Journal of MultiDisciplinary Evaluation
Volume 9, Issue 20, 2013

JMDE
Journal of MultiDisciplinary Evaluation

ISSN 1556-8180
http://www.jmde.com

The Role of Random Allocation in Randomized Controlled Trials: Distinguishing Selection Bias from Baseline Imbalance

Allyn Fives
National University of Ireland Galway

Daniel W. Russell
Iowa State University

Noreen Kearns
National University of Ireland Galway

Rena Lyons
National University of Ireland Galway

Patricia Eaton
National University of Ireland Galway

John Canavan
National University of Ireland Galway

Carmel Devaney
National University of Ireland Galway

Aoife O'Brien
National University of Ireland Galway

Background: This paper addresses one threat to the internal validity of a randomized controlled trial (RCT), selection bias. Many authors argue that random allocation is used to ensure baseline equality between study conditions in a given study and that statistically significant differences at pretest mean that randomisation has failed.

Purpose: The purpose of this study was to clarify the role of random allocation in an RCT study. Is the role of random allocation to protect against selection bias? And does it have a

Research design: The allocation procedure adopted in this study was stratified and blocked random allocation.

Data collection and analysis: Data were collected using standardised and criterion-referenced tests of reading ability. Data were collected by qualified Speech and Language Therapists. Independent-samples t-tests were used to analyse pretest data.

Findings: The role of random allocation is to protect against

further role, namely to ensure baseline equality and the absence of statistically significant differences between study conditions at pretest?

Setting: The participants for this study were 229 children in 1st and 2nd grade and data were collected as part of an RCT evaluation of a volunteer reading programme piloted in Ireland, Wizards of Words (WoW).

Intervention: Not applicable.

selection bias, and statistically significant baseline differences can result even when random allocation has been successful. Whether or not random allocation has been successful is determined by the generation of the random allocation sequence and the steps taken to ensure its concealment. The size of differences between study conditions at pretest can be important for the analysis of posttest data but does not by itself determine whether random allocation was successful. In addition, there are serious concerns about the appropriateness of tests of significance when comparing two study conditions at pretest.

Keywords: *baseline equality; baseline imbalance; random allocation; selection bias; test of statistical significance*

Introduction

The randomized controlled trial (RCT) is viewed as the most rigorous and reliable study design for evaluation research (Boruch, Weisburd, Tutner, Karpyn, & Littel, 2009; Oakley, Strange, Toroyan, Wiggins, Roberts, & Stephonson, 2003; Shadish, Cook, & Campbell, 2002). This is the case as an experimental design, where participants are randomly allocated to either a control group or an intervention group, can generate statistically significant findings about the size of a programme's effect (multiple control and/or intervention groups are also possible). Random allocation or random assignment is commonly used under the assumption that it removes selection bias. Selection bias occurs when there is "any influence on the allocation of treatment by the investigator (either subconscious or deliberate)" (Altman, 1991, p. 1481), and refers to systematic differences over conditions in respondent characteristics that could also cause the observed effect (i.e., a third variable) (Shadish et al., 2002). If selection bias is removed, any differences in outcomes between study conditions after participants have received their respective treatments can be attributed to the differences between these treatments and not other confounding variables. However, there are a number of issues with random assignment that researchers should consider. The issues addressed in this paper concern the precise role of random allocation and the distinction between selection bias and baseline imbalance. The participants for this study were 229 children in 1st and 2nd grade and data were collected as part of an RCT evaluation of a volunteer reading programme piloted in Ireland, Wizards of Words (WoW).

The Role of Random Allocation

The internal validity of an RCT study is the extent to which systematic error (i.e. 'bias') is minimised. A study is internally valid when the differences between the two study conditions observed at the completion of the treatment (i.e., at posttest) can be ascribed to the different treatments (along with random error) and not to other variables (Juni, Altman, & Egger, 2001). Random allocation has a vital role to play in protecting the internal validity of a comparison between groups, but precisely what is that role? It has been concluded by some (GSR, 2007; IES, 2011; Maynard and Holley, 2008; Oakley et al. 2003; Rossi, Lipsey, & Freeman, 2004; Savage, Carless, & Stuart, 2003; SPR, 2004; Tierney, Grossman, & Resch, 1995) that the purpose of random allocation is to create 'baseline equality' between study conditions, and that baseline differences on such variables are evidence of randomization failure.

According to Rossi et al.'s (2004) textbook on evaluation research, the benefit of random allocation is that it ensures 'equivalence' between the two study conditions at pretest. As allocation of participants to treatment conditions is random, the two groups will have "identical composition," that is, the "same mix" of persons in terms of "their programme-related and outcome-related characteristics," and "identical predispositions." That is, they are "equally likely, without the intervention, to attain any given outcome status" (p. 239). This is important as baseline differences "that are related in any way to the outcomes under investigation will cause errors in estimates of programme effects" (p. 239).

Others share the view that the role of random allocation is to ensure baseline equality. According to the 'Procedures and Standards Handbook' of the *What Works Clearinghouse*, "carried out correctly, random allocation results in two groups that are similar on average in both observable and unobservable characteristics" (IES, 2011, p. 12).

This means that, in any study sample where allocation has been random, the two groups will be similar. The Government Social Research Unit (UK) paper on social experiments concludes that the benefit of random allocation is to ensure against 'systematic differences,' as there should be "no systematic relationship between membership of the programme or the control groups, and the observed and/or unobserved characteristics of the units of study" (2007, p. 4). While that point is not contested here, the authors of the paper infer that a systematic difference can be detected by comparing the baseline data of the study groups. "Sufficiently rich data" should be collected at pretest as this "allows one to determine whether the allocation process has distributed cases randomly" (ibid.).

Recent publications of findings from RCT studies would suggest that researchers share the view that random allocation should ensure baseline equality. In reports of research carried out in social settings, it has been argued that random allocation "distributes pre-treatment variation in subjects evenly" (Maynard & Holley, 2008, p. 29) and randomisation is a "method capable of generating socially equivalent intervention and control groups" and it ensures "that factors that may affect the outcomes of interest are equally distributed between intervention and control groups" (Oakley et al., 2003, p. 175; see Savage et al., 2003, p. 219; Tierney et al., 1995, p. 8).

We have seen that some of its proponents assume that random allocation should create baseline equality. Those who take a largely critical view of RCT methodology also accept that random allocation should produce baseline equality, but they go on to note that baseline equality is difficult to guarantee, in particular in social settings where unobserved as well as observed differences may exist, and for that reason RCT designs are at best questionable in such settings. Thus, in a social setting it is not possible to create 'experimental conditions,' and by this it is meant that it is not possible to have "truly equivalent groups" due to the limitations on our ability "to identify, isolate, control and manipulate the key variables" (Morrison, 2001, p. 72). A related argument is that differences within groups are considered unimportant by those who run RCT studies (on the basis that the role of random allocation is to create groups that are 'equal' with respect to such variables) whereas it is just such differences that may explain observed outcomes. Random allocation cannot guarantee against "heterogeneity in the samples," and this "lack of background information means that important correlates of programme effectiveness may be missed" (Ghate, 2001, p. 27).

It is the case that random allocation, if successful, removes selection bias. However, it does not follow that random allocation should ensure baseline equality. There would seem to be two reasons for the belief that randomization does ensure baseline equality. The first is when there is not a clear distinction made between the results of any one random allocation and the results of all possible random allocations. For example, Rossi et al. (2004) note correctly that over all possible random allocations the two study conditions would be equal at pretest, but they then conclude that in any one study the two groups should be equal at pretest. However, this conclusion does not follow. It may be that over all randomizations the groups are balanced, but "for a particular randomization they are unbalanced" (Senn, 1994, p. 1716), for "in a given trial the subjects with some particular characteristic will not be split equally between groups" (Altman, 1985, p. 126). Shadish et al.'s (2002) classic textbook on experimental studies makes clear that random assignment "equates groups on expectation at pretest" and that this "does not mean that random assignment equates units on *observed* pretest scores" (p. 250; see also Jaded, 1998).

The second reason for the belief that randomization does ensure baseline equality is when *systematic difference* is equated with *baseline inequality*. Selection bias refers to systematic differences over conditions in respondent characteristics at pretest that could also cause the observed effects at posttest. Selection bias is ruled out 'by definition' in a randomized controlled trial as allocation is not based on any systematically biased method (Shadish et al., 2002; see Altman, 1991). However, after securing against selection bias, it is still possible that there will be imbalances (inequalities) between the two study conditions (see Simon, 1979). For example, in the RCT study of the WoW programme, whose objective was to make small improvements in reading skills for young readers, a systematically biased method of allocation would have been to assign participants to groups on the basis of some variable known to influence outcomes, such as baseline reading ability or gender. Random allocation instead ensured the two groups were not systematically biased, although the average reading ability of one group was higher than the other at pretest on various measures (see Table 1). The two groups were not systematically different (their allocation was not biased) even though they were unequal

(their random allocation resulted in an imbalance in mean scores on measures of reading ability).

Table 1
Pretest Comparison of Control and Intervention (Data from the WoW evaluation)

Measure	Study Condition	*N*	Mean	Std. Deviation	*p*	Cohen's *d*
SWR	Control	111	81.05	9.556	.100	.218
	Intervention	118	79.14	7.929		
RA	Control	100	89.41	8.471	.265	.155
	Intervention	112	88.05	9.121		
RC	Control	100	98.35	8.611	.301	.144
	Intervention	108	97.09	8.865		
RR	Control	39	79.92	8.533	.397	.207
	Intervention	29	81.79	9.488		
Spell	Control	110	82.20	9.171	.975	.004
	Intervention	116	82.16	8.526		
BPVS	Control	111	95.51	10.347	.493	.090
	Intervention	118	94.60	9.765		
PA	Control	110	32.53	9.161	.246	.155
	Intervention	118	31.16	8.565		
PK	Control	110	27.24	4.816	.975	.004
	Intervention	117	27.26	4.955		

Notes: SWR = WIAT Single Word Reading, RA = York Reading Accuracy, RC = York Reading Comprehension, RR = York Reading Rate, Spell = WIAT Spelling, BPVS = British Picture Vocabulary Scale, PA = Phonemic Awareness, PK = Phonic Knowledge

Selection bias can result when researchers (or other participants) can predict when one treatment group is about to be assigned and as a result are in a position to exercise 'strategic selection' of participants (Berger & Weinstein, 2004). This is made all the more likely if allocation is based on non-random 'systematic' occurrences, such as date of birth, case record number, date of presentation, or alternate assignment (Schulz & Grimes, 2002). That is, systematic allocation procedures create the possibility of someone predicting what treatment is to be allocated next and on that basis ensuring some specific participant receives one or the other treatment. For that reason, the statement on the Consolidated Standards of Reporting Trials (CONSORT) recommends that the success of the random allocation process depends on both the generation of the random allocation sequence and the steps taken to ensure its concealment (Altman et al., 2001).

In the WoW study, random allocation was 'stratified' and 'blocked' (see Jones, Gebski, Onslow, & Packman, 2001). Stratification is achieved by creating subsets of participants based on variables strongly related to outcomes (prognostic variables) and then performing a separate randomization procedure within each stratum. Randomization is then 'blocked' by ensuring equal numbers within each stratum (see Altman, 1991; Beller, Gebski, & Keech, 2002; Kernan, Viscoli, Makuch, Brass, & Horwitz, 1999; Schulz and Grimes, 2002; Simon, 1979). In the WoW study, there were 16 strata, as children were grouped in 1st grade or 2nd grade in each of 8 schools in each of two cohorts. After a random start (the flick of a coin) treatments were assigned to all the participants within the stratum. In order to conceal the allocation sequence only one member of the research team assigned ID numbers to participants and then carried out the allocation sequence. The research team also assigned ID numbers before data on reading

ability or personal information were collected from participants.

Inappropriateness of Tests of Significance When Comparing Study Conditions at Pretest

As we have already seen, many influential voices in the field of social research support the view that the role of random allocation is to create 'baseline equality' between study conditions. A further and related claim is that tests of significance should be used at pretest to show whether or not randomization has been successful (i.e., has created baseline equality).

Boruch et al. (2009), in their guide to randomized controlled trials for evaluation and planning, note that analysis of baseline data can "reassure the trialist about the integrity of the random allocation process" (p. 169). The use of 'numerical tables' is necessary for quality assurance; that is, to establish that the "randomised groups do not differ appreciably from one another prior to the intervention" (p. 172). Others go one step further, arguing that if random allocation was successful, there should be no statistically significant baseline differences. This was one conclusion of a recent RCT study of a volunteer reading programme:

> Tests of significance indicated that, for all background measures but one, there were no statistically significant differences between the treatment and control groups: the randomisation had worked to create pre-treatment equivalence (Ritter & Maynard, 2008, p. 8).

The evaluation of the *Experience Corps* reading programme also concluded that randomization "was effective in creating two equal groups' as 'none of the group differences were statistically significant" (Morrow-Howell, Jonson-Reid, McCrary, Lee, & Spitznagel, 2009, p. 10), while Pullen, Lane, & Monaghan's RCT study of a volunteer tutoring program noted that "pretest data" confirmed "the experimental and control samples are statistically equivalent" (Pullen et al., 2004, p. 27). In their study of reading support programmes, Savage et al. (2003) analysed baseline data "to check that a pre-test balance was achieved," and the results of their ANOVA analysis "revealed no significant main effect of group" (p. 219; also see Hutchings, Bywater, Eames, & Martin, 2008, p. 19; Baker, Gersten, & Keating, 2000, p. 504).

One explanation for the belief that researchers should use tests of statistical significance on baseline data to ensure that randomisation has been successful is simply an extension of the assumption that random allocation ensures baseline equality in any given study. If the scores on outcome measures of the study groups should be equal at pretest it follows they should not be significantly different at pretest.

A further explanation is the belief that a test of statistical significance can establish whether or not the characteristics of the groups at baseline are the result of chance, and therefore, the result of random allocation. The rationale behind this belief is that, as a statistically significant difference is one that is sufficiently unlikely to be explained by chance, a statistically significant baseline difference must be accounted for by something other than chance, namely a systematic or non-random cause. Therefore, a significant baseline difference would show that the allocation to group was not random.

However, the use of tests of significance for this purpose has been criticised by the CONSORT statement, on the grounds that it is 'inappropriate' when comparing two study conditions at pretest (Altman, Moher, Schulz, Egger, Davidoff, Elbourne, Gotsche, & Lang, 2001; see Senn, 1994; Assman, Pocock, Enos, & Kasten, 2000; Pocock, Assman, Enos, & Kasten, 2002). This is the case as the test "assesses the probability ... that the observed difference, or a greater one, could have occurred by chance when *in reality* there was no difference" (Altman, 1985, p. 126; emphasis added). It should be remembered that the question asked by a test of significance is as follows: would the difference observed in this study sample also be observed in the population (i.e. 'in reality') or instead is it the result of chance, i.e. an artefact of this study sample? If we find a statistically significant baseline difference between study conditions this means that it is not the result of chance *and in addition* we would observe this difference in the population as well. However, the problem arises because the relevant population is the population of all possible random allocations, and in this population there would be no difference between the means of the study conditions. It is for this reason alone that tests of significance are inappropriate at pretest to determine if allocation was random.

This argument can be explained by examining the null hypothesis and the alternative hypothesis that are evaluated when conducting a test of significant differences between groups using baseline data. At pretest, the following should be the null and alternative hypotheses:

$$H_o: \mu_A = \mu_B$$
$$H_1: \mu_A \neq \mu_B$$

where, crucially, μ_A and μ_B are population means, that is "means over all allocations possible under the assignment mechanism" (Senn, 1994, p. 1716). The 'inappropriateness' emerges when we observe that, were random allocation to have taken place, then the null hypothesis would be true: over all random allocations possible, the two groups would be equal. It is true by definition that if enough random allocations were performed the result would be two equivalent groups. However, it does not follow that, if randomization was successful, the observed means of the groups in any one study are equal. So the test of significance is inappropriate at baseline because if randomization has occurred the null hypothesis cannot be rejected.

What is more, a non-significant test result at baseline does not entail there has been no interference in the allocation process (i.e. selection bias). It is possible for there to have been interference in the allocation process and yet for the results of a statistical test to show no significant difference between the scores for the two groups. One can imagine an unscrupulous investigator interfering with the allocation of participants to condition in an effort either to generate two study conditions that were equivalent, or to ensure that one study condition was 'stronger' than another but not significantly different. This is in an effort "to cheat without being discovered" (Senn, 1994, p. 1717); that is, an effort to create the appearance of a successful randomization process.

In many RCT studies, tests of statistical significance are used to assess whether randomization resulted in baseline equality. The reasons why this practice is considered inappropriate by Senn (1994) and others were outlined above. However, Senn goes on to claim that tests of significance can be used to test the 'random allocation mechanism' itself. The argument is that the only circumstances where a difference "could not be assigned to chance were if we had failed to use a chance mechanism in allocating" participants (Senn, 1995, p. 176). According to Senn, a significant difference at pretest is evidence that the researchers have "either bungled or cheated" (p. 177). However, the role that Senn here is giving to tests of statistical significance is also problematic. This argument should fail for the same reasons that the argument calling for the use of tests of statistical significance to assess baseline equality also failed. As noted above, using a test of statistical significance where the null hypothesis is known to be true is questionable. Senn himself pointed out that selection bias need not result in a difference (significant or otherwise) between study conditions, and indeed it may result in baseline equality. The unscrupulous researcher mentioned above interfered with the allocation sequence to ensure that no statistically significant differences would be observed when comparing the study conditions at pretest. This illustrates that a non-random allocation process need not result in statistically significant differences. But it is still the case that a random process may result in significant differences.

How to Deal with Baseline Imbalance

When randomization is successful selection bias is avoided. However, randomization does not guarantee the equality of the study groups. Nonetheless, differences at pretest are of concern to researchers as "it is generally felt desirable to establish the comparability of randomized groups" (Altman, 1985, p. 132) and even a successful random allocation process can result in a difference between the study conditions on "a major factor that predicts outcome" (Torgerson and Torgerson, 2003, p. 39); that is, a baseline imbalance on a major prognostic variable. Because of this, Slavin (2008) has argued there should not be a 'large' baseline imbalance, a difference greater than 50% of a standard deviation.

One way to deal with baseline imbalance is to attempt to reduce the likelihood of it occurring in the first place by making modifications to the random allocation process. One such approach, an option adopted in the WoW study, is to use 'stratified' and 'blocked' randomization (see Jones at al., 2001). Cohort, school, and class year are important prognostic variables influencing outcomes, and for that reason were appropriate for use as strata in random allocation. Stratified randomization is recommended so as to increase study power. It is recommended for studies with a 'small' sample size as a way to reduce differences between the two groups on the relevant variables without having to increase sample size, as "study power is inversely related to variance of the difference between two means" (Kernan et al., 1999, p. 20 & 21; see Schulz and Grimes, 2002, p. 518). Stratification reduces the error variance associated with statistical tests of differences between the groups and therefore increases power. However, as allocation of participants in a stratified study is still random, it does not

guarantee equivalence on relevant variables. Indeed, the claim that stratified randomization is to "ensure that the treatment groups are 'comparable' with regard to factors other than treatment that may affect response" has been described as "vague and problematical" (Simon, 1979, p. 505). Instead, what can be said is that it increases similarity between groups on the variable(s) used for stratification, but not on other variables (Kernan et al. 1999). Other methods can be used to increase study power by reducing the variance of the difference between the two groups' means, including minimization, replacement randomisation, and propensity score matching (see Schulz and Grimes, 2002). However, they too are not judged successful on the basis of whether or not they lead to baseline equality.

Another approach to dealing with baseline imbalance relates to the strategy for analysis of posttest data. It has been argued that, if baseline group differences "are observed, it is essential to adjust for these differences statistically (e.g. covariance analysis) before conducting other analyses" (SPR, 2004, p. 4). However, this is not strictly the case, as baseline imbalance does not by itself dictate the type of posttest data analysis to be conducted. On the one hand, even a "significant [baseline] imbalance will not matter if a factor does not predict outcome" (Assmann et al., 2000, p. 1067). To make such distinctions what is needed is "prior knowledge of the prognostic importance of the variables," "clinical knowledge," and "common sense" along with knowledge of the relevant research (Altman, 1985, p. 130 - 132). On the other hand, a major prognostic variable should be adjusted for using covariance analysis, especially if it was used to stratify random allocation, and regardless of the presence or absence, or significance or non-significance, of baseline imbalances on that variable (Altman, 1985; Assmann et al., 2000,; Pocock et al., 2002,). For instance, pretest scores should be controlled for because they are an important prognostic variable, even if there was no baseline imbalance on this variable. Employing the pretest score on the dependent variable as a covariate will enhance the power of the analysis as it will serve to lessen the error variance in the posttest score when testing for differences between the groups. In addition, it should be noted that error in the covariate will lead to an 'under adjustment' for the covariate (see Field, 2009). It is therefore wise to conduct such an analysis using latent variable modelling to remove the biasing effect of random measurement error on the results (Russell, Kahn, Spoth, & Altmaier, 1998).

What this illustrates is that the implications of baseline imbalance are very different from those of selection bias. If selection bias has occurred, no amount of statistical analysis can undo the damage done to the study. The presence of selection bias implies that the allocation of participants to groups was not random, a necessary assumption for the analysis of data in an RCT experiment (Assmann, et al., 2000). In contrast, a baseline imbalance in itself is not contrary to the assumptions of the statistical analyses used in an RCT study. Indeed, the use of covariance analysis for posttest data can greatly diminish the importance of any baseline imbalances.

This paper has reported baseline imbalances in terms of effect sizes along with the p values for those findings. This was done as pretest scores are a major prognostic variable and pretest scores help give a picture of the study sample prior to treatment. Also, in reviewing a study's findings it is important to know whether there were large baseline differences between study conditions on major prognostic variables as the presence of such differences rightly raise the question of the comparability of the two study conditions. However, it is also the case that there are situations in which baseline imbalance on outcome scores may be a cause for concern, even when random allocation was successful and even when pretest scores can be adjusted for. For instance, data from various programmes of support for weaker readers suggest that the child's rate of progress is related to initial baseline levels, and volunteer reading programmes may not be appropriate to children with severe needs (Brooks, 2002). Therefore, the ability of a study to measure the effectiveness of a volunteer reading programme would be put in doubt where there was a large difference in pretest scores and in one study condition there was a greater preponderance of participants who because of their pretest scores were less likely to respond well to the programme.

Conclusion

In an RCT study the success of the random allocation process depends on both the generation of the random allocation sequence and the steps taken to ensure its concealment. Researchers should focus their attention on getting this part of the research design right and on ensuring their study is implemented in line with this requirement. The role of the random allocation process is to protect against selection bias. However, even a successful random allocation may lead to statistically significant differences in

pretest scores. In addition, there are serious concerns about the appropriateness of tests of significance when comparing two study conditions at pretest. Therefore, on the one hand, researchers should not assume that the presence of statistically significant differences in pretest scores means that random allocation has not succeeded; and on the other hand, the absence of statistically significant differences will not compensate for a failed random allocation process.

Baseline imbalance should be distinguished from selection bias, as the former need not undermine the internal validity of an RCT study. As the study conditions should be comparable at pretest then information on the size of baseline imbalances is important and should be presented when studies are reported. At the same time, not all baseline imbalances are equally important. Using clinical knowledge, relevant research, and common sense, researchers can identify prognostic variables, and estimates of programme impacts should be adjusted for baseline values on such variables.

Acknoweldgements

This paper is based on findings from the evaluation of the Wizards of Words (WoW) reading programme. The programme is run by Barnardos, Ireland, and the Child and Family Research Centre, NUI Galway were commissioned by Barnardos to conduct the evaluation. Funding was received from The Atlantic Philanthropies to carry out this research.

References

Altman, D. G. (1985). Comparability of randomized groups. *The Statistician, 34*, 125-136.

Altman, D. G. (1991). Randomization: Essential for reducing bias. *British Medical Journal, 302*, 1481-1482.

Altman, D. G., Moher, D., Schulz, K. F., Egger, M., Davidoff, F., Elbourne, D., Gotsche, P., & Lang, T. (2001). The CONSORT statement: Revised recommendations for improving the quality of reports of parallel group randomized trials. *Annals of Internal Medicine, 134*(8), 657-662.

Assmann, S. F., Pocock, S. J., Enos, L. E., & Kasten, L. E. (2000). Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet, 335*(9235), 1064-69.

Axford, N., Morpeth, L., Little, M., & Berry, V. (2008). Linking prevention science and community engagement: a case study of the Ireland Disadvantaged Children and Youth Programme. *Journal of Children's Services*, *3*(2), 40-54.

Baker, S., Gersten, R., & Keating, T. (2000). When less may be more: A 2-year longitudinal evaluation of a volunteer tutoring program requiring minimal training. *Reading Research Quarterly, 35*(4), 494-519.

Beller, E. M., Gebski, V., & Keech, A. C. (2002). Randomization in clinical trials. *Medical Journal of Australia, 177*(10), 565-567.

Berger, V. W., & Weinstein, S. (2004). Ensuring the comparability of comparison groups: is randomization enough? *Controlled Clinical Trials, 25*(5), 515-524.

Boruch, R., Weisburd, D., Turner III, H. M., Karpyn, A., & Littell, J. (2009). Randomized Controlled Trials for Evaluation and Planning. In L. Bickman & D. J. Rog (Eds.), *The Sage Handbook of Applied Social Research*. Second edition (pp. 147-181). London: Sage.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioural Sciences*. Second edition. New Jersey: Lawrence Erlbaum Associates.

Dunn, L., Dunn, L., Whetton, C., & Burley, J. (1997) *The British Picture Vocabulary Scale*. Second edition. London: GL Assessment.

Field, A. (2009). *Discovering Statistics Using SPSS*. Third edition (London: Sage).

Ghate, D. (2001). Community-based evaluations in the UK: Scientific concerns and practical constraints. *Children and Society, 15*(1), 23-32.

Government Social Research Unit (GSR) (2007). *Magenta Book Background Papers. Paper 7: why do social experiments?* London: HM Treasury. Available at http://www.civilservice.gov.uk/Assets/chap_6_magenta_tcm6-8609.pdf (accessed January 2011)

Hatcher, P. J., Hulme, C., Miles, J., Carroll, J., Hatcher, J., Gibbs, S., Smith G., Bowyer-Crane, C., & Snowling, M. (2006). Efficacy of small group reading intervention for beginning readers with reading-delay: A randomized controlled trial. *Journal of Child Psychology and Psychiatry, 47*(8), 820-827.

Hutchings, J., Bywater, T., Eames, C., & Martin, P. (2008). Implementing child mental health interventions in service settings: Lessons from three pragmatic randomised controlled trials in Wales. *Journal of Children's Services*, *3*(2), 17-27.

Institute of Education Sciences (2011). *What Works Clearinghouse. Procedures and Standards Handbook*. Version 2.1. Available

at: http://ies.ed.gov/ncee/wwc/documentsum.aspx?sid=19

Jaded, A. (1998). *Randomised Controlled Trials: A User's Guide*. London: BMJ Books.

Jones, M., Gebski, V., Onslow, M., & Packman, A. (2001). Design of randomized controlled trials: Principles and methods applied to a treatment for early stuttering. J*ournal of Fluency Disorders, 26*(4), 247-267.

Juni, P., Altman, D. G., & Egger, M. (2001). Assessing the quality of controlled clinical trials. *British Medical Journal, 323*(7303), 42-46.

Kernan, W. N., Viscoli, C. M., Makuch, R. W., Brass, L. M., & Horwitz, R. I. (1999). Stratified randomization for clinical trials. *Journal of Clinical Epidemiology, 52*(1), 19-26.

Lachin, J. M., Matts, J. P., & Wei, L. J. (1988). Randomization in clinical trials: Conclusions and recommendations. *Controlled Clinical Trials, 9*(4), 365-374.

McIlroy, D. (2010). The fundamental importance of baseline comparisons in a clinical trial. *The Journal of Thoracic and Cardiovascular Surgery, 139*(3), 801-801.

Morrison, K. (2001). Randomised controlled trials for evidence-based education: Some problems in judging "What Works." *Evaluation and Research in Education, 15*(2), 69-83.

Morrow-Howell, N., Jonson-Reid, M., McCrary, S., Lee, Y. S., & Spitznagel, E. (2009). *Evaluation of Experience Corps Student Reading Outcomes*. Available at: http://www.mathematica.mpr.com/publications/SearchList2.aspx?jumpsrch=yes&txtSearch=%22Susan%20Sprachman%22%20or%20%22S.%20Sprachman%22 (accessed March 2010).

Oakley, A., Strange, V., Toroyan, T., Wiggins, M., Roberts, I., & Stephenson, J. (2003). Using random allocation to evaluate social interventions: Three recent U.K. examples. *The ANNALS of the American Academy of Political and Social Science, 589*(1), 170-189.

Pawson, R., & Tilley, N. (1997). *Realistic Evaluation*. London: Sage.

Pocock, S. J., Assmann, S. E., Enos, L. E., & Kasten, L. E. (2002). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting: current practice and problems. *Statistics in Medicine, 21*(19), 2917-2930.

Pullen, P. C., Lane, H. B., & Monaghan, M. C. (2004) Effects of a volunteer tutoring model on the early literacy development of struggling first grade students. *Reading Research and Instruction, 43*(4), 21-40.

Ransohoff, D. F. (2005). Bias as a threat to the validity of cancer molecular-marker research. *Nature Reviews, 5*(2), 142-149.

Ritter, G., & Maynard, R. A. (2008). Using the right design to get the 'wrong' answer? Results of a random assignment evaluation of a volunteer tutoring programme. *Journal of Children's Services, 3*(2), 4-16.

Ritter, G., & Holley, M. (2008). Lessons for conducting random assignment in schools. *Journal of Children's Services, 3*(2), 28-39.

Ritter, G., Barnett, J. H., Denny, G. S., & Albin, G. R. (2009). The effectiveness of volunteer tutoring programs for elementary and middle school students: A meta-analysis. *Review of Educational Research, 79*(1), 3-38.

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation: A Systematic Approach*. Seventh edition. London: Sage.

Russell, D. W., Kahn, J. H., Spoth, R., & Altmaier, E. M. (1998). Analyzing data from experimental studies: A latent variable structural equation modeling approach. *Journal of Counseling Psychology, 45*(1), 18-29.

Savage, R., Carless, S., Stuart, M. (2003). The effects of rime- and phoneme-based teaching delivered by Learning Support assistants. *Journal of Research in Reading, 26*(3), 211-233.

Schulz, K. F., & Grimes, D. A. (2002). Generation of allocation sequences in randomized trials: chance, not choice. *The Lancet, 359*(9305), 515-519.

Senn, S. J., (1994) Testing for baseline balance in clinical trials. *Statistics in Medicine, 13*(17), 1715-1726.

Senn, S. J. (1995). Base logic: Tests of imbalance in randomized clinical trials. *Clinical Research and Regulatory Affairs, 12*(3), 171-182.

Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. New York: Houghton Mifflin Co.

Simon, R. (1979). Restricted randomization designs in clinical trials. *Biometrics, 35*, 503-512.

Slavin, R. (2008). What Works? issues in synthesizing educational program evaluations. *Educational Researcher, 37*(1), 5-14.

Snowling, M., Stothard, S., Clarke, P., Bowyer-Crane, C., Harrington, A., Truelove, E., & Hulme, C. (2009). *York Assessment of Reading for Comprehension*. University of York: Cantre for Reading and Language.

Society for Prevention Research (2004). *Standards of Evidence: Criteria for efficacy, effectiveness and dissemination*. Available at: http://www.preventionresearch.org/StandardsofEvidencebook.pdf

Tierney, J. P., Grossman, J. B., & Resch, N. L. (1995). *Making a Difference: An Impact Study of Big Brothers Big Sisters*. Public/Private Ventures. Available at: http://www.ppv.org/ppv/publication.asp?search_id=7&publication_id=111&section_id=0

Torgerson, C. J. (2001). The need for randomized controlled trials in educational research. *British Journal of Educational Studies, 49*(3), 316-328.

Torgerson, D., & Torgerson, C. (2003). Avoiding bias in randomized controlled trials in education research. *British Journal of Educational Studies, 51*(1), 36-45.

Wasik, B. A. (1998). Volunteer tutoring programs in reading: A review. *Reading Research Quarterly, 33*(3), 266-292.

Wasik, B. A. & Slavin, R. (1993). Preventing early reading failure with one-to-one tutoring: A review of five programs. *Reading Research Quarterly, 28*(2), 179-200.

Yu, L-M., Chan, A-W., Hopewell, S., Deeks, J. J., & Altman, D. G. (2010). Reporting on covariate adjustment in randomized controlled trials before and after revision of the 2010 CONSORT statement: a literature review. *Trials, 11*, 59 (Open Access).

Appendix: Measures Used

In measuring reading achievement, three normed tests were used to assess reading comprehension, reading accuracy, vocabulary, and spelling: the Single Word Reading and Spelling tests from the WIAT-II^{UK}-T; the Reading Accuracy, Reading Comprehension, and Reading Rate tests from the York Assessment of Reading for Comprehension Passage Reading Test (Snowling et al., 2009); and the vocabulary test, the British Picture Vocabulary Scale (Dunn et al., 1997). The research team also employed the criterion-referenced tests of phonological sensitivity, Phonemic Awareness and Phonic Knowledge, developed by Prof. Morag Stuart of the University of London.