Sources of bias in outcome assessment in randomised controlled trials: A case study

to appear in

Educational Research and Evaluation: An International Journal on Theory and Practice

Hannah Ainsworth[a], Catherine E Hewitt[a], Steve Higgins[b], A. Wiggins[c], David J Torgerson[a] and Carole J Torgerson[b]*.

(The * following Carole Torgerson indicates that she is corresponding author).

[a]Ground Floor, ARRC Building
Department of Health Sciences
University of York
Heslington
York
YO10 5DD

[b]Durham University
School of Education
Leazes Road
Durham
DH1 1TA

[c]CEM
Rowan House, Mountjoy Centre
Stockton Road, Durham
United Kingdom
DH1 3UZ

Abstract

Randomised controlled trials (RCTs) can be at risk of bias. Using data from a RCT we considered the impact of post-randomisation bias. We compared the trial primary outcome, which was administered blindly, with the secondary outcome which was not administered blindly.

522 children from 44 schools were randomised to receive a one-to-one maths tuition programme that was assessed using two outcome measures. The primary outcome measure was assessed blindly whilst the secondary outcome was delivered by the classroom teacher and therefore this was un-blinded.

The effect sizes for primary and secondary outcomes were substantially different (0.33 and 1.11 respectively). Test questions that were similar between the two tests this did not explain the difference. There was greater heterogeneity between schools for the primary outcome, compared with the secondary outcome. We conclude that, in this trial, the difference between the primary and secondary outcomes was likely to have been due to lack of blinding of testers.

Running head: Sources of bias in outcome assessment in educational RCTs

Key words: randomised trials: methodology; blinding; treatment inherent measures.

Background

The randomised controlled trial (RCT) is widely regarded as the 'gold standard' research method in health for determining whether a cause and effect relationship exists between a proposed intervention and identified outcome (Cook & Campbell, 1979; Shadish, et al., 2002; Torgerson & Torgerson, 2008). When randomised controlled trials are possible they are usually the gold standard measure to establish effectiveness as they are the only design, when undertaken rigorously, that can offer the potential to eliminate selection bias. Other designs, no matter how well conducted, are always susceptible to selection effects. Some argue that RCTs are not the gold standard (Berk, 2005; Cartwright & Hardie, 2012). Cartwright and Hardie, in particular, argue that RCTs should not be seen as the key for evidence based policy and that because many are not generalizable other forms of evidence should be considered. In contrast, Berk (2005), whilst arguing that the RCT is not the 'gold-standard' accepts that nothing is, and the RCT is the best form of evidence there is. In this paper, we do not engage in this debate except to note it in passing. Rather we highlight an issue that if not taken into account in the design of a RCT will reduce its internal validity. A trial with poor internal validity will, by definition, have poor external validity or generalizability, as we cannot rely on the results in any context.

In educational research RCTs are increasingly being viewed as the design of choice for answering questions of effectiveness (Cooper, Levin & Campbell, 2009). The process of randomisation deals with a number of sources of bias which can

undermine the validity of an experiment, leading to incorrect conclusions being drawn. Selection bias is one of the main threats to the internal validity of an experiment. Selection bias occurs when participants are selected to receive the intervention on the basis of a variable associated with outcome (Shadish et al., 2002). Randomisation eliminates selection bias; however, there are a number of other sources of bias which can occur after randomisation, such as attrition bias (caused by the loss of participants post-randomisation) and dilution bias (occurring when participants in the intervention or control group get the opposite treatment, a form of contamination) (Torgerson & Torgerson, 2003; Torgerson & Torgerson, 2008). This paper focuses specifically on bias associated with outcome assessment.

Outcome assessment

In any RCT it is important that outcomes are assessed objectively and represent a 'fair test' of the intervention under evaluation. In educational trials the outcome of interest is usually a form of educational test. Often several educational tests are given to assess outcome. It is important, however, that a single test is identified as the main outcome variable *before* the experiment has been completed or the data examined. This is to reduce the problem of a Type I error, concluding that a difference between the groups exists, when in reality it does not. For instance, if we assume there is no difference between two groups and we test multiple outcomes we will eventually observe a difference that is statistically significant simply by chance (Bland & Altman, 1995). Consequently, we need to state in advance our main outcome and not have that choice driven by the data (i.e. data dredging).

After the main outcome test has been chosen it is usually appropriate to examine other 'secondary' outcome measures. These will usually correlate with the main outcome. Therefore, if we observe a difference between the groups in the main outcome we will usually see a difference in secondary outcomes in the same direction.

In a recent trial of a numeracy programme – *Numbers Count* within the *Every Child Counts* United Kingdom (UK) national mathematics policy – we found that whilst both our primary and secondary outcomes of numeracy found a difference favouring the intervention, the effect size was approximately four times greater for the secondary outcome than for the primary outcome.

In this paper we have undertaken exploratory analyses to ascertain some of the possible reasons for this difference in order to inform future RCTs about the selection and conduct of their outcome assessments.

Background to Every Child Counts evaluation

Details of the study have been published elsewhere. For further detailed description of the trial design and analysis see Torgerson et al., 2011a; Torgerson et al., 2011b; Torgerson et al., 2013. However, in brief, in 2009 an independently conducted pragmatic RCT investigated the effectiveness of *Numbers Count* (*NC*) compared to normal classroom practice. *Numbers Count* (Edge Hill University et al., 2008) is an intensive one-to one maths intervention within the *Every Child Counts* strategy for children performing in the lowest 5% nationally in maths at Key Stage 1 (age 6-7). The trial involved 44 schools; each of which identified approximately 12 children

meeting the inclusion criteria (n = 522 in total – see Figure 1). In each school, participating children undertook a pre-test, the Sandwell Early Numeracy Test – Revised (SENT– R) test A (Arnold et al., 2011), after which they were randomly allocated into three groups: Group 1 received *Numbers Count* in the autumn term (term 1), Group 2 received *Numbers Count* in the spring term (term 2); and Group 3 received *Numbers Count* in the summer term (term 3). All of the children were post-tested using the primary outcome, Progress in Maths 6 test (PIM 6) (Clause-May et al., 2004) at the beginning of the spring term (January 2010). All children were also post-tested using the secondary outcome, Sandwell Early Numeracy Test–R test B, at the end of the first term (December 2009). [See below for detailed discussion of the reasons each test was selected.]

INSERT Figure 1: Trial Design Diagram

The Progress in Maths 6 test (Clause-May et al., 2004) (administered in January 2010) was selected by the Trial Steering Committee as the primary outcome measure for the main randomised comparison between intervention and control children for a number of reasons (see below). The Trial Steering Committee made the pragmatic decision that the evaluation would also include, as a secondary outcome, the Sandwell Test (see below).

The Progress in Maths 6 test (Clause-May et al., 2004) was developed (and re-standardised) from the NFER/Nelson 5-14 Mathematics assessment and is a widely used commercial mathematics test. The Progress in Maths 6 version is appropriate for children who are six years of age. The assessment covers a wide range of mathematical skills and mirrors the National Curriculum assessments at key stage 1

(KS1) and key stage 2 (KS2) (as well as the international assessment (TIMSS) for 9-10 year olds). The key areas assessed are: algebra; numbers and the number system (the focus of *Numbers Count*); calculating; using and applying mathematics; shape, space and measures; handling data. Progress in Maths 6 can be administered to more than one child at once.

This test is programme-independent; in other words, it is not closely aligned to the *Numbers Count* programme, being neither used diagnostically nor as a teaching element of the programme. Skills covered by the Progress in Maths 6 are routinely taught during normal classroom practice.

The Sandwell Early Numeracy Test was originally developed for exclusive use by the Sandwell Inclusion Support Service, but it went on to be adopted by the Every Child a Chance Trust for use within the *Numbers Count* element of *Every Child Counts* both as a diagnostic feature and as a post-test following completion of the programme. The test is commercially available, but its use outside *Every Child Counts* (and Sandwell) is relatively limited. Two similar versions (A and B) are offered. The assessment covers National Curriculum levels P6 to 2a, and focuses on the following areas of number: identification of numbers; oral counting; object counting; value and computation; language. The use of P Scales and National Curriculum levels 1 and 2 covers a spread of attainment suitable for average pupils between the ages of 5-7. Performance scales (P scales 1-8) support assessment of children who are working below level 1 of the English national curriculum. Typically children who are working at level 1 are six years old and those at level 2 are seven years old. The intervention was targeted at low performing 6 year olds so a lower

baseline than the English national curriculum level 1 was needed. The Sandwell test largely mirrors the underlying approach of *Numbers Count* and is treatment inherent. The focus on number is based on the principle that gains in number will lead to gains in other areas of mathematics (e.g., space and shape) (Edge Hill University et al., 2008, p11). The test is administered by the NC teacher, or other teachers/trained members of staff, prior to the child starting the programme. The test is also administered on exit, and three and six months after the end of the programme by a link teacher.

The policy decision to use Progress in Maths 6 as the primary outcome measure was based on the following reasons: Progress in Maths 6 is a well-recognised and reliable standardised test; it is not part of the *Numbers Count* programme and it could, therefore, be administered independently of the programme; the evaluators could ensure that the people administering and marking the test were blinded to the groups (*NC* or control); it was a programme independent measure; it is a broad measure of mathematics achievement; and it could be administered to more than one child at once (i.e., it was cost effective in terms of the budget for independent testing).

The Sandwell assessment was selected as the secondary outcome for the following reasons: the testing could not be undertaken independently (due to it being part of the *NC* programme at both pre- and post-test – and the funder did not agree to fund independent administration of the Sandwell test at post-test) so the administration and marking of tests was not undertaken blind to the group allocations. The test itself was a programme inherent measure, and assessed a narrower range of

mathematics skills. The teaching was determined by weaknesses identified by the Sandwell test and therefore we would expect particularly good progress to be made in these areas, but the Trial Steering Committee wanted the evaluators to measure the broader mathematical impact of the programme because the *Numbers Count* programme works on the principle that equivalent gains in other areas of mathematics will be made. In conclusion, the Sandwell test was used to aid the diagnostic process in the programme, but it did not provide a good measure of programme impact.

As can be seen in Table 1, the mean Progress in Maths 6 mathematics test score for the children receiving *Numbers Count* in the autumn term was 15.8 (SD 4.9) and for the control children who had yet to receive *Numbers Count* it was 14.0 (SD 4.5). The effect size was 0.33 (95% CI 0.12 to 0.53) indicating strong evidence of a difference between the two groups (1.47 95% CI 0.71 to 2.23, $p<0.0005$). This shows that children who received *Numbers Count* scored significantly higher on the Progress in Maths 6 mathematics test compared with children in the control group who had not yet received *Number Count*.

The mean Sandwell B mathematics test scores for children receiving *Numbers Count* in the autumn term was 45.0 (SD 11.1) and for the control children who had yet to receive *Numbers Count* it was 32.5 (SD 10.2). The effect size for this measure was 1.11 (95% CI 0.91 to 1.31).

The effect size is approximately four times greater for the secondary outcome than for the primary outcome. There are a number of potential explanations for this

difference: the tests measure different mathematical constructs; the Sandwell Early Numeracy - R test focuses on number whilst the Progress in Maths 6 measures more general mathematical skills, including number. The tests also cover different national curriculum levels and it is possible that a floor effect is present in the Progress in Maths 6. The tests are also delivered in different ways and at different times; the Progress in Maths 6 is delivered to a group of children and was delivered in January 2010, the Sandwell Early Numeracy - R test is delivered individually and was conducted in December 2009. 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.

However, there are also potential sources of post-randomisation bias which may be systematically impacting the results. One potential source of post-randomisation bias is un-blinded outcome ascertainment. It is possible that knowledge of group allocation may have resulted in a conscious or unconscious tendency, by the testers, to award higher marks to children who had received *Numbers Count* when undertaking the Sandwell Early Numeracy - R test. A tendency to award higher marks to such children could be due to teachers believing the *Numbers Count* programme is more effective than normal classroom teaching. It could also be due to the fact that the tests were conducted by teachers who knew, and had in some cases been working individually with, the children; there may have been an element of giving children who had received *NC* the 'benefit of the doubt', because teachers

had previously seen the child demonstrate their ability to master a particular skill during the course of the intervention period.

The term 'blinding' refers to keeping trial participants, investigators, or assessors unaware of the assigned intervention, so that they will not be influenced by that knowledge (Shultz & Grimes, 2002). In education it is extremely difficult for participants, particularly teachers, to be blind to allocation. However blinding of assessors (those collecting outcome data) is possible and should be considered, as bias in test marking, particularly at post-test is a significant potential issue (Howlin, Gordon, Pasco, Wade & Charman, 2007; Torgerson, 2009). Blinding of outcome assessors usually reduces differential assessment of outcomes (ascertainment or information bias: Kelly & Perkins, 2012, p. 57).

A second potential source of post-randomisation bias is the use of a programme inherent measure, the Sandwell Early Numeracy - R test, in assessing outcomes. This test is used as a diagnostic tool within the *Numbers Count* programme, it follows that much of the teaching was determined by weaknesses identified by the Sandwell Early Numeracy - R test and therefore we would expect particularly good progress to be made in these areas.

A programme inherent measure or test inherent to the experimental intervention, would be one that assessed the knowledge and or skills taught as part of the experimental intervention but not ordinarily taught or taught to the control group. The test may be very closely related to the content of the experimental intervention, more

so than the content the control group will receive. The test in itself may also form part of the experimental intervention (Slavin & Madden, 2011).

The opposite of a treatment or programme inherent measure is a treatment or programme independent measure; such measures assess skills or content taught to both the control and experimental group (i.e. normal class teaching).

Cheung (2013) found that, for educational technology studies, measures inherent to the experimental treatment tended to report larger effect sizes (p 28). Slavin and Madden (2011), in their comparison of studies included in the What Works Clearinghouse (http://ies.ed.gov/ncee/wwc/), found that the average effect size for studies of mathematics interventions using programme (treatment) inherent measures was +0.45, whereas for programme independent measures the effect size was -0.03 (p. 377).

Methods

We were unable to explore all of the potential explanations for the difference in effect size found between the primary and secondary outcome. However we were able to conduct two additional analyses. Firstly, to explore whether the different mathematical constructs in the 2 tests (Progress in Maths 6 and Sandwell Early Numeracy test– R B) accounted for the difference in effect size between the primary and secondary outcome, we conducted 2 further regressions. Two sub-scores were calculated from the total Progress in Maths 6 score, one which included the number

only questions within the Progress in Maths 6 test and one which included the other mathematical constructs of shape, space and measures and data handling.

Secondly, to explore any potential impact of un-blinded outcome ascertainment we treated each school as a separate 'mini trial' and then combined the results in a meta-analysis for both the primary outcome measure (Progress in Maths 6) and the secondary outcome measure (Sandwell Early Numeracy Test- R B).

A priori, we made the following hypothesis:

If the difference in effect sizes between the Progress in Maths test and the Sandwell Early Numeracy Test were due to the latter being a more 'treatment inherent' measure then when the effect sizes of the 'treatment inherent' questions of the Progress in Maths tests were calculated separately from the non-treatment inherent questions there should have been a similar overall effect size.

We might expect differences in heterogeneity in the meta-analysis using the Progress in Maths 6 due to blinding compared with the Sandwell Early Numeracy Test- R B. 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.

There may also be less heterogeneity with the Sandwell Early Numeracy Test-R B test, compared with Progress in Maths 6, if the intervention reduces the variation of the teaching of the skills tested. Because the Progress in Maths 6 tests a broader range of mathematical skills, which the *Numbers Count* programme does not develop, there may be more variation in teaching in these non-*Numbers Count* 'core' skills.

Results

As can be seen from Table 1, the questions focused on number within the Progress in Maths 6 are clearly driving the effect size finding using this test. However this does not account for the difference in effect size between the primary and secondary outcome, with the effect size on number only questions within the PIM 6 being 0.39 (95% CI 0.16 to 0.61), still considerably different to the effect size of 1.11 (95% CI 0.91 to 1.31) found with the Sandwell Early Numeracy-R B test.

INSERT Table 1: Primary and Secondary Outcome Effect Sizes

Figure 2 presents the forest plot of a meta-analysis with each individual school treated as a 'mini trial' using the Progress in Maths 6 total score as the outcome measure. From figure 2 it can be seen that there is variation between the schools with some showing a programme benefit and some showing either no difference or a benefit of usual teaching over the intervention programme (as would be expected due to chance). Heterogeneity I^2 is 63% (Table 2).

Figure 3 presents the forest plot of a meta-analysis with each individual school treated as a 'mini trial' using the Sandwell Early Numeracy- R B test score as the outcome measure. From figure 3 it can be seen that there is still variation between the schools but all the schools are showing a programme benefit, apart from one. Heterogeneity I^2 is 48.3% (Table 2)

In figures 2 and 3 we can see that there are 8 'discordant' schools (A, C, FF, I, M, O, U, and W). These schools appear to show a positive effect of *Numbers Count* using the Sandwell Early Numeracy - R B test score as the outcome measure but a negative effect using the PIM 6 score as the outcome measure.

INSERT Table 2: Meta-analyses using individual schools as 'mini trials'

INSERT Figure 2: Meta-analysis PIM 6 total score

INSERT Figure 3: Meta-analysis

Discussion

The data presented in this paper highlight the difference found between the primary and secondary outcomes in a RCT investigating the effectiveness of an intensive one-to-one maths intervention and seeks to explore the possible underlying causes for such a difference. We have explored two potential explanations for the differences in effect sizes observed in the trial (test content and un-blinded outcome assessment).

Difference in test content does not appear to explain all of the difference in effect size between the primary outcome and the secondary outcome. A sub-score using number only questions within the Progress in Maths 6 test still results in an effect size of 0.39 (0.16 to 0.61) compared with an effect size of 1.11 (0.91 to 1.31) using the secondary outcome, which focuses entirely on number.

When the schools are treated as 'mini trials' and combined in a meta-analysis for each outcome, heterogeneity, which all things being equal, we would expect to be the same, is different between the primary and secondary outcome, (I^2 63.0 and 48.3 respectively). Heterogeneity is lower in the meta-analysis using the Sandwell Early Numeracy - R B test, this could be due to assessor bias acting consistently (therefore reducing heterogeneity), with the possibility that most teachers consciously or unconsciously gave higher marks to the intervention group because of the knowledge that they were receiving the intervention. Using the Sandwell Early Numeracy – R B test only 1 school shows a 'negative effect' compared with 9 schools when the PIM 6 test is used, suggesting a bias towards the intervention, as we would expect given the very small samples per school some negative results even if the intervention is effective. Qualitative work conducted as part of the independent evaluation demonstrates that all the *Numbers Count* teachers interviewed were positive about the programme, highlighting the impact on the children's mathematical and wider skills (reported in full in Torgerson et al., 2011b). Lower heterogeneity using the Sandwell Early Numeracy - R B test could also be due to the intervention reducing the variation of the teaching of the skills tested in this test compared with the Progress in Maths 6.

Therefore, 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. However other explanations are possible; indeed we know from previous studies (Cheung, 2013; Slavin & Madden, 2011) that programme (treatment) inherent measures are likely to inflate the effect size compared to programme independent measures. Although the additional analyses looking at test content (one indicator of whether a test is programme inherent or independent) suggest that, in this trial, variation in the programme independence of the primary and secondary outcomes may be limited as an explanation for the difference in effect size, a case can still be made that the Sandwell Early Numeracy– R B test remains more treatment inherent than the number only questions within the Progress in Maths 6 test (since Sandwell Early Numeracy test– R B was used diagnostically as part of the intervention). The only way to reliably determine that the difference in effect size can be explained solely by blinded outcome assessment (and therefore to rule out other potential explanations) would be to have a randomised comparison between children allocated to be tested blind and children allocated to be tested un-blind to allocation.

Interpreting effect sizes as a measure of programme effectiveness is always challenging (Hill, Bloom, Black & Lipsey, 2008); further potential biases can exacerbate these problems. With careful attention to design and conduct in this trial we were able to successfully minimise the possible impact of two post randomisation biases, associated with outcome assessment, on the conclusions of programme effectiveness. If conclusions were to rely solely on the results from the secondary outcome measure in this study, then without careful attention being paid to its

weakness; being conducted un-blind to group allocation and its programme inherent nature, an overestimation of the estimated effect size of the *NC* programme on children's mathematical skills could be made. However inclusion of a programme independent test, conducted blind to allocation, as the primary outcome measure avoided overestimation of programme effectiveness. The findings from this paper illustrate the vital importance of conducting blinded outcome assessment as a matter of standard practice in educational trials – without doing so greatly increases the chances of bias being introduced.

Acknowledgements

The trial described in this article was funded by the then-Department for Children Schools and Families (DCFS), now the Department for Education (DfE) in the United Kingdom. We also acknowledge the University of York and Durham University for additional funding to support the trial. The Every Child Counts independent evaluation team included: C.J.Torgerson, A. Wiggins, D.J.Torgerson, H.Ainsworth, P. Barmby, C. Hewitt, K. Jones, V. Hendry, M. Askew, M. Bland, R. Coe, S. Higgins, J. Hodgen, C. Hulme, and P. Tymms.

Notes on contributors

Hannah Ainsworth is a Research Fellow within the York she has worked on a number of randomised trials in education. She was the trial manager for the ECC trial and has subsequently managed other large trials in education.

Catherine Hewitt is a Senior Statistician and Deputy Director of York Trials Unit based in the Department of Health Sciences at the University of York. She is an experienced Statistician with a strong Mathematical and Statistical background. Her research has focused on developing, refining and applying statistical methods in the conduct of randomised controlled trials and systematic reviews. Currently, she is working across a number of health and education trials.

Steve Higgins is Professor of Education at Durham University. A former primary school teacher, he has an interest in the synthesis of intervention and evaluation findings through meta-analysis and the implications for policy and practice.

Carole Torgerson is Professor of Education in the School of Education at Durham University. She has completed a number of RCTs in education, and is currently principal investigator on nine RCTs funded by the Education Endowment Foundation.

David Torgerson is Director of the York Trials Unit and a Professor in the Department of Health Sciences. He is involved in a number of education and social science randomised trials.

Andy Wiggins is Associated Director of Research and Evaluation at the Centre for Evaluation and Monitoring (CEM) at Durham University. Over the last 12 years at CEM he has led and contributed to a wide range of educational research and trials for organisations such as the Educational Endowment Fund and the Scottish Government.

References

Arnold, C., Bowen, P., Tallents, M., Walden, B., & Sandwell inclusion support service. (2011). *Sandwell early numeracy test - revised (SENT-R).* GL Assessment. http://www.sandwellearlynumeracytest.co.uk/sent-r/

Berk, R.A. (2005) Randomized experiments as the bronze standard. *Journal of Experimental Criminology,* 1,:4-17-433.

Bland, J. M. & Altman, D. G. (1995). Multiple significance tests: the Bonferroni method. *British Medical Journal,* 310,170.

Cartwright, N., & Hardie J. (2012) *Evidence-Based Policy: A practical guide to doing it better.* OUP, Oxford.

Cheung, A. (2013). Effects of Educational Technology Applications on Student Achievement for Disadvantaged Students: What Forty Years of Research Tells Us. *Cypriot Journal of Educational Sciences,* 8(1), 19-33.

Clause-May, T., Vappula, H. & Ruddock, G. (2004). *Progress in Mathematics 6.* GL Assessment.

Cook, T.D., & Campbell, D. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Boston: Houghton Mifflin.

Cooper, A., Levin, B., & Campbell, C. (2009). The growing (but still limited) importance of evidence in education policy and practice. *Journal of Educational Change*, 10(2-3), 159-171.

Edge Hill University, Lancashire County Council & Every Child Counts. (2008). *Numbers Count Handbook 2008 – 2009.* Edge Hill University.

Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. *Child Development Perspectives,* 2(3), 172-177.

Howlin, P., Gordon, R. K., Pasco, G., Wade, A., & Charman, T. (2007). The effectiveness of Picture Exchange Communication System (PECS) training for teachers of children with autism: a pragmatic, group randomised controlled trial. *Journal of Child Psychology and Psychiatry,* 48(5), 473-481.

Kelly, B., & Perkins, D. F. (Eds.). (2012). *Handbook of implementation science for psychology in education.* Cambridge: Cambridge University Press.

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

Schultz, K. F. & Grimes, A. D. (2002.) Blinding in randomised trials: hiding who got what. *Lancet,* 359: 696–700

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

Torgerson, C. J. (2009). Randomised controlled trials in education research: a case study of an individually randomised pragmatic trial. *Education,* 3–13, 37(4), 313-321.

Torgerson, D. J., & Torgerson, C. (2008). *Designing randomised trials in health, education and the social sciences: an introduction*. Palgrave Macmillan, Basingstoke UK.

Torgerson, D.J., & Torgerson, C.J. (2003). *Avoiding bias in randomised controlled trials in educational research.* British Journal of Educational Studies, 51, 36–45.

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. (2011a). *Every Child Counts: The independent evaluation executive summary.* Department for Education (DfE) DFE-RBX-10-07. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/193101/DFE-RBX-10-07.pdf

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. (2011b). *Every Child Counts: the independent evaluation Technical report.* Department for Education (DfE) Research Report DFE-RR091a. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/182404/DFE-RR091A.pdf

Torgerson, C.J., Wiggins, A., Torgerson, D., Ainsworth, H., Hewitt, C. (2013). Every Child Counts: Testing policy effectiveness using a randomised controlled trial, designed, conducted and reported to CONSORT standards. *Research in Mathematics Education* 15, 2, 141-153.

What Works Clearinghouse website: (http://ies.ed.gov/ncee/wwc/).