This is an Accepted Manuscript of a paper to appear in *Educational Research and Evaluation*

Unmasking the unasked: correcting the record about assessor masking as an explanation for effect size differences.

Adrian Simpson
School of Education, Durham University

Abstract

Ainsworth et al.'s paper "Sources of bias in outcome assessment in randomised controlled trials: a case study" examines alternative accounts for a large difference in effect size between two outcomes in the same intervention evaluation. It argues that the probable explanation relates to masking: only one outcome measure was administered by those aware of participants' treatment assignment. This paper shows this conclusion is not substantiated by the evidence: the original paper fails to exclude alternative explanations and what it takes as positive evidence for the preferred explanation is actually negative. While accepting the importance of masking in RCTs, this paper concludes that the original question was based on a misconception about effect sizes: seen correctly as a measure of whole study design, the question of effect size difference between different outcome measures does not need asking.

Introduction

There is no doubt that, for randomised controlled trials to act as a 'gold standard' in being able to attribute the cause of differences in outcomes to differences in intended treatment, study designers need to ensure that those intended treatments are the only differences between randomly allocated groups. Allowing any other factor to vary between the groups voids that causal attribution.

One set of potential factors relates to masking. Howick (2011) argues that as many as seven different groups may need to be masked[1] from the treatment assignment in an RCT: allocators, participants, treatment implementers, data collectors, assessors, analysts and authors. Failure to do this can lead to selection, performance, assessment and/or reporting biases. Field RCTs in education cannot obviously mask participants or treatment implementers to treatment assignment: pupils, parents, fellow pupils, teachers, teaching assistants and many others besides are likely to know whether individuals have been assigned to an intervention or a control treatment. However, the opportunity to mask others may be available.

Even in the pharmacological paradigm from which 'evidence based education' takes its inspiration, full masking is sometimes impossible. If a cancer treatment has a strong side effect (such as severe sickness or hair loss), it is difficult to conceive of an ethical placebo which could be developed to preserve masking for doctors and patients. Moreover, a rapidly effective cure is easily detected, leading Ney, Collings and Spencer (1986) to describe

[1] Following Howick, this paper uses the less loaded term 'masking' instead of the more common 'blinding'.

the ‘Philip’s Paradox’: the more powerful the treatment, the more difficult it is to evaluate according to standards requiring completely masked trials.

In education, it is likely to be impossible to mask many key groups of people involved in a study, but it could be argued that we should still do the best we can and mask wherever possible and it is important to understand the potential consequences which may result when the mask slips.

Ainsworth et al. (2015) details an interesting study in which two different outcome measures were used: one in which the administrators and markers were masked to treatment assignment and another where they were not. The two different assessments were associated with very different standardised effect sizes: the unmasked assessor condition had an effect size more than three times that of the masked assessor condition. The paper explores three alternative explanations for this: the timing of the assessment, the ‘treatment inherence’ of the different measures and the assessor masking condition. It present two forms of evidence: comparing subsets of the less treatment inherent measure and examining the heterogeneity of the results of the two measures across schools. The paper evaluates each piece of evidence as supporting or undermining each explanation and, on this basis, concludes that the probable explanation is assessor masking, justifying the advice to make assessor masking standard practice in educational field trials.

While recognising the potential importance of masking, this paper will show the argument is not valid. Evidence taken to support the assessor masking explanation is actually evidence against it and alternative explanations for treatment inherence and heterogeneity issues are not explored. Ultimately, however, the fact that different assessments lead to quite different effect sizes is not a matter that needs explanation unless one makes the category error of assuming that effect size is a measure of the effectiveness of an intervention. If, instead, one recognises that the intervention plays only a partial role in the definition of effect size, one should only expect effect sizes to be the same if the whole study is the same: the intervention and control treatments, the sample and, most crucially in this case, the exact nature of the test. Even tests focussed on the same topic areas can lead to radically different effect sizes, irrespective of the masking status of assessors.

Context

As part of an assessment of the ‘Numbers Count’ intervention within a wide ‘Every Child Counts’ numeracy programme (Torgerson et al., 2011) an RCT was designed involving over 500 pupils across over 40 schools. The programme involved an intensive one-to-one mathematics intervention intended for children in Key Stage 1 (ages 6-7) performing at the lowest 5% level nationally. The first phase of the evaluation took place between September and December 2009.

Part of the analysis involved two different tests. The primary test was the Progress in Mathematics 6 (PIM) test conducted in January 2010. The secondary test was the Sandwell Early Numeracy Test – Revised Test B (SENT-R-B) conducted in December 2009. PIM was used as the primary outcome measure because it was felt to cover a wide range of age appropriate mathematics, was easier to administer and had been carefully developed and

standardised. SENT-R-B was considered to be a weaker choice of measure because it was an integral part of the Numbers Count programme and appears narrowly focussed.

SENT-R-B covers basic numeracy including object counting, oral counting, value and computation, identification of numbers and language and is conducted one-to-one. PIM covers algebra, numbers and the number system, calculating, shape, space and measures, data handling and using and applying mathematics and is conducted in groups.

The PIM test was administered and marked by assessors who were masked from treatment assignment, while SENT-R-B was marked and administered by those who would have been aware of the treatment assignments.

The effect size (in the form of standardised mean difference) for the PIM test was 0.33 and for the SENT-R-B test was 1.11. Ainsworth et al. (2015) is predicated on the assumption that these numbers are estimates for the effectiveness of the intervention, should be expected to be similar and, therefore, that the apparently large difference needs explanation.

Alternative Explanations

Ainsworth et al. (2015) suggests three alternative explanations for the difference in effect size: the timing of the tests, their treatment inherence and the nature of masking. The paper presents various arguments as evidence in favour or against each alternative, concluding that "the evidence from this study suggests that the difference in effect size between the primary and secondary outcome is probably due to lack of blinding and non-independence of teachers administering the tests" (p.12), albeit maintaining that the only way to be sure of this explanation would be to conduct an RCT comparing masked and non-masked assessor conditions.

This paper examines each of the explanations and the evidence presented for and against them in Ainsworth et al. (2015).

In relation to whether the difference in effect size may have been due to the timing of the test, Ainsworth et al. (2015) says only "The timing of the test could be a possible explanation for the difference in effect size, with the possibility of any immediate benefit of the Numbers Count programme quickly diminishing over the Christmas holiday period. However, without also having results from a Progress in Maths 6 test conducted before the Christmas holidays, we cannot explore this." (pp 7-8). While in effect dismissing this explanation because the authors cannot address its likelihood any further, this argument cannot be taken as evidence either for or against the explanation.

In relation to treatment inherence, Ainsworth et al. (2015) notes that there is considerable evidence that effect size is impacted by the nature of the chosen measure. In particular, the overlap between focus of the teaching and the questions in the test appears crucial: while PIM is a general test of mathematics (including shape, space, measures and data handling questions in addition to numeracy questions), SENT-R-B contains questions focussed tightly on numeracy and was a part of the usual assessment regime of the Numbers Count programme. This then is an argument in favour of treatment inherence as an explanation for the effect size difference.

However, an argument to counter this is presented. The scores for two sub-sets of the PIM question set (one for number-only questions and one for other mathematical constructs) are calculated. By noting that, while slightly larger, the effect size for the PIM number variant (0.39) was still not as large as for the SENT-R-B test, the paper concludes treatment inherence does not account for all of the difference in effect sizes.

In relation to non-masking of assessors, Ainsworth et al. (2015) argues that some existing literature suggests that masking usually acts to reduce effect size. The paper presents as evidence for the non-masking explanation that there is a difference in heterogeneity of raw mean differences across the schools. By treating the results from each school as a mini-study of its own and applying meta-analytic techniques to combine these mini-studies, the paper provides measures of heterogeneity (Higgins & Thompson, 2002), noting that the PIM results (I^2 = 63%) have higher heterogeneity than SENT-R-B (I^2 = 48.3%). The paper argues "if most teachers consciously or unconsciously gave higher marks to the intervention group because of the knowledge that they were receiving the intervention, this might have decreased heterogeneity, as assessor bias may be more likely to act consistently. However, if only a few teachers were consciously or unconsciously giving higher marks to the intervention group, then heterogeneity would have increased." (p.9).

Tangentially the paper notes that one might expect reduced heterogeneity with SENT-R-B compared to PIM if the variation in skills tested was reduced as a result of the intervention. If, as might be expected, a numeracy intervention leads to more consistent teaching of numeracy than other areas of mathematics, when evaluated across schools, there would be more variation measured with PIM than with SENT-R-B.

Evaluating the evidence base.

The original paper thus presents two key pieces of evidence to distinguish between alternative explanations: the difference between the effect size for the full PIM test and the number subset, and the difference in heterogeneity between the PIM and SENT-R-B tests.

Ainsworth et al. (2015) appears to argue that if the issue was treatment inherence, we would expect the effect sizes for SENT-R-B and the PIM-number subset to be similar. While much existing literature does indeed show that tests more closely aligned to the material taught in the intervention are associated with larger effect sizes (e.g. Slavin & Madden, 2011; Cheung & Slavin, 2016), it is worth examining whether there are alternative explanations which account for this issue. In particular, there is a clear alternative explanation in the sensitivity of a test to the precise choice and nature of the questions and how those questions are organised within a test. In fact, effect size can be extremely sensitive to the precise nature of questions, even when focussed narrowly on the same issue.

Figure 1 illustrates one example of this. Participants were randomly assigned to one of two groups. One group was asked to explain the definition of a well-defined, familiar category and then to classify fifteen objects according to whether they belonged to the category or not. The other group was only asked to undertake the classification without the explanation

instruction (Alcock and Simpson, 2017). The resulting effect size between the explanation group and the non-explanation group, on the researcher-designed test of 15 objects, was 1.01.

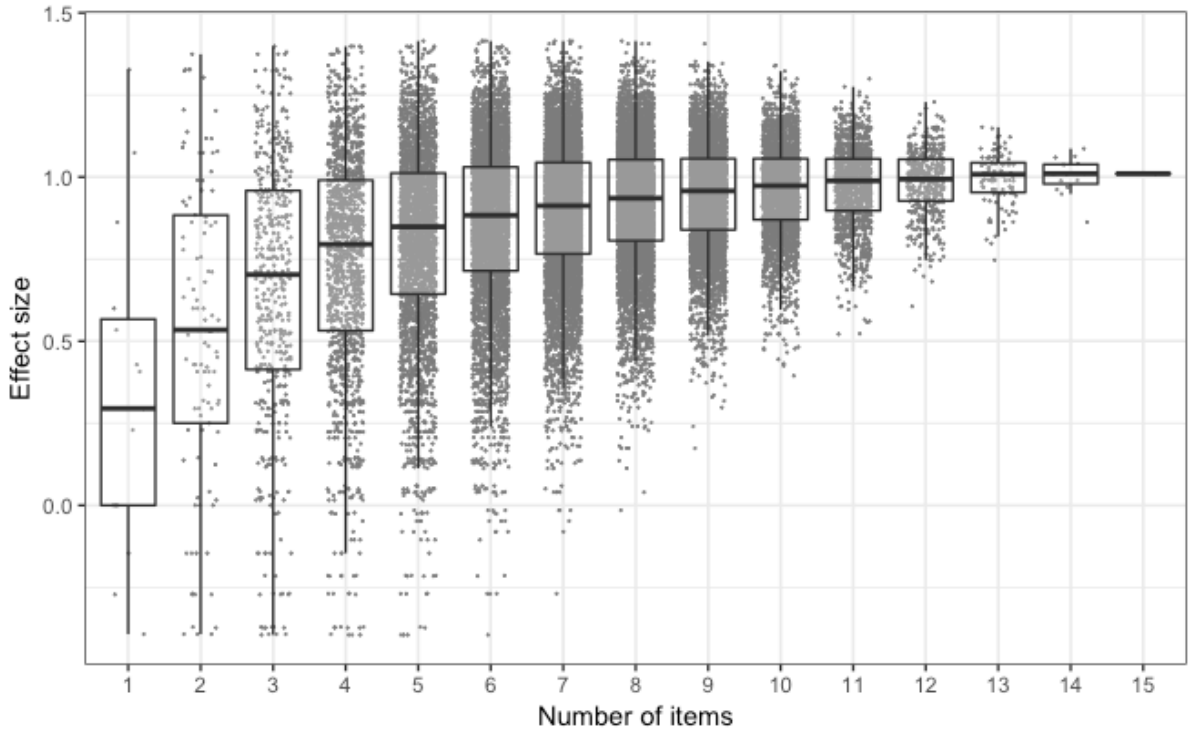

Figure 1: Impact of question choice on effect size

However, the choice of the fifteen objects was a design decision and many other design decisions would also have been acceptable. The researchers could have chosen a greater or lesser number of objects to classify. In figure 1, each dot represents one of the 32767 different alternative tests which could have been constructed from the fifteen items the researchers actually used, located according to the number of questions in the test and the effect size obtained by coding the real data on the questions in the test. The box plot for each test length gives a sense of the distribution of those effect sizes for each alternative test length.

Two things should be clear: first, averaged across all possibilities, subsets of tests would be expected to have lower effect sizes than full tests; second, the precise choice of questions can have a very large impact on effect size. For example, effect sizes associated with tests of ten questions vary from 0.39 to 1.34, with mean 0.96.

So, while superficially it makes sense to compare a subset of PIM focussed on number with SENT-R-B, on the one hand we would expect a random subset of PIM to have a smaller effect size, and on the other we would expect very large variation depending on the precise choice of questions. Given that the questions of SENT-R-B and PIM (and the PIM-numeracy subset) are disjoint, there is no reason to believe their effect sizes should be comparable. Moreover, it is worth noting that Ainsworth et al. (2015) acknowledges the different aims of the two tests, with PIM focussed on a wider spectrum of UK national curriculum levels

including material intended for older children, while SENT-R-B is intended directly for pupils covered by the Numbers Count curriculum.

The second piece of evidence Ainsworth et al. (2015) presents is the difference in heterogeneity of raw effect sizes across schools. This is used directly as evidence in favour of the non-masking explanation. The warrant for this is that non-masked teachers, acting consistently, would decrease heterogeneity. This is not the case. External assessors, trained in the use of the assessment instrument and masked to the treatment assignment would be expected to be more consistent between schools than teachers who might (consciously or unconsciously) deviate from the mark scheme to favour intervention pupils (unless the marks were adjusted so much that floor or ceiling effects came in to play). At best, teachers acting entirely consistently by adding a constant amount to each intervention group participant would inflate (raw) effect size while leaving heterogeneity across schools unchanged (and such a possibility would commit us to some conspiratorial mechanism by which teachers would act so consistently across schools).

So, decreased heterogeneity is not evidence in favour of non-masking as the cause of an increased effect size. Indeed, if anything, it is the opposite: all other things being equal, non-masked assessors who favoured the intervention group but who varied in their level of favouritism would lead to increased heterogeneity.

As with the first piece of evidence, it is instructive to consider what other factors may have led to Ainsworth et al.'s (2015) observation about the difference in I^2 values between tests. Heterogeneity is rather sensitive to missing data: just a single study omitted from a meta-analysis can substantially reduce I^2 (Patsopoulos, Evangelou, & Ioannidis, 2008). In the case of the data as illustrated in Figures 2 and 3 of Ainsworth et al. (2015), it is notable that data for SENT-R-B has two extra schools (G and GG) and one omitted school (OO) compared to the PIM data set. However, the loss of just one school can have a dramatic effect on heterogeneity: on the basis of the data as displayed, the PIM-6 I^2 for all 41 schools in Ainsworth et al.'s (2015) figure 2, is 60% (using the DerSimonian-Laird estimator for heterogeneity). Deleting just school W (which is unusual in having a negative raw mean difference and narrow confidence interval) reduces I^2 to 23%. This is a much larger reduction than between the PIM-6 and SENT-R-B analysis discussed in Ainsworth et al. (2015). That is, the level of missing data apparent in the Numbers Count data set could more than account for the discrepancy in heterogeneity.

Moreover, as noted above, Ainsworth et al. (2015) provides an alternative scenario in which I^2 might be reduced: if the teaching of numeracy was more consistent between schools than the teaching of other areas of mathematics, the heterogeneity of raw effects across schools would be smaller for a numeracy focussed test (SENT-R-B) than for a less focussed mathematics test (PIM).

So reduced heterogeneity is in fact an evidential mark *against* the non-masking explanation, in favour of the treatment inherence explanation and, by examining sensitivity, also offers support for an alternative explanation which is not addressed in Ainsworth et al. (2015) – different samples. While there is presumably considerable overlap between the groups of pupils who took the two tests, of the 522 participants who were assigned, 409 were

analysed for the PIM (and PIM-numeracy) effect size and 464 were analysed for the SENT-R-B effect size. One would expect the difference in the samples to contribute in some way to different effect sizes, even for identical tests. It is noticeable that 2% of the intervention group were missing from the SENT-R-B test (compared to 12% of the control group) while 17% of the intervention group were missing from the PIM test (compared to 21% of the control group); so somewhat more of the intervention-control difference was captured by SENT-R-B than by PIM.

Note neither of the presented items of evidence impacts on the timing explanation: the possibility that taking SENT-R-B immediately after the treatments, but delaying the PIM test until after a long Christmas holiday, might be a cause of effect size difference. Indeed, one would expect many interventions to have a notable drop off in effect size caused by a delay between treatment and measurement (Bailey, Duncan, Odgers & Yu, 2017).

Other Evidence of Masking Impact

Despite the evidential status of the items presented, Ainsworth et al. (2015) concludes that the masking condition is probably the cause of the effect size difference. It is instructive, then, to see if this issue is observable elsewhere.

The Numbers Count evaluation which led to Ainsworth et al. (2015) is just one of an increasing number of studies which purport to evaluate the effectiveness of education intervention projects. The Education Endowment Foundation (EEF) has an ongoing scheme to fund such projects and to date has published evaluation reports for over eighty interventions, generally giving an admirable level of detail. While the vast majority of these studies use administrators and markers who are blinded to the treatment allocation, this is not the case for all of them. Given the similarity of evaluation process between Numbers Count and these other projects, one might expect that, if masking is the issue Ainsworth et al. (2015) claims, we should see a consistent positive difference between masked and non-masked assessments.

However, this is not the case. In the each of the following examples (taken from the EEF evaluations), researchers used multiple outcome measures, some masked and some not, without evidence of more positive outcomes for non-masked measures:

- *Success for All* (Miller, Biggart, Sloan & O'Hare, 2017). The primary outcome was the Woodcock Reading Mastery Test III which was administered by assessors masked to treatment assignment while the secondary outcome was the national 'phonics check' which was administered by the teacher. The two effect sizes were very similar (0.07 and 0.06 respectively).
- *Lesson study* (Murphy, Weinhardt, Wyness & Rolfe, 2017). For one cohort, the primary measure was the combined Key Stage 2 maths and reading scores (based on national tests) while the secondary measure was the teacher assessed Key Stage 2 science level. Again, there was no evidence of a large effect size increase between masked and non-masked situations (0.02 and -0.06 respectively).
- *SPOKES* (Tracey, Chambers, Bywater & Elliott, 2016). Assessors masked to the treatment assignment conducted a battery of tests at different times as primary outcomes (with effect sizes 0.08, 0.05, 0.03, 0.15, 0.11, 0.11, 0.13, 0.27 and 0.25)

while a secondary outcome came from the Key Stage 1 assessment of literacy which is administered and scored by the non-masked teacher (effect sizes 0.02 [phonics], 0.26 [reading/writing] and 0.33 [reading]).

- *Improving Numeracy and Literacy* (Worth, Sizmur, Ager & Styles, 2015). The primary outcomes were administered and marked by those masked from treatment assignment (effect sizes -0.05, 0.20), while the secondary outcomes were administered and marked by those not masked to treatment assignment (effect sizes -0.06, -0.03).

So there is little evidence in these other projects, designed to similar standards and where assessments varied by masking condition, that being aware of treatment assignment leads to strongly positively shifted effect sizes. If, as Ainsworth et al. (2015) contends, non-masked assessors act consistently to exacerbate the difference between intervention and control outcomes in Numbers Count, the paper would need to explain why this is not apparent in these other relatively well designed RCTs.

Summary

So, the arguments that the difference in effect sizes is probably the result of assessors not being masked to treatment allocation are not warranted. Ainsworth et al. (2015) fails to address two alternative explanations: the timing of the tests – one carried out at the end of the intervention and one carried out after a Christmas break – and the difference in the sample. The paper gives as positive evidence for the masking explanation the decrease in heterogeneity when in fact this is evidence *against* the masking explanation. It fails to consider the sensitivity of the heterogeneity measure of missing school level data (where a single missing school can more than halve the proportion of heterogeneity). The paper states as evidence against the treatment inherence that restricting the broader measure to a subset of numeracy questions does not lead to a substantial effect size increase, but fails to address the issue of expected deflation of effect size with reduced test length and the extreme sensitivity of effect size to the precise selection of questions.

There is little doubt that masking is an issue that evaluators should be concerned about. Everything which happens to the intervention group after allocation and everything which happens to the control group after intervention should be considered a part of the 'blob of cause' – an interacting network of factors which may be causal, may support an effect or may suppress an effect. The more factors we can remove from that blob, the more chance we have of being able to conclude that the cause of the difference in outcomes is the difference in intended treatments (and not unintended differences).

Knowledge of treatment assignment is one part of the blob of cause. Whether it is teachers who are more motivated when they know they are assigned to the intervention; parents who take a closer interest when they know their offspring are getting special treatment; pupils who attend more when they view themselves as singled out for support or assessors who consciously or subconsciously mark intervention and control group scripts differently, all forms of masking are potential elements in the blob. However, we cannot reasonably argue that, on the basis of the evidence presented, the difference in effect size between PIM and SENT-R-B measures in this case is probably caused by differences in assessor masking conditions.

Conclusions

As with so much of the 'evidence based education' literature, the argument in Ainsworth et al. (2015) is grounded in the idea that effect size is a measure of the effectiveness of the intervention. From this viewpoint, large differences in effect size for the same intervention require explanation.

However, this interpretation of effect size is fundamentally flawed. Effect size is a measure associated with the study as a whole, not simply the intervention, and to treat effect size as a measure associated only with the intervention is a category error. Standardised mean difference depends on the whole set of design decisions: intervention treatment, control treatment, outcome measure and sample. While keeping any three of these the same, varying the other can lead to very different effect sizes (Simpson, 2017). As noted above, simply taking a different choice of questions from a bank of alternatives can lead to very different effect sizes.

So the fact that two different tests, with different designs and disjoint questions (conducted on slightly different samples, with an intervening holiday) have very different effect sizes should not be a surprise to anyone who avoids the effect size category error. In the end, Ainsworth et al. (2015) attempts to answer a question that never actually needed to be asked, addresses it with flawed logic and reaches a conclusion which should already have been appreciated.

References

Ainsworth, H., Hewitt, C. E., Higgins, S., Wiggins, A., Torgerson, D. J., & Torgerson, C. J. (2015). Sources of bias in outcome assessment in randomised controlled trials: a case study. *Educational Research and Evaluation*, *21*(1), 3–14.

Alcock, L., & Simpson, A. (2017). Interactions between defining, explaining and classifying: the case of increasing and decreasing sequences. *Educational Studies in Mathematics*, 94(1), 5–19.

Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. *Journal of Research on Educational Effectiveness*, 10(1), 7-39.

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

Higgins, J. P. T., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, 21(11), 1539–1558. http://doi.org/10.1002/sim.1186

Howick, J. (2011). *The Philosophy of Evidence-based medicine*. Chichester: Wiley.

Miller, S., Biggart, A., Sloan, S. & O'Hare, L. (2017). *Success for All: Evaluation report and executive summary*. London: EEF.

Murphy, R., Weinhardt, F., Wyness, G., & Rolfe, H. (2017). *Lesson Study: Evaluation report and executive summary*. London: EEF.

Ney, P. G., Collins, C., & Spensor, C. (1986). Double blind: Double talk or are there ways to do better research. *Medical Hypotheses*, 21(2), 119–126.

Patsopoulos, N. A., Evangelou, E., & Ioannidis, J. P. A. (2008). Sensitivity of between-study heterogeneity in meta-analysis: Proposed metrics and empirical evaluation. *International Journal of Epidemiology,* 37(5), 1148–1157.

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

Slavin, R. E., & Madden, N. A. (2011). Measures inherent to treatments in program effectiveness reviews. *Journal of Research on Educational Effectiveness*, 4, 370–380.

Torgerson, C. J., Wiggins, A., Torgerson, D., Ainsworth, H., Barmby, P., Hewitt, C., Jones, K., Hendry, V., Askew, M., Bland, M., Coe, R., Higgins, S., Hodgen, J., Hulme, C., & Tymms, P. (2011). *Every Child Counts: The independent evaluation executive summary*. London: Department for Education.

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

Worth, J., Sizmur, J., Ager, R. & Styles, B. (2015). *Improving Numeracy and Literacy: Evaluation report and executive summary*. London: EEF.