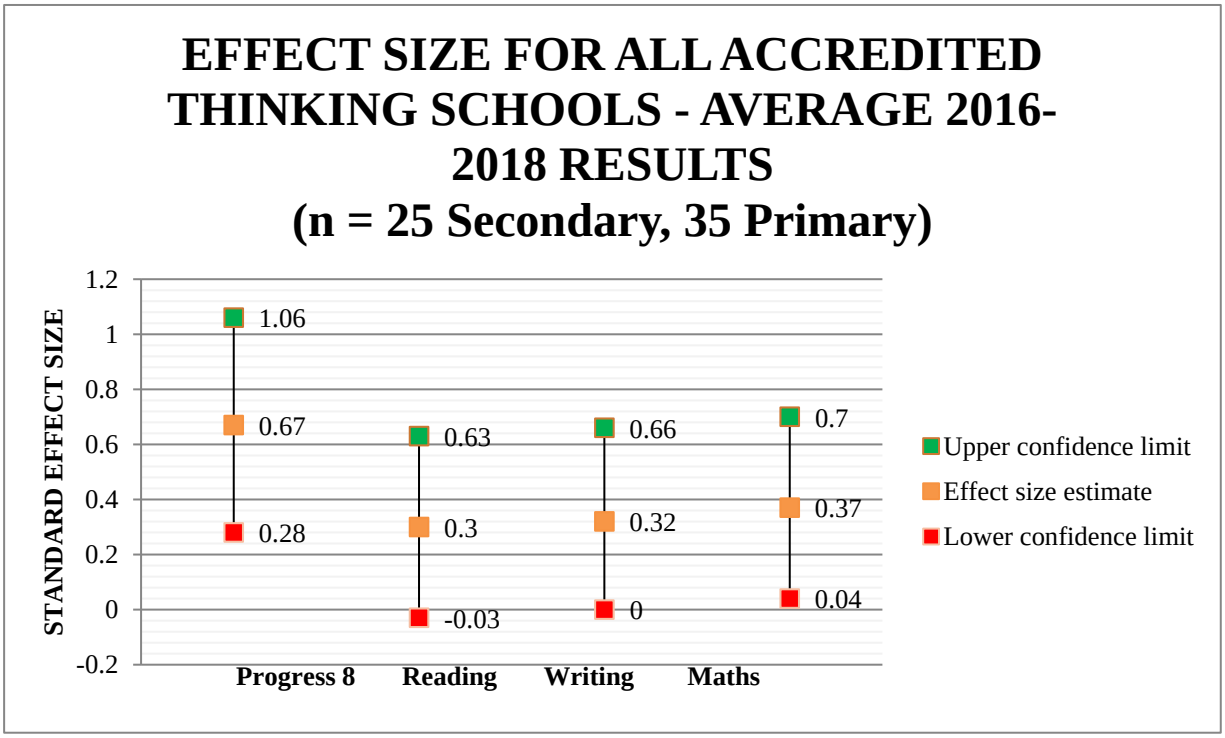
What follows next are the results (or reaction) together with guidance in the form of a lens through which to view the outcomes.

6. Results and Reaction

Table 1 presents the average effect sizes 2016 to 2018 for all accredited primary and secondary Thinking Schools and Figure 1 represents this graphically. See over:

Table 1. Average effect sizes 2016 to 2018 for all accredited primary and secondary Thinking Schools – (n = 35 primary and 25 secondary).

STANDARDISED EFFECT SIZE	Progress 8	Reading progress	Writing progress	Maths progress
Upper confidence limit	1.06	0.63	0.66	0.7
Effect size estimate	**0.67**	**0.3**	**0.32**	**0.37**
Lower confidence limit	0.28	-0.03	0	0.04



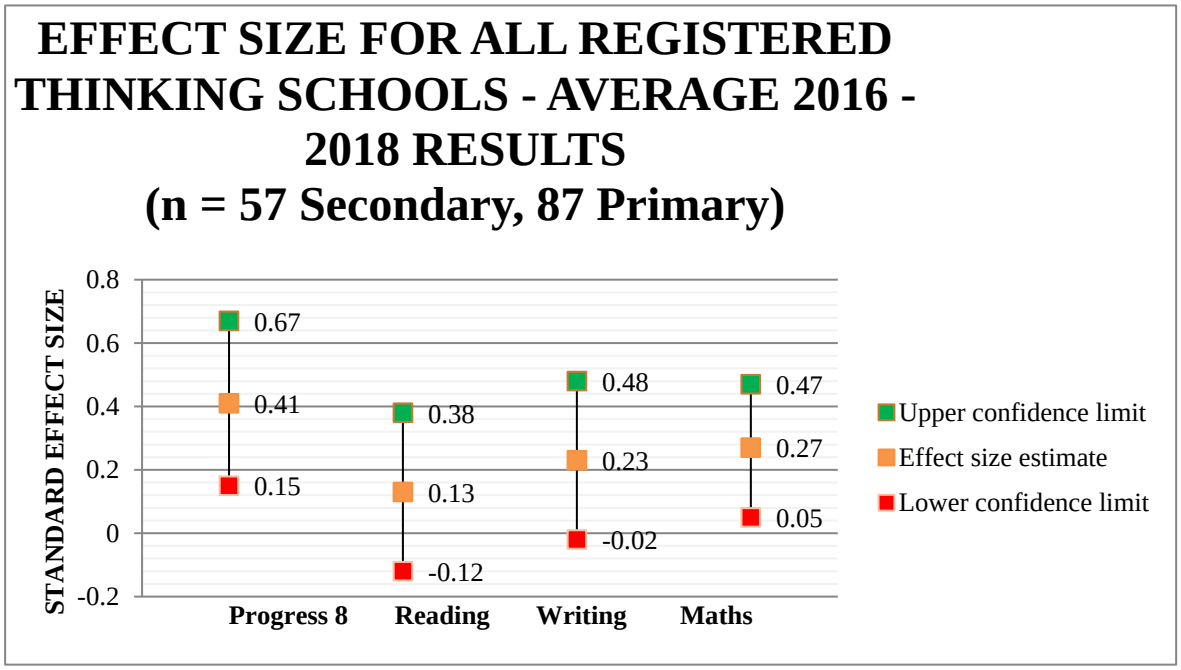
(Average effect size across all measures = 0.42)

Figure 1. Average effect sizes 2016 to 2018 for all accredited Thinking Schools.

Table 2. Average effect sizes 2016 to 2018 for all registered primary and secondary Thinking Schools (n = 87 primary, n = 57 secondary)

STANDARDISED EFFECT SIZE	Progress 8	Reading progress	Writing progress	Maths progress
Upper confidence limit	0.67	0.38	0.48	0.47
Effect size estimate	**0.41**	**0.13**	**0.23**	**0.27**
Lower confidence limit	0.15	-0.12	-0.02	0.05

Table 2 presents the average effect sizes 2016 to 2018 for all *registered* primary and secondary Thinking Schools and Figure 2 represents this graphically.



(Average effect size across all measures = 0.26)

Figure 2. Average effect sizes 2016 to 2018 for all registered Thinking Schools.

Before analysing and interpreting these data, it is important to understand some key benchmarks for effect sizes. What follows is a comprehensive lens through which to view the effect sizes in this evaluation based on these established benchmarks.

6.1. Effect size comparison/ interpretation data

So, we have a calculated effect size. How should we interpret this? To come up with an 'objective' and widely acknowledged benchmark we need to use three main considerations:

- When we look at many major longitudinal databases – the Progress in International Reading Literacy Study (PIRLS), the Program for International Student Assessment (PISA), the Trends in International Mathematics and Science Study (TIMSS), the National Assessment of Educational Progress (NAEP) and the National Assessment Programme – Literacy and Numeracy (NAPLAN), they all lead to a similar estimate of an effect size of **0.40** for a year's input of schooling. For example, using NAPLAN (Australia's national assessments) reading, writing, and maths data for students moving from one year to the next, the average effect size across all students is **0.40**.
- The average of 900+ meta-analyses carried out by [4], based on over 250 million students show an average intervention effect size of **0.40**.
- The Sutton Trust Education Endowment Foundation also indicates the overall distribution of effects found in education research with an average around **0.40** [5].

Therefore, an effect greater than **0.40** is seen as above the norm and indicates a more than anticipated impact. In other words, the innovation is working better than expected. Further, [14] describes an effect size of **0.29** as not being perceptible to the naked eye and equal to the difference between the height of a 5' 11'' and a 6' 0'' person. In addition, [14] describes an effect size of **0.50** as being perceptible to the naked eye and therefore 'medium'. [14] goes on to describe an effect size of **0.8** as 'grossly perceptible and therefore large', equating it to the difference between the heights of 13-year-old and 18-year-old girls.

6.2. Converting effect size to grade growth

If we then consider how effect sizes relate to actual improvements in examination grades, the following interpretation may prove useful by way of an illustration. If we take Coe's (2002) analysis that the distribution of GCSE grades in compulsory subjects (ie English and mathematics) have standard deviations of between 1.5 – 1.8 grades, so an improvement of one GCSE grade represents an effect size of 0.5 – 0.7. In a secondary school therefore, introducing an innovation whose effect size was known to be 0.6 would be likely to yield an improvement of about a grade for each pupil in each subject. For a school in which 50% of pupils were previously gaining a grade 5 or more in English and mathematics (or indeed overall), this percentage (other factors being equal, and assuming that the effect applied equally across the range of subjects offered) would rise to 73%! For a school with a prior progress 8 score of zero (the national average – pupils' grades were in line with national grades for pupils with the same starting points), progress 8 would move up to 1. This would place the school in the 'Outstanding' inspection category in England for this significant measure. For a school with a prior progress 8 score of -0.5 (pupils' grades being half a grade lower than national grades for pupils with the same starting points) and in an 'Inadequate' inspection category in England, progress 8 would move to 0.5 (pupils' grades being half a grade higher than national grades for pupils with the same starting points) and place the school in the 'Good' inspection category in England. Further, for a school with a prior progress 8 score of 0.5 (pupils' grades being half a grade higher than national grades for pupils with the same starting points) and in a 'Good' inspection category in England, progress 8 would move to 1.5 (pupils' grades being a grade and a half higher than national grades for pupils with the same starting points) and place the school in a notionally 'Exceptional' inspection category in England. This pattern of increases in attainment and progress would apply for primary assessment measures as well as international student outcome measures in compulsory subjects or subjects with large percentage entries.

6.3. Effect size of metacognitive strategies

The following benchmark figures summarise two main meta-analyses:

- [4] - Effect size = **0.69**
- [5] – Effect size = **0.62 to 0.69**

6.4. And so how might we interpret the results of this study given the afore-mentioned lenses of interpretation?

Table 1 and Figure 1 show the average of these results across the 2016 to 2018 period for accredited Thinking Schools and indicate what would be termed 'High' impact for P8 (0.67), yielding over a whole grade extra growth. The impact measures for primary reading, writing and mathematics would be viewed as 'Moderate' overall (around 0.5 extra grade growth). The confidence intervals for P8 and

mathematics progress *do not* include zero and indicate that these results are *significant*. The confidence intervals for primary reading and writing progress tip slightly towards the inclusion of zero and therefore are very 'close to' being significant overall.

Table 2 and Figure 2 show the results across the 2016 to 2018 period for all registered Thinking Schools (inclusive of accredited and those pursuing accreditation) and indicate what would be termed 'Moderate' for P8 (0.33), yielding over a third of a grade extra growth. The impact measures for primary writing and mathematics progress would be viewed as 'Moderate' overall (also around a third of a grade extra growth). Once again, the confidence intervals for P8, writing and mathematics progress *do not* include zero and therefore indicate that these results are *significant*. The confidence intervals for primary reading progress tips towards a positive and so therefore also leans towards being significant.

If we compare the data for accredited Thinking Schools with that of all registered Thinking Schools for the period 2016 to 2018, we can see that the picture is one of clear and significant growth in terms of impact (all effects show a marked increase from registered to accredited status). In addition, primary reading and mathematics progress impact measures show growth from registered to accredited status. In addition, secondary P8 impact and primary reading impact shows clear growth from 2016 to 2017. It should be noted that 2018 student outcome accountability data methodology changed and so impact on trend analysis.

The following section adds more of a story to these initial interpretations and offers some overall evaluative conclusions.

7. Evaluation (How can we effectively develop thinking in specific contexts?)

By referring back to each of the previous sections of the SPARE wheel model, it is possible to draw some tentative conclusions and recommendations for further consideration.

Without attempting to draw *absolute* cause-effect claims, the evidence suggests that taking the Thinking Schools approach impacts greatly on the progress of pupils in both primary and secondary schools in England as measured by P8, reading progress, writing progress and mathematics progress. Indeed, the effect sizes are consistent with the high impact on pupils' achievement of meta-cognitive strategies illustrated by the research of [4] and [5]. The very high impact on P8 is of particular note as this measure spans a *wide* range of subject disciplines, not just the traditional canon of achievement represented by mathematics and English.

Comparing the data for accredited Thinking Schools with registered Thinking Schools illustrates clear impact growth in all progress areas examined, once the criteria for accreditation has been met. This would tend to reflect the importance of fully embedding cognitive education, a process that would normally take at least three years given the profile of the schools in this study. The very high impact on P8 would tend to suggest that further research into possible 'latent' achievement development, as pupils move through primary and secondary phases, may further add to the body of knowledge in the areas covered by this study.

Given the importance of the context for specific schools and that schools naturally apply the Thinking Schools approaches in different ways, there is a need to not only share 'what works' *generally* across all schools but also enable schools to develop 'what works' for their own *particular* context. For example, under the original model of school training in the use of cognitive tools provided by TM, schools were provided with three, interrelated pathways to pursue:

- Visual Tools for Thinking – tools that explicitly support thinking processes
- Dispositions for Mindfulness – intelligent learning behaviours
- Questioning for Enquiry – skills for effective questioning and enquiry

The selection of pathways, including the order of progression, was left for the schools to decide. Some would start with Questioning for Enquiry and then move onto Visual Tools for Thinking. Others would start with Visual Tools for Thinking and then move onto Dispositions for Mindfulness. The corollary of the pathway model has been that schools develop in a rather disjointed manner where the cognitive tools are seen as discrete components. Further, schools began to describe themselves as a Visual Tools school or Questions for Enquiry Schools. In short, schools were showing good impact as illustrated by this evaluation but were adopting a narrow approach to the teaching of thinking. TM, having recognised this and in their pursuit of building further on the already impressive impact of their approach, have created a more integrated model of teaching thinking that is based on the dynamic and interrelated nature of the thinking process and the use of Grounded Practice principles [15].

In addition, through the sharing of best practice offered by 'hub' Thinking Schools, and application

of formative evaluative processes to compliment the more summative evaluation offered by accreditation (although it is advisable to use this more 'summative' evaluation in a formative style on a three year cycle), schools can monitor what works best in their *own* specific context. This new wave of development brings cognitive education and assessment together in order to secure self-improving organisations with a common conception of impact, independent but embracing of any accountability measure. Given the higher impact on achievement for accredited Thinking Schools, where formal training in cognitive education has been undertaken, this also needs consideration if schools are to fully realise the potential of taking a whole school approach to the teaching of thinking.

Finally, the picture of impact provided by this evaluation of 2016 to 2018 national student outcomes in England shows a relatively stable positive picture for accredited and registered Thinking Schools. This would tend to suggest that this positive impact is not cohort dependent as the data is representative of different cohorts of pupils. Also, given the increased rigour and demand of the new 9 – 1 GCSE grading system in secondary schools in England, and that most secondary schools saw their results drop in 2017/18 [16], the stability apparent in the Thinking Schools would tend to indicate that they were not impacted negatively by this change of the examination system. Together, these factors point to positive resilience to large-scale external changes in the educational landscape.

7.1. Future avenues of enquiry

Although this study presents a snapshot of 2016 to 2018 outcomes in England and does not track all individual schools' development as they move through the Thinking Schools process, continuing this evaluative approach would extend our understanding of how Thinking Schools develop in a changing national and international educational landscape. Further, it would also allow the new TM model of teaching thinking to be monitored and evaluated as it gains ground.

In addition to continuing with an ongoing snapshot evaluative approach, a future research focus in England would seem to lie in a longitudinal study of how pupils and schools develop over time. Evidence of long lasting 'far transfer', in terms of accelerated pupil growth, as indicated by multiple indicators (not *just* academic achievement), would be of particular interest here.

On a more international level, the challenge would be to replicate the evaluation offered by this study more globally using the key outcome measures specific to particular countries. Again, once a snapshot of impact has been illustrated, countries across the globe may wish to fund more longitudinal studies in order to grow their Thinking Schools further in their own particular contexts. The following case studies chart the three year growth of three state secondary thinking schools and three state primary thinking schools provide a stimulus for 'getting underneath' the positive headlines to gain an insight into contextual influences that impact positively on the achievement of pupils. They span a variety of contexts, particularly in terms of the percentage of 'disadvantaged pupils' on roll. In short, future research questions seem to centre on the nature of specific school culture and success in growing a thinking school.

8. References

[1] R.L. Burden, "Illuminative evaluation", *Educational and Child Psychology*, 15 (3), 1998, pp. 15-23.

[2] P.D.Ellis, *The Essential Guide to Effect Sizes: Statistical Power, Meta-Analysis and the Interpretation of Research Results*, London: Cambridge, 2010.

[3] R.L. Burden and L. Nichols, "Evaluating the process of introducing a thinking skills programme into the secondary school curriculum", *Research Papers in Education*, 15 (3), 2000, pp. 293-306.

[4] J.A. Hattie, Visible Learning – *A Synthesis of over 800 Meta-Analyses Relating to Achievement*, London: Routledge, 2009.

[5] S. Higgins, et.al., *The Sutton Trust – Education Endowment Foundation Teaching and Learning Toolkit: Technical Appendices*, London: The Sutton Trust, 2013.

[6] G. Stobart, "The validity of formative assessment", in J. Gardner (ed), *Assessment and Learning*, London: Sage, pp. 133-146, 2006.

[7] S. Clarke, *Formative Assessment in the Classroom*, London: Hodder Murray, 2005.

[8] R. Coe, "It's the effect size, stupid: What effect size is and why it is important", *Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, UK*, 2002, pp. 12-14.

[9] DfE, *Primary Progress Measures: How the Primary progress Measures are Calculated*, London: DfE, 2016.

[10] DfE, *Progress 8 and Attainment 8: Guide for Maintained Secondary Schools, Academies and Free Schools*, London: DfE, 2017.

[11] S. Harford, *School Inspection Update*, London: Ofsted, 2017.

[12] M. Waters, *Thinking Allowed on Schooling*, London: Crown House Publishing, 2013.

[13] L. Hedges and I. Olkin, *Statistical Methods for Meta-Analysis*, New York: Academic Press, 1985.

[14] J. Cohen, *Statistical Power Analysis for the Behavioural Sciences* (2nd ed.), Hillside, New Jersey: Lawrence Erlbaum Associates, 1988.

[15] D. Walters, "Grounded practice: putting the 'self' back into self-evaluation", *Educational Action Research*, 22 (1), 2014, pp. 72-92.

[16] M. Treadway, Key stage 4 performance tables: Closing the gap just got harder, (https://educationdatalab.org.uk/2018/01/key-stage-performance-tables-2017-closing-the-gap-just-got-harder/), London: FFT Datalab, 2017.

