Churches, R., Dommett, E., Devonshire, I.M., Hall, R., Higgins, S. & Korin, A. (Accepted) Translating laboratory evidence into classroom practice with teacher-led randomised controlled trials - a perspective and meta-analysis. *Mind, Brain and Education* (forthcoming)

1.
2. Complete title of the manuscript

Translating laboratory evidence into classroom practice with teacher-led randomised controlled trials - A perspective and meta-analysis

3. Short running title (not more than 60 characters)

Translating laboratory evidence into classroom practice

4. Name of all the authors, their affiliations including institution or university and email address

Dr Richard Churches

Education Development Trust

rchurches@educationdevelopmenttrust.com

Dr Eleanor J. Dommett

Institute of Psychiatry, Psychology, and Neuroscience (IoPPN), King's College London

eleanor.dommett@kcl.ac.uk

Dr Ian M. Devonshire

University of Nottingham Medical School

Ian.Devonshire@nottingham.ac.uk

Robin Hall

British Science Association, Education Section

Robin.HALL@education.gov.uk

Professor Steve Higgins

School of Education, Durham University

s.e.higgins@durham.ac.uk

Astrid Korin

Education Development Trust

akorin@educationdevelopmenttrust.com

5. Contact details for the corresponding author

Dr Richard Churches, Education Development Trust, Reading RG1 4BS, England; e-mail: rchurches@educationdevelopmenttrust.com

6. Funding information

This research was funded by the Wellcome Trust with a grant.

7. Disclosure statement

Dr Richard Churches and Astrid Korin are employees of Education Development Trust, who received the grant from the Wellcome Trust. Dr Eleanor Dommett and Dr Ian Devonshire were paid consultant fees to work on the programme by Education Development Trust. No financial arrangement existed with Professor Steve Higgins and Robin Hall who gave their time pro bono to contribute to the journal article and make independent contributions to the meta-analysis and analysis of the results. On the training programme teachers were given books to use as textbooks by Dr Richard Churches, Dr Eleanor Dommett and Dr Ian Devonshire (*Teacher-Led Research* and *Neuroscience for Teachers*), these books are published and marketed commercially by Crown House Publishing in the UK and US and by Hawker Brownlow in Australia.

RESEARCH ARTICLE

Submission for: Special issue SIG22 conference 2018, *Mind, Brain, and Education* journal

Translating laboratory evidence into classroom practice with teacher-led randomised controlled trials - a perspective and meta-analysis

<Footnotes 1 – 6, author names removed for review>

ACKNOWLEDGMENTS

We would like to thank the Wellcome Trust for funding this project alongside five other science of learning projects. We would also like to thank the teachers and schools in England that were involved in the project, who gave their time so freely and without remuneration.

Abstract: 149/150 words

KEYWORDS

Education, translation, evaluation

ABSTRACT

We initiated and structured a single programme that supervised teachers, some with neuroscience or psychology degrees, to collaborate and explore the effects of science of learning-translated pedagogy. This paper reports on the 34 findings from teacher-led randomised controlled trials (RCTs) and replications. Teachers designed trials, looking at areas such as attention, memory, and spaced learning. Overall, positive effects were found over short periods [one to six weeks] (r = 0.15, p < .0001 [d = 0.30], N = 2,157). However, retrieval practice (testing as a learning experience) had differential effects mediated by age, approach, and lesson content. Results suggest science of learning-translated pedagogy needs extensive replication to establish how best to use laboratory evidence in classrooms. Multiple planned replication of teacher-led RCTs has potential as an

evaluation tool, combining high levels of mundane realism with strong internal validity and the potential to build cost effective large samples for meta-analysis.

<Research Article text: 4,687. Extension to word length agreed by email to account for revisions and additional detail>

THE TRANSLATION OF LABORATORY EVIDENCE INTO CLASSROOM PRACTICE – PROBLEMS AND POSSIBILITIES

The idea that neuroscience, cognitive psychology and education could form effective partnerships is not a new one. The suggestion was originally met with much excitement in the 1960s (Gaddes, 1968), when it was proposed that neuropsychological approaches could be integrated into diagnostic understanding and educational planning for children with learning disorders. This was followed by a further flurry of enthusiasm, with Fuller and Glendening (1985) proposing that so-called 'neuroeducators' should serve both communities to apply knowledge about the brain to the learning process. Later, Bruer (1997) argued for the development of 'an interactive, recursive relationship among research programs in education, cognitive psychology, and systems neuroscience' (p. 15).

Twenty years on, although the volume of evidence related to the neuroscience and cognitive psychology of learning has increased exponentially (Churches, Dommett & Devonshire, 2017; Howard-Jones, 2010; 2014), use of such knowledge in everyday teacher practice remains limited. Of course, there are many reasons why this is the case - including the general challenges associated with teachers engaging with education research evidence (McAleavy, 2016). Despite this, it is also clear that teachers are highly enthusiastic about engaging with neuroscience and the science of learning, as was shown in research commissioned by the Wellcome Trust (Simmons, 2014). This suggested that eight out of ten teachers would collaborate with neuroscientists doing research in education. Alongside this, some important emerging links are beginning to be forged (see Brookman-Byrne and Thomas, 2018).

Many challenges exist regarding collaboration between neuroscientists and educators (Churches, Dommett & Devonshire, 2017; Dommett & Devonshire, 2010; Dommett et al., 2013). Three key ones are particularly relevant here. First, on a theoretical level, education and neuroscience are fundamentally different in their overall objectives and the way these objectives are pursued (the ‘goal problem’). Neuroscience is a natural science that investigates the workings of the brain, functional architecture of the mind and the way that the brain and mind map together. In contrast, education aims to develop pedagogies; and therefore, arguably, has more in common with the way in which medicine uses biology to ground research and practice during the development of treatments; or the way in which architecture uses physics.

Secondly, neuroscience research can take place at several levels (Figure 1) not all of which have the same applicability to classrooms. The lowest microscopic level of analysis neuroscience can look at is individual genes and molecules. At the highest level of analysis, a neuroscientist may examine the workings of the whole brain in – most commonly but not exclusively – healthy adults and often in laboratory settings. In contrast, education research often starts at the level of the individual and progresses up to examine social processes, culture, and meaning (at levels above those illustrated in Figure 1 – in other words, the two professions only meet in passing at the behavioural level).

Thirdly, and by extension, there is a translation problem. Specifically, the outputs of neuroscience research often translate poorly into something that is useable in education. For instance, knowing a brain region is important for a certain skill does not actually tell you what to do about that in an educational setting. However, we believe neuroscience has a future in education akin to the way biology underpins medicine; and architecture is underpinned by physics. In both cases, although there is great scope for creative and developmental practice, neither profession can function against the laws of their respective fields. Although all the parallel principles are far from established yet in educational neuroscience, it can only be a matter of time (with the right research) before a framework can be established that can help teachers to ground their work in the biology of learning. In turn, as with medicine (and its foundations in biology), understanding the core biology of learning and related cognitive neuroscience, could help teachers avoid the sort of blind allies and mistakes in practice that are increasingly recognised as 'neuromyths' (see Howard-Jones, 2017; Kim and Sankey, 2017, for discussions). The key focus of this paper is therefore to explore whether small-scale, teacher-led RCTs, based on findings from neuroscience and cognitive scientists, are feasible for teachers to run in the classroom and what impact they might have on the learning of the students in those classes.

Bridging the gap between classroom 'clinical practice' and the science of learning evidence – can from 'bench to bedside' become from 'neuroscience lab to whiteboard'?

As we outlined above, a critical challenge facing neuroscience and education is how to translate evidence from laboratories into classrooms (Dommett & Devonshire, 2010; Dommett et al., 2013). From the mid-nineteenth century, similar challenges faced the medical profession as it aspired to become a 'natural science' grounded in biology. Firstly, laboratories are not classrooms, just as biological experiments are not clinical practice. Secondly, wide replication to control for pupil individual differences as well as school context is likely to be essential. However, the challenge is greater than just the need to accumulate real-life classroom evidence. It includes dealing with wider and deeper challenges within the education system that affect the way in which knowledge is mobilised and evidence is disseminated and used (see for a recent discussion, Coldwell et al., 2017).

Most notably, there is a 'democratic deficit' in education research (Biesta, 2007). In contrast to medicine and healthcare, where often serving clinicians publish studies on clinical practice, in education, few practitioner studies reach journals or get disseminated. Further, those researchers who do study or design pedagogy often no longer practice as teachers. This problem is compounded by the fact that most teacher education has, until recently, limited itself to teaching teachers to do small-scale qualitative action research, rather than controlled quantitative evaluation. This deficit, we would argue, has had effects across the spectrum of evidence use and leaves a key gap between 'laboratory and whiteboard' - to adapt the phrase used in medicine and healthcare (from 'bench to bedside' (Horton, 1999)).

In medicine and healthcare, the 'bench to bedside' concept expresses the linkage between basic laboratory research, through applied research and evaluation of programs, to evidence-based treatments and training to deliver them. Of course, there are also processes and structures at the system level to support this, that are also lacking in education. For example, in health, organisations like the National Institute for Health Research in England work to link the different organisations in the translation pipeline (NIHR, 2016/2017), with similar institutions taking on a parallel role in other countries. Such networks ultimately mobilise a global sense of endeavour around the best available clinical evidence. Underpinning these structures, is the way in which the serving medical profession itself is deeply engaged in the creation of the clinical practice evidence - ensuring the buy-in of the profession.

With the aim of directly exploring how this gap in the education system could be filled, we sought to place teachers directly in the 'clinical researcher space', between laboratory and whiteboard. So, in a Wellcome Trust funded project, delivered by Education Development Trust, teachers who previously conducted randomised controlled trials (Churches, 2016; Churches, Higgins & Hall, 2017), and teachers with a psychology or neuroscience degree, came together to design and deliver a series of trial protocols.

The neuroscience-informed, teacher-led randomised controlled trial project

A total of 31 individual schools, and Teaching School Alliances (a program set up to parallel the role of Teaching Hospitals (DfE, 2010)), were involved in this project. These included five Education Endowment Foundation/Institute for Effective Education Research Schools (Gu et al., 2018). The schools received an RCT design day in October 2017 and pre-reading material about RCT design (Churches & Dommett, 2016) and the neuroscience and cognitive psychology of learning (Churches, Dommett & Devonshire, 2017)[7]. Specifically, 15 distinct trial protocols were implemented, completed and analysed. These were replicated with 16 different pupil age groups (and/or with content variations) in parallel, but within identical treatment windows. Conference posters reported replications on the same poster. The aim of the analysis presented in this paper is to establish proof of concept for, and the viability of, multiple planned teacher-led RCTs reported within a single meta-analysis. The overall estimate of effect on learning outcomes can also provide an indicative estimate of the possible benefit from this approach.

Teacher research protocols were signed off by an educational neuroscience team and informed consent included the head teachers and principals of the schools. Over four months, the teachers implemented their RCTs, coming together again in February 2018 for an analysis, interpretation, and write up day. They then produced conference posters to support the dissemination of their findings. These posters have now been presented at a wide range of research conferences[8]. The educational neuroscience team consisted of the two neuroscientists and an education expert. This team also delivered the initial one-day training programme to support translation of the evidence covered in supplied pre-reading material (Churches & McAleavy, 2015; Churches & Dommett, 2016; Churches, Devonshire & Dommett, 2017). This group also reviewed individual RCT protocols and gave feedback on the intervention and control conditions. The trials were led by teachers that had previously attended similar randomised controlled trial training. The teachers chose the areas for research, designed the protocols and all research materials. The teachers also decided on, or created, appropriate dependent variables.

In the same way that clinical practitioners might design the clinical application of a new drug within a trial process, because they are better placed to consider the in-practice aspects of a treatment, such a model could be used in the future to define the relationship between education practitioners and the purely academic researcher. This does not, however, imply that academic roles might become obsolete. Rather evidence from teacher-led trials could be used to feedback into purely academic research as part of ongoing hypothesis testing.

Most of the teachers previously designed and implemented RCTs as part of a Department for Education program (in England), *Closing the Gap: Test and Learn* (Churches, 2016). It was therefore not surprising that a very wide range of research designs were chosen by the teachers, to test an equally wide range of hypotheses. Studies included both between-participant (independent measures) designs and within-participant (repeated measures) designs, with both two and three levels to the independent variable. Some teachers chose to adopt matched pair designs and, in one case, a 2 x 2 factorial design. A wide range of randomisation strategies were also used. These included case-matching prior to randomisation and stratified randomisation. Treatment windows were short compared to most RCTs implemented in education with the teacher trials varying in length from single lesson studies (Bryant-Khachy, 2018a; 2018b) to 42 days between pre-test and post-test (Siddle, 2018). In total, 2,157 children were able to be included in the analyses with an overall attrition rate of 12.6%. In many of the shorter trials there was no attrition.

META-ANALYSIS OF THE TEACHER FINDINGS

In undertaking a meta-analysis of impact on learning the aim is to find out whether these small-scale teacher-led RCTs, based on findings from neuroscience and cognitive science, are beneficial for the learning of the students in those classes. Meta-analysis is a statistical method that allows for the synthesis of quantitative evidence from related research in a way that can summarise that body of evidence. Statistical synthesis takes one of two forms: a fixed effect model or a random effects model. In the fixed effect model the assumption is that all the results that have been combined are an estimate of the same 'fixed' treatment effect. In contrast, a random effects model allows for difference in treatment effects between studies (Riley, Higgins & Deeks, 2011; Higgins, 2018). In this case a random effects model is clearly indicated (Borenstein, 2009).

Figure 2 shows a forest plot for all 34 results from all of the trials and year group replications that were completed by the teachers (k= 34; *N* (of students) = 2,157). The plot was constructed using spreadsheets from Suurmond and colleagues (2017). Most of the trials produced data that were not normally distributed and, as a result, the effect sizes were calculated using Rosenthal's *r* rather than Cohen's *d* (Rosenthal, 1994)[9]. Where other effect size measures were used in the trials, such as Hedges' *g* or η_p^2, the effect sizes were converted to Rosenthal's *r* so that the trial results used in the meta-analysis were presented on the same scale.

The trial results display a high level of heterogeneity, with $p_Q < .0001$ and $I^2 = 79.9\%$. Meta-analysis was carried out using a random effects model (which was indicated by the variation in interventions and the high level of heterogeneity. Unlike the fixed effect model, the random effects model does not assume that all the effect sizes are estimates of a single true effect. Despite the present analysis's heterogeneity, it is still useful and valid to bring these studies together in a single meta-analysis as these trials were all part of the same overall study investigating the effect of teachers designing their own pedagogical approaches, based on neuroscience and cognitive psychology evidence from the same source. As Hak and colleagues note, 'the assumptions underlying meta-analytic hypothesis testing in the social sciences will usually not be met under real-life conditions. This is the reason why meta-analysis is increasingly conducted with a different aim, based on more realistic assumptions. . .to explore the dispersion of effect sizes (2016, p.1). In interpreting the findings in the meta-analysis, we have adopted a similar position. The spread is as important as the pooled effect. Here the overarching 'intervention' is conceived of as the support provided to teachers to design and implement small-scale RCTs with outcome measures appropriate for the ages of pupils and the curriculum subjects being taught by the teachers involved.

The findings are grouped by type of intervention and include instances where replications adopted the same protocol but with different ages of children. Each point represents the effect size. The error bars represent the 95% confidence intervals. The relative size of the dot shows the weighting assigned to that effect size within the random effects model. Positive effect sizes, to the right of the chart ($r > 0.00$), indicate that the treatment resulted in improved pupil outcomes compared to the control, while negative effect sizes to the left indicate that the control group performed better. The equivalent standardised mean difference (Cohen's *d*) is presented in the final column[10]. On the left-hand side of the plot is a narrative summary of the area that was being explored in the trial; on the opposing side, information about the school year the children were in and the subject area[11].

When meta-analyses are included in systematic reviews, an analysis of publication bias is usually also included. This is to account for the fact that not all trial results get published. There may be trial results that were not available to the researchers that could affect the combined effect size produced by the meta-analysis. In this case, all the results of the completed trials were included in the meta-analysis and so publication bias is not relevant as there are no missing trial results.

Overall, teacher translations of the neuroscience and cognitive psychology evidence had positive significant effects on pupil outcomes ($r = 0.15$, 95% CI [0.09 – 0.21], $p < .0001$), an effect size equivalent to a Cohen's *d* of 0.30. A clear majority of treatment effects were positive (85.3%). The largest positive effect, associated with a single strategy, was found for the use of novelty to enhance salience and attention ($r = 0.41$, [$d = 0.90$]) (Morris, 2018). The largest negative effect ($r = -0.38$, [$d = -0.82$]) was for multiple choice testing alone as a means of learning new spellings compared to Look, Cover, Write, Check (LCWC) (a simple strategy using multiple rehearsal in working memory) (Baker and Hindley, 2018a). However, combining LCWC with multiple choice testing completely reversed this effect, producing the largest overall effect ($r = 0.51$, [$d = 1.19$]). The difference in approaches being evaluated and the differing outcome measures used mean that some caution is needed in interpreting the overall pooled effect, as is also indicated by the high overall heterogeneity. This pooled effect needs to be interpreted at the overall level of aggregation of the meta-analysis (Higgins, 2018) and indicates that, on average, the neuroscience/cognitive psychology informed interventions were effective in improving educational outcomes for students in the intervention groups, compared with the controls. Separate meta-analysis of the retrieval practice related protocols yielded a smaller overall pooled effect (r = 0.14, 95% CI [0.06 – 0.23] [d = 0.28], p = .001).

It is early days for the use of this type of approach and, as noted, the results should be interpreted with a degree of caution. That said, it is worth drawing out some preliminary conclusions. It is noteworthy that retrieval practice (the term now commonly applied to the use of 'testing as a learning experience' (Bjork & Bjork, 2011) had differential effects. These differences in effect appeared to be mediated by factors such as pupil age, subject area and the way in which testing as a learning event were applied. This contrasts with the evidence from a wide range of predominantly laboratory-based cognitive psychology research supporting the approach (Adesope, Trevisan & Sundarajan, 2017). Makarova's study (2018) suggests prior attainment might mediate outcomes. As discussed above, Baker and Hindley (2018a) showed that, for the learning of new spellings, LCWC was better than multiple choice testing alone. However, combining retrieval practice with LCWC produced stronger effects.

Here, perhaps, lies the nub of the problem. As we have pointed out before, the laboratory is not the classroom and thus by extension a laboratory protocol is not necessarily an effective pedagogy. Retrieval practice, particularly in the form of multiple-choice testing, is often operationalized in a way that would be of little direct benefit in the classroom over a whole one-hour lesson period, for example. Bluntly expressed, and because of the complexity of much learning that takes place in the classroom, no serving teacher practitioner would advocate learning just though the use of tests. Whereas, often in laboratory-based retrieval practice studies, the experimental condition is exactly that – the use of just testing around a very straight-forward task - which is then compared to no testing, to see if there has been improvement in performance on that task.

A whole lesson, to be effective or enhanced, might well include a test but would always need to include elements that appear in 'clinical practice' terms to be important for effective classroom practice (such as engagement, feedback, guided learning, high levels of instruction, questioning, modelling, praise, review, scaffolding, good subject knowledge, giving time for practice etc. (Coe et al., 2014; Hattie, 2009; Ko, Sammons & Bakkum, 2014; Muijs & Reynold, 2011; Rosenshine, 2012). It is not inconceivable that testing might, in some circumstances, work in opposition to otherwise effective classroom strategies – indeed in this project this may have been the case within some teacher trials. In the real world (as opposed to the laboratory condition), it is highly likely that the application of a test, as useful as this might be in theory, will have to be contextually defined and applied through the lens of best existing education practice, the context, age of the children, the subject being taught and the point that has been reached in the learning process. It cannot be good enough to imply that testing will always work, for every teacher, in every situation, with all children – nor can it be acceptable to jump to similar conclusions about other evidence from the science of learning. Indeed, from a developmental (and educational) perspective, it is not at all surprising that the same thing does not work in the exact same way across all ages and content. In fact, it would be rather surprising if it did. Those promoting educational neuroscience classroom solutions to teachers should, in our view, be more cautious in the way in which they present such approaches until clearer evidence has been accumulated.

To support the assessment of the project as a proof of concept, teacher trial posters and protocols were evaluated using the Jadad score (Jadad et al., 1996) (Table 1). This method is used to evaluate the methodological quality of clinical controlled research. It does this in relation to three key methodological features (randomisation, masking (blinding), withdrawals) that can be scored from zero (very poor) to five (rigorous), overall. The 15 teacher-led trial designs (and 16 replications) scored a combined average score of 2.73 (moderately rigorous). Main areas for improvement were: a) inclusions of the reporting of withdrawals and attrition rates when these were zero; b) design of trials in a way that might allow for double blinding.

In the case of the present studies double blinding was not possible, as the aim was to develop teacher skills regarding trial design and delivery. Therefore, the teachers were actively involved in the design, implementation and analysis of their own research. However, such approaches could be feasible going forward, now that the group of teachers involved have developed the skills to lead others in the delivery of such translational designs. Indeed, such a cohort of teachers, could not only manage the trial protocol on the ground to ensure masking but could also analyse each other's results as independent and blind evaluators.

CONCLUSIONS, NEXT STEPS AND IMPLICATIONS FOR POLICY MAKERS

Recently, Connolly and colleagues (2018) concluded the first systematic literature review of the 1,071 large-scale RCTs that have taken place in education over the last two to three decades. As well as showing many of the early criticisms of the use of RCTs in education to be largely unfounded, they identify several remaining issues. Relevant to the findings above, they point to the need for 'more nuanced and sophisticated trials that explicitly seek to contribute to theory testing and development and that are acutely aware of the contingent and context-specific nature of educational interventions' (2018, p. 14). Similarly to Fitzgibbon (1985), who argued that meta-analysis of small-scale studies, could be a viable and desirable option in education, Connolly and colleagues conclude that 'through the synthesis of data from a range of RCTs conducted across a variety of contexts, there is the genuine possibility of beginning to move on from the notion of 'what works' towards what works for whom, under what conditions, and in what circumstances' (Connolly, Keenan & Urbanka, 2018, p. 15).

The evidence presented above suggests that collaborative multiple planned replication by teachers could not only resolve many of the challenges that remain in RCT programs in education (Connolly, Keenan & Urbanska, 2018) but could, in addition, help to ensure the buy-in of the teaching profession and create evidence pipelines paralleling 'bench to bedside' processes in medicine and healthcare (Horton, 1999). Returning to the results from this project, we believe teacher-led RCTs have greater potential to control extraneous variables (variations in implementation) compared to larger scale trials – which face even greater challenges, in this regard, compared to laboratory studies. They also have the potential for higher levels of mundane realism ('everyday-ness' - reducing the way that participants may react to being in the trial). In addition, teacher-led RCTs offer up the possibility of breaking complex multifaceted pedagogical interventions down into their component parts; and, by extension, it becomes more feasible to study the effects of interventions on different children in different contexts and with subtle and systematic variations in protocol.

Future similar deliveries may wish to make use of meta-regression to moderate contextual factors with a larger number of replications, perhaps involving 15 or more studies coded for the same moderating variables (see Valentine (2010), for a discussion of power-related issues that can arise during the meta-analysis of small-scale studies). This should be specified in an analysis plan for the protocol and seen as offering correlational indications across the small-scale randomised trials. Meta-regression is not usually undertaken with less than 10 studies coded for each of the moderators investigated (Higgins & Green, 2011). Provisional exploration might be undertaken with subgroup analysis for categorial variables, but both approaches risk being underpowered (Baker et al., 2009), especially if the underlying studies are themselves underpowered (see also Tipton, 2015).

Regarding the future assessment of the quality of teacher-led RCTs an adapted version of the CONSORT checklist could be used to support the process. This worked well in medicine (Moher et al., 2001) and some preliminary development work has already been undertaken in an ESRC funded project (Grant et al., 2013). This could involve reports being published on a moderated website where those submitting are required to review other submissions. An additional possibility would be to work with existing educational organisations such as the College of Teaching or the Education Endowment Foundation who already support similar work. Although beyond the intended scope of this project and aim of the analysis, further similar collaborations between neuroscientists and educators should consider focusing on specific areas for investigation so that full conceptual background can be established for the key research question(s), research question(s) and hypotheses.

As in the type of research conducted by serving clinical practitioners in medicine and healthcare, serving teachers are perhaps better placed to come up with interventions and variants on current practice than educationalists who no longer practice the art of teaching on a daily basis, or commercial organisations driven by the goal of finding a standardised cost-effective product, rather than the sort of context-specific outcomes that are demanded in evidence-based practice. Elsewhere (Churches, 2016; Churches and Dommett, 2017), we have noted that it was teachers themselves who started producing trial results after attending training – the aim of which was simply to develop their understanding of the large-scale trials that they had been recruited to support. Looking back, we should perhaps not have been surprised at the enthusiasm of the teachers who have been involved. Teacher are constantly 'experimenting' informally, allocating children to different tasks, testing, and reporting what they see to improve pupil outcomes. Experimental research design gives them a structure to hang these activities on. Beyond that, the findings from this current project suggest that with systematic organisation teacher-led experimental research could become a national and global means for teachers to create, communicate, and share findings.

In terms of next steps for the approach outlined in this paper several possible routes are becoming clear, none of which are mutually exclusive. Firstly, there is much potential for the growing number of educational neuroscience PhD students to connect with teachers experienced in conducting teacher-led RCTs, to build collaborations around the translation of laboratory evidence as it grows and develops. In turn, we are looking to build an online space where teachers can share their conference posters and findings. There is also the possibility of teachers beginning to collaborate at scale in a similar way to that shown to work in the emerging field of citizen science (Gura, 2013; Hand, 2010). For example, teachers could be given access to well-designed protocols that they then deliver in their local context contributing to large-scale data sets that could be analysed in a similar way to the teacher trials above. Finally, multiple planned teacher-led RCTs exploring a single intervention (with the synthesis of findings in a meta-analysis as the outcome), could have much potential in adaptive programming environments – where testing, learning and iteration are required to find solutions (Ramalingam, Wild and Buffardi, 2019). Interventions that are being explored in an adaptive way could have each adaption bounded and segmented by pre- and post-testing so that alongside context differences different adaptations could be compared within different sequential treatment windows.

Scaling will inevitably have its challenges. Effective change of this nature needs national buy-in, collaboration, and investment. Turning the world of education research on its head (so that teachers become the producers of the research rather than just the passive consumers of it), requires changes in behaviour at all levels of the system and within a wide range of stakeholder groups (including academia, government, and school leadership teams). At the same time, maintaining quality at scale will require robust systems of accountability (including links to accreditation and peer review) that ensure scrutiny of research methods and results to the same standard expected in academic research.

REFERENCES

Adesope, O.O., Trevisan, D.A., & Sundararajan, N. (2017). Rethinking the use of testing: a meta-analysis of practice testing. *Review of Educational Research*, *87*(3), DOI: 10.3102/003465431668930

Baker, S., & Hindley, C. (2018a, March). Combining retrieval practices with Look, Cover, Write, Check improves pupils' progress in spelling. Poster session presented at Third Space: Science of Learning Conference, Chartered College of Teaching, University of York, York.

Baker, S., & Hindley, C. (2018b, June). Retrieval practices (multiple choice testing) and recall of multiplication tables – a Year 4 trial and parallel Year 5 replication. Poster session at European Conference on Education, IAFOR, Brighton.

Baker, W.L., White, C.M., Cappelleri, J.C., Kluger, J., & Coleman, C.I. (2009). Understanding heterogeneity in meta-analysis: the role of meta-regression. *International journal of clinical practice*, *63*(10), 1426-1434.

Biesta, G. (2007). Why 'what works' won't work: evidence-based practice and the democratic deficit in educational research. *Educational Theory*, *57*(1), 1–22.

Bjork, E.L., & Bjork, R.A. (2011). Making things hard on yourself, but in a good way. In M.A. Gernsbacher, R.W. Pew, & L.M. Hough (Eds.), *Psychology and the Real World* (pp. 56–64). New York: Worth Publishers.

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

Brookman-Byrne, A., & Thomas, M.S.C. (2018). Neuroscience, psychology and education: Emerging links. *Impact* [College of Teaching]*, 2,* 5–8

Bruer, J.T. (1997). Education and the brain: A bridge too far. *Educational Researcher*, *26*(8), 4-16.

Bryant-Khachy, F. (2018a, June). A randomised controlled trial of spaced learning in a primary school context. Study A: KS1 parallel replication (geography) [year group replications reported: Y1, Y2]. Poster session at Academy of Principals, 9th Global Educational Leadership Conference, Singapore.

Bryant-Khachy, F. (2018b, June). A randomised controlled trial of spaced learning in a primary school context. Study B: KS2 parallel replication (history) [year group replications reported: Y3, Y4, Y5, Y6]. Poster session at Academy of Principals, 9th Global Educational Leadership Conference, Singapore.

Churches, R. (2016). Closing the gap: test and learn: research report. London: Department for Education/National College for Teaching and Leadership [paper]. Retrieved from www.gov.uk/government/publications/closing-the-gap-test-and-learn

Churches, R. (2016, Winter). Closing the Gap: Test and Learn. London: Department for Education/National College for Teaching and Leadership [paper]. Retrieved from www.gov.uk/government/publications/closing-the-gap-test-and-learn

Churches, R., & Dommett, E. (2016). *Teacher-Led Research*. Camarthen: Crown House. www.educationdevelopmenttrust.com/~/media/EDT/Reports

Churches, R., Dommett, E., & Devonshire, I. (2017). *Neuroscience for Teachers*. Camarthen: Crown House.

Churches, R., Higgins, S., & Hall, R. (2017). The potential of teacher-led randomised controlled trials in education research. In A. Childs, & I. Menter (Eds), *Mobilising Teachers Researchers* (pp. 113–140). Abingdon: Routledge.

Churches, R., & McAleavy, T. (2016). Evidence That Counts. Reading: Education Development Trust [paper]. Retrieved from

Coe, R., Aloisi, C., Higgins, S., & Elliot Major, L. (2014, October). What makes great teaching? London: Sutton Trust [paper]. Retrieved from www.suttontrust.com/wp-content/uploads

Coldwell, M., Greany, T., Higgins, S., Brown, C., Maxwell, B., Stiell, B., Stoll, L., Willis, B., & Burns, H. (2017, July). Evidence-informed teaching: an evaluation of progress in England: Research report [paper]. London: Department for Education. Retrieved from www. assets.publishing.service.gov.uk/government/uploads

Connolly, P., Keenan, C., & Urbanska, K. (2018). The trials of evidence-based practice in education: a systematic review of randomised controlled trials in education research 1980–2016. Educational Research, DOI: 10.1080/00131881.2018.1493353.

Day, L. (2018, March). The effect of 'think-alouds' and other metacognitive strategies on pupil performance in maths problem solving. Poster session presented at Third Space: Science of Learning Conference, Chartered College of Teaching, University of York, York.

DfE (2010). The importance of teaching: The schools white paper 2010. London: Department for Education [paper]. Retrieved from www.assets.publishing.service.gov.uk/government/uploads

Dommett, E., Devonshire, I.M., & Churches, R. (2018). Bridging the gap between evidence and classroom 'clinical practice', *Impact* [College of Teaching], *2*, 64–67.

Dommett, E.J., & Devonshire, I.M. (2010). Neuroscience: viable applications in education. *The Neuroscientist*, *16*(4): 349–356.

Dommett, E.J., Devonshire, I.M., Sewter, E. & Greenfield, S. (2013). The impact of participation in a neuroscience course on motivational measures and academic performance. *Trends in Neuroscience and Education*, *2*(3–4), 95–138.

Dunford, E., & Rhoades, T. (2018, September). Using an app that provides 10 minutes of retrieval practice every day improves the speed of times table recall in a rural primary school in England – a preliminary study with parallel Year 2 and Year 3 replications. Poster session at BERA Conference, Northumbria University, Newcastle.

Elliott, R., & Wyatt, R. (2018, March). Using retrieval practice improves pupil progress in times table tests with Year 4 pupils – a randomised controlled trial. Poster session presented at EARLI SIG22, Neuroscience and Education Conference, Wellcome Trust, London.

Fitz-Gibbon, C.T. (1985). The implications of meta-analysis for education research. *British Educational Research Journal*, *11*(1), 45-49.

Fuller, J., & Glendening, J. (1985). The neuroeducator: professional of the future. *Theory into Practice*, *24*, 135–137.

Gaddes, W. (1968). A neuropsychological approach to learning disorders. *Journal of Learning Disabilities*, *1*(9), 523–534.

Grant, S., Mayo-Wilson, E., Melendez-Torres, G.J., & Montgomery, P. (2013). The reporting quality of complex psychological and social intervention trials: a systematic review of reporting guidelines and trial publications, *PLoS ONE*, *8*(5), e65442.

Greenfield, J., Noden, J., & Siddle, J. (2018, September). The effect of interleaving and retrieval practice on times table retention a randomised controlled trial. Poster session at BERA Conference, Northumbria University, Newcastle.

Gu, Q., Rea, S., Sammons, P., & Hodgen, J. (2018). Evaluation of the Research Schools Program in new opportunity areas: investigating the impact of Research Schools in promoting better outcomes in schools. Protocol Version: 1.0 [paper]. London: Education Endowment Foundation. Retrieved from: www.educationendowmentfoundation.org.uk

Gura, T. (2013). Citizen science: Amateur experts. *Nature, 496*(7444): 259–261.

Hand, E. (2010). Citizen science: People power. *Nature, 466*(7307): 685–687.

Hak, T., Van Rhee, H.J., & Suurmond, R. (2016). How to interpret results of meta-analysis. (Version 1.3) Rotterdam, The Netherlands: Erasmus Rotterdam Institute of Management [paper]. Retrieved from www.erim.eur.nl/research-support/meta-essentials/downloads

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

Higgins, J.P.T., & Green, S. (2011) *Cochrane Handbook for Systematic Reviews of Interventions* [online book resource]. Retrieved from https://handbook-5-1.cochrane.org

Higgins, S. (2018). *Improving Learning: Meta-analysis of intervention research in education*. Cambridge: Cambridge University Press. DOI: 10.1017.9781139519618

Horton, B. (1999). From bench to bedside…research makes the translational transition. *Nature*, *402*,213–215.

Howard-Jones, P.A. (2010). *Introducing Neuroeducational Research*. Oxford: Routledge.

Howard-Jones, P.A. (2014, January). Neuroscience and education: A review of educational interventions and approaches informed by neuroscience. London: Education Endowment Foundation [paper]. Retrieved from www.educationendowmentfoundation.org.uk/public/files/Publications

Howard-Jones, P.A. (2017) Neuromyths [paper]. Geneva: International Bureau of Education-International Brain Research Organisation/ United Nations Educational, Scientific, and Cultural Organization. www.ibe.unesco.org/sites/default/files/resources

Jadad, A.R., Moore, R.A., Carroll, R., Jenkinson, C., Reynolds, D.M., Gavaghan, D.J., Henry, J. & McQuay, D.M. (1996). Assessing the quality of reports of randomised clinical trials: is blinding necessary? *Controlled Clinical Trials*, *17*, 1–12.

Kim, M., & Sankey D. (2017). Philosophy, neuroscience and preservice teachers' beliefs in neuromyths: A call for remedial action, *Educational Philosophy and Theory*, DOI: 10.1080/00131857.2017.1395736

Ko, J., Sammons, P. & Bakkum, L. (2014). Effective teaching: a review of research and evidence [paper]. Reading: Education Development Trust. Retrieved from www.educationdevelopmenttrust.com/~/media/EDT/Reports

Maberly, L. (2018, March). A preliminary study into the effectiveness of using revision flashcards to improve performance in GCSE end-of-topic assessment. Poster session presented at Third Space: Science of Learning Conference, Chartered College of Teaching, University of York, York.

Makarova, D. (2018, October). A small-scale randomised controlled trial into the effect of practice testing on attainment in KS4 Science. Poster session, International Mind, Brain, and Education Society Conference, University of Southern California, Los Angeles.

McAleavy, T. (2016). Teaching as a research-engaged profession: Problems and possibilities. Reading: Education Development Trust [paper]. Retrieved from https://www.educationdevelopmenttrust.com/~/media/EDT/Reports/Research

Moher, D., Jones, A., Lepage, L., & Consort Group (2001). Use of the CONSORT statement and quality of reports of randomised trials: a comparative before-and-after evaluation. *Jama*, *285*(15), 1992–1995.

Morris, C. (2018, March). Retrieval practice, the use of novelty and normal classroom practice – a small-scale randomised controlled trial using a 2x2 factorial design, in an English rural primary school. Poster session at Academy of Principals, 9th Global Educational Leadership Conference, Singapore.

Muijs, R.D., & Reynolds, D. (2011). *Effective teaching: evidence and practice*, London: Sage.

NIHR (2016/2017). Annual Report – Improving the health and the wealth of the nation through research [paper]. National Institute for Health Research. Retrieved from www.nihr.ac.uk/about-us/documents

Pemberton, R. (2018, September). Preliminary evidence for the effectiveness of software based on retrieval practice, compared with peer-based learning and current ordinary school practice. A pilot within-participant design in a small rural primary school. Poster session at BERA Conference, Northumbria University, Newcastle.

Quinn, L., & Lamb, J. (2018, March). The effect of retrieval practice on retention and recall of vocabulary with Year 8 English students. Poster session presented at Third Space: Science of Learning Conference, Chartered College of Teaching, University of York, York.

Ramalingam, B. Wild, L. & Buffardi, A.L. (2019, April) Making adaptive rigour work [paper]. UKaid/USAID. Retrieved from https://www.odi.org

Ramsay, B., & Boothby, I. (2018, October). Structured collaborative problem-solving may improve performance more than unstructured collaboration and solitary learning –preliminary evidence from a randomised controlled trial. Poster session, International Mind, Brain, and Education Society Conference, University of Southern California, Los Angeles.

Riley, R.D, Higgins, J.P.T., & Deeks, J.J. (2011). Interpretation of random effects meta-analyses. British *Medical Journal*, 342, d549.

Rosenshine, B. (2012). Principles of instruction: Research based principles that all teachers should know, *American Educator*, Spring 2012. Retrieved from www.aft.org/sites/default/files/periodicals

Rosenthal, R. (1994). Parametric measures of effect size. In H. Cooper & L. V. Hedges (Eds.), *The Handbook of Research Synthesis* (pp. 231-244). New York: Russell Sage Foundation.

Siddle, J. (2018, October). The effect of two types of retrieval practice on vocabulary recall with different aged primary school pupils – a randomised controlled trial incorporating four parallel replications (EYFS, Year 2, Year 3 and Year 5). Poster session, International Mind, Brain, and Education Society Conference, University of Southern California, Los Angeles.

Simmonds, A. (2014). How neuroscience is affecting education [paper]. London: Wellcome Trust. Retrieved from www.wellcome.ac.uk/sites/default/files

Suurmond, R., van Rhee, H., & Hak T. (2017). Introduction, comparison and validation of Meta-Essentials: A free and simple tool for meta-analysis. *Research Synthesis Methods*, *8*(4), 537–553.

Tipton, E. (2015). Small sample adjustments for robust variance estimation with meta-regression. *Psychological Methods*, *20*, 375–393.

Valentine, J.C., Pigott, T.D., & Rothstein, H.R. (2010). How many studies do you need? A primer on statistical power for meta-analysis. *Journal of Educational and Behavioral Statistics*, *35*(2), 215-247.

FIGURE LEGENDS

Fig. 1. Levels of research in neuroscience and their applicability to classroom practice

Fig. 2. Forest plot of effect sizes and 95% confidence intervals from the teacher-led RCTs and year group replications

TABLE LEGENDS

Tab.1. Analysis of the robustness of the trials reported on in this paper using Jadad scores

FOOTNOTES

[1] Education Development Trust

[2] Institute of Psychiatry, Psychology, and Neuroscience (IoPPN), King's College London

[3] University of Nottingham Medical School

[4] British Science Association, Education Section

[5] School of Education, Durham University

[6] Education Development Trust

[7] This book contains discussions of a wide range of evidence from the science of learning. It includes a chapter connecting the cognitive psychology evidence (often cumulatively referred to as the 'desirable difficulties' research (Bjork & Bjork, 2011)) to the broader discussion about the use neuroscience in education.

[8] The first presentation of each poster is recorded in the references below.

[9] Rosenthal's *r* is a non-parametric equivalent of Cohen's *d*. The thresholds 0.2, 0.5, and 0.8 are considered to represent small, medium and large effects, respectively. The thresholds 0.1, 0.3, and 0.5 are the equivalent thresholds when interpreting *r*.

[10] We have included the conversion of *r* to *d*, in the interest of communication with the education profession. Much education research dissemination has made use of this effect size as a standard for discussing the effects of different interventions, irrespective of distribution (see for example, Hattie (2009)).

[11] In state schools in England, each year that children attend school is given a number. For example, the Early Years Foundation Stage and Year 1 to Year 2 cover the ages five to seven etc. (abbreviated in Figure 2 as EYFS, Y1 and Y2 etc.).

Table 1. Analysis of the robustness of the trials reported on in this paper using Jadad scores

	Scoring protocol	Baker & Hindley (2018a)	Baker, S. & Hindley (2018b)	Bryant-Khachy (2018a)	Bryant-Khachy (2018b)	Day, L. (2018)	Dunford & Rhoades (2018)	Elliott & Wyatt (2018)	Greenfield, Noden & Siddle (2018)	Maberly (2018)	Makarova (2018)	Morris (2018)	Pemberton (2018)	Quinn & Lamb (2018)	Ramsay & Boothby (2018)	Siddle (2018)	
Was the study described as randomised (this includes words such as randomly, random, and randomisation)?	0/1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Was the method used to generate the sequence of randomisation described and appropriate (table of random numbers, computer-generated, etc)?	0/1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	
Was the study described as double blind?	0/1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Was the method of double blinding described and appropriate (identical placebo, active placebo, dummy, etc)?	0/1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Was there a description of withdrawals and dropouts?	0/1	1	1	1	1	1	0	0	1	0	0	1	1	1	1	1	
Deductions																	
Deduct one point if the method used to generate the sequence of randomisation was described and it was inappropriate (patients were allocated alternately, or according to date of birth, hospital number, etc).	0/-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	
Deduct one point if the study was described as double blind but the method of blinding was inappropriate (e.g., comparison of tablet vs. injection with no double dummy).	0/-1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**Average trial robustness**
TOTAL		3	3	3	3	3	2	2	3	2	2	3	3	3	3	3	2.73

Figure 1: Levels of neuroscience research

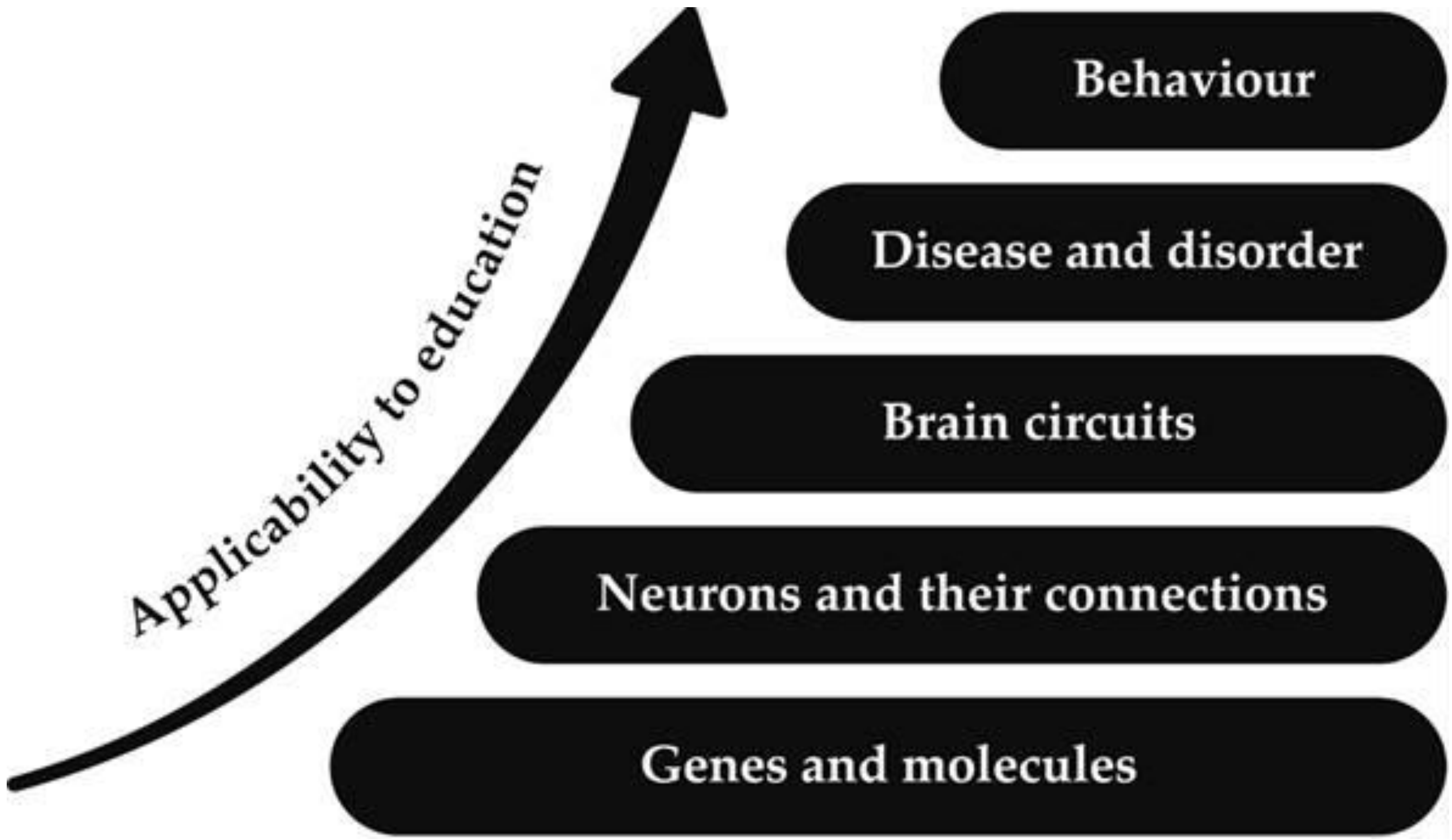

Figure 2: Forest plot for all 34 results