Optimising the ingredients for evaluation of the effects of intervention

Lyndsey Nickels[1], Wendy Best[2] & David Howard[3]

[1]ARC Centre of Excellence in Cognition and Its Disorders (CCD) & NHMRC Centre of Clinical Excellence in Aphasia Rehabilitation, Department of Cognitive Science, Macquarie University, Sydney, Australia

[2]Division of Psychology and Language Sciences, UCL, London, UK

[3]Centre for Research in Linguistics and Language Sciences and Tavistock Aphasia Centre North-East, Newcastle University, Newcastle-upon-Tyne, UK

Contact details

Professor Lyndsey Nickels,

ARC Centre of Excellence in Cognition and Its Disorders (CCD,

Department of Cognitive Science,

Macquarie University, Sydney,

NSW 2109,

Australia

Email: lyndsey.nickels@mq.edu.au

Tel: +61-2-9850-8448

Authors' note

During the preparation of this paper Lyndsey Nickels was funded by an Australian Research Council Future Fellowship (FT120100102). Wendy Best was in receipt of a grant from the Economic and Social Research Council (RES-062-23-2721) and a visitor's travel grant from the Australian Research Council Centre for of Excellence in Cognition and Its Disorders, Macquarie University.

Abstract

Background:

In Howard, Best and Nickels (2015, this issue) we presented a set of ideas relevant to the design of single case studies for evaluation of the effects of intervention. These were based on experience with intervention research, methodological expertise and a set of simulations. Our discussion and conclusions were not intended as guidelines (of which there are several in the field) but rather had the aim of stimulating debate and optimising designs in the future. Our paper achieved the first aim - it received a set of varied commentaries, not all of which felt we were optimising designs, and which raised further points for debate.

Aims:

In this paper, respond to the commentaries, and examine the guidelines for evaluation of the design of single case studies proposed by Tate, Perdices, Rosenkoetter, Wakim, Godbee, Togher & McDonald (2013). We aim to further the discussion our target article has started and extend the scope more broadly to issues that were not discussed in our target article (e.g. replication)

Main Contributions & Conclusions:

It is clear that there is a strong consensus that adequately designed single case studies of intervention are an appropriate and important tool in our quest for effective interventions with people with cognitive disorders. It is also the case that many agree that there is no single design that is appropriate for every intervention, every participant or every question. However, whichever design is used it must be able to discriminate between the true effect of an intervention on behaviour, and other potential reasons for change (e.g. practice effects, spontaneous recovery, Hawthorne effects, placebo effects). We have suggested that, depending on the conditions and question to be addressed, this can be achieved using a combination of design features. These may include: multiple pre-treatment baselines, treated and untreated (or subsequently treated) items/processes/tasks, control tasks (not predicted to be affected by treatment even when generalisation is expected), and a cross-over phase (replication across items/tasks). In addition, the outcome of treatment should be evaluated statistically.

We note that generalisation which is clinically desirable, can lead to particular difficulties in attributing change to intervention unless appropriate controls have been included, and that when items are selected on the basis of poor pre-treatment performance, apparent treatment-related gains may in fact be due to regression to the mean and discuss the implications of this for future research.

In Howard, Best and Nickels (2015, this issue), we discussed of some of the features of different designs for evaluating the outcome of intervention at the level of the individual. These included aspects of the different phases of a study (pretherapy, therapy phase and post therapy), assignment of items to conditions (and the problem of regression to the mean) and different methods of analysis, including presentation of a statistical method based on λ coefficients that we called 'Weighted Statistics' (WEST).

As Hillis (2015, this issue) notes we did not, nor did we intend to, provide comprehensive guidelines nor definitive criteria regarding single case experimental design for intervention. Rather we hoped to stimulate discussion and provide pointers for the future. We are pleased that so many of those prominent in the field chose to provide commentaries and develop the discussion beyond our starting point. In this response to the commentaries, we will take the opportunity to clarify our approach and discuss some ways in which it may be sensibly extended. We also will make explicit some aspects of our philosophy that were not the focus of our original article.

We structure this response around a welcome move in the field - the first attempt to develop clear guidelines for rating the methodological quality of single case experimental designs for intervention: The Risk Of Bias in N-Of-1 Trials (ROBiNT) rating scale (Tate, Perdices, Rosenkoetter, Wakim, Godbee, Togher & McDonald, 2013). Using this scale as a starting point, we reflect on the different views on experimental design for intervention, their similarities and differences. A key issue is the extent to which different designs maintain 'experimental control' and we discuss the requirements for experimental control and the different ways this may be achieved. 'Experimental control' means, at least as a first approximation, that it is clear that the change of performance by a participant can only be as a result of an intervention, and not due to other factors such as spontaneous recovery or change, measurement error or poor experimental design.

The ROBiNT rating scale (Tate et al., 2013) built on the earlier Single-Case Experimental Design Scale (Tate, McDonald, Perdices, Togher, Schulz and Savage, 2008). The 2008 scale was focused upon weaknesses of single-subject designs in terms of their threats to validity. They cite 6 key problems with single case designs and map these on to possible solutions and criteria used in the scale (Tate et al., 2008, Table 1): Understanding of the problem issue, determining treatment efficacy, variability in behaviour, observer bias, verification of treatment efficacy and generalisation. The ROBiNT scale amended and refined the Tate et al. scale adding five new items.

Table 1 summarises the items in this scale and, many, we are confident, are uncontroversial. For example, few would disagree that using independent assessors blind to the phase of intervention will reduce the risk of even unintentional bias in the interpretation of results (ROBiNT 5: Blinding assessors). Similarly, measures of inter-rater reliability in coding of observations will ensure that the results are reliable (ROBiNT 6: Inter-rater reliability). Like McDonald (2015, this issue), we also fully support the need for the raw data to be provided (ROBiNT 12: Raw data record)[1]. If, for example, the presentation is limited to means of results from each phase, this can obscure the variability and

[1] Figure 1 in Howard et al. only provides the mean of three pre-therapy baselines and two post-therapy baselines. This was done in order to focus attention on the problem of regression to the mean. If a figure like this were to be presented in a report of a single case treatment study, it would be essential that raw data for every testing point was also provided (most probably in a table).

overall trend in the data. Moreover, we would go further in suggesting that item-level scores should always be made available, at least for the primary outcome measure. This can often now be done by providing supplementary data to a paper, and allows for the possibility of meta-analysis (Beeson, 2015, this issue).

TABLE 1 : Subscales of the Risk Of Bias in N-Of-1 Trials (ROBiNT) rating scale (Tate et al., 2013).

Items 2, 4, 7, 9 and 11 were additions to the ROBiNT scale, and did not appear in the earlier Single-Case Experimental Design Scale (Tate et al, 2008).

Internal validity subscale	
1.	Design
2.	Randomisation
3.	Sampling behaviour (all phases)
4.	Blinding patient/therapist
5.	Blinding assessors
6.	Inter-rater reliability
7.	Treatment adherence
External validity and interpretation subscale	
8.	Baseline characteristics
9.	Therapeutic setting
10.	Dependent variable (target behaviour)
11.	Independent variable (intervention)
12.	Raw data record
13.	Data analysis
14.	Replication
15.	Generalisation

Blinding of the participant and therapist (ROBiNT point 4: Blinding patient/therapist) is a complex issue. It may be feasible and appropriate to blind to the nature of the intervention for some interventions such as computer delivered programmes but it is rarely possible to blind to the phase of the study (intervention or assessment). Indeed, it may not be appropriate in many cases. For example, in many clinical interventions, participants will be involved in selecting their own rehabilitation targets and the process of working towards these is likely to be explicit specifically to increase motivation. This may be true for a range of interventions where a focus on impairment (such as in writing) may be to reach a functional goal (e.g. composing e-mails to a relative). Similarly, there are also a set of approaches where the metacognitive or metalinguistic aspect is core to the therapy, including for example, self-cueing (e.g., Nickels, 1992), use of a cueing-aid (e.g., Best, 2005; Bruce & Howard, 1987) or conversational approaches (e.g., Beeke, Beckley, Johnson, Heilemann, Edwards, et al, in press). In these approaches, blinding would never be possible, and a study without such blinding should not be penalised.

We also agree that there is a need to demonstrate that the positive results of treatment can be replicated across different individuals and therapists (ROBiNT 14: Replication). But, this replication does not affect the strength of the conclusion that can be drawn regarding the presence of treatment effects for a particular individual. We return to the issue of replication in the discussion of design below. Finally, it is clearly valuable for an intervention study to enable evaluation of the extent of any generalisation (ROBinT: 15 Generalisation) from the target task (e.g word retrieval in picture naming) to other, perhaps more functional tasks (e.g. word retrieval in conversation). Or, alternatively, to investigate generalisation from the targeted setting (e.g. conversation with a spouse) to another setting (e.g. conversation with a different familiar or unfamiliar partner).

It is issues to do with design (ROBiNT 1: Design, ROBiNT 3: Sampling behaviour, ROBinT 14: Replication) and analysis (ROBiNT 13 Analysis) where there is the most debate and controversy and this is reflected in the commentaries on our target article (Howard et al., 2014, this issue). Before we discuss these issues we will first briefly reflect on some of the other points in the scale.

Specification of Intervention and its target behaviour

It is uncontroversial that both the intervention (ROBiNT 11: Independent variable) and the target behaviour (ROBiNT 10: Dependent variable; e.g. accuracy of reading aloud of nonwords; accuracy of picture naming; grammatical complexity in conversation) need to be precisely described.

The intervention should always be described in the detail needed to allow replication of the technique. Nevertheless, it is surprising how often, when one comes to attempt precise replication of a treatment, it becomes clear that the detail necessary is simply not provided. For example, all of us have been involved in (independent) studies where the treatment was to be based on Semantic Feature Analysis. Our literature searches found many studies using Semantic Feature Analysis but none of them specified the treatment in the detail we needed to develop clear treatment protocols (and it was clear that 'Semantic Feature Analysis' was in fact many different treatments because the way in which it was implemented varied (Boyle, 2011)).

Coltheart (1983) writes that the treatment should be "described in sufficient detail to allow it to be exactly duplicated by other therapists working with other patients" (pp. 198-9). As Howard (1986) points out, "The only trials that can be used to modify and improve clinical therapy are those where the treatment involved is specified in at least 'cook book' detail" (p. 95). Over the last couple of years there have been some initiatives from those involved in Randomised Controlled Trials to set minimum standards for the description of behavioural interventions (Boutron, Moher, Altman, Schulz & Ravaud, 2008; Hoffmann et al., 2014) . However, these are just minimum standards and they don't always require full detail on behavioural interventions. Included in the description of intervention should be specification of what was done, and exactly how it was done. For example, the number of sessions, the timing and duration of those sessions, and the number of 'treatments' per item per session (dose) should all be clear. However, there also needs to be specification of what may seem like minutiae. For example, if the task was judging whether a picture and a word matched – was the picture provided first and the participant given time to think about the name, before the word is provided? Or were picture and word provided simultaneously. It is possible that the two different presentations of the same task may lead to different processing demands (attempted

naming in the former but perhaps not in the latter). These different processing demands may lead to different outcomes of intervention.

Similarly, all the different response possibilities should be clearly specified. For example what happened when the person failed to respond, responded incorrectly, asked for repetition? What feedback (if any) was given regarding their response? These factors may or may not influence the outcome. We have no idea without empirical data how important these factors might be. For example, Nickels and Best (1996), showed that for one man, AER, the same treatment technique produced different outcomes depending on whether or not there was feedback on and discussion of the responses made.

With the move towards treatment manuals and publication of study protocols we hope that the ability to truly replicate a treatment exactly will improve. However, even when the treatment it is fully specified it also important to document to what extent the protocol was followed (ROBiNT 7: Treatment adherence). This includes adherence by the treating clinician: For example, if the protocol says 'no feedback', how often did the clinician, under pressure from their therapeutic training or the client, relent and provide feedback? Or if the picture and word were to be provided simultaneously for judgement, how often was the picture provided first? Delivery of treatment using computer programmes and apps will increase the standardisation and reproducibility of treatment delivery, but still does not guarantee full adherence to the planned protocol. Treatment adherence also includes adherence by the participant and those around them. For example, when asked to judge whether a picture and written word match, but not to read the written word aloud, does the participant comply or do they read the word aloud on some sessions or trials? Or might, as we experienced in a home programme, a well-meaning friend write down the names of the items in the treatment set as an additional cue, and change the target word for some stimuli (Mason, Nickels, McDonald, Moses, Makin and Taylor, 2011)!

The setting of the therapy may influence outcomes and should also be specified (ROBiNT 9: Therapeutic setting). This may seem less important for some interventions (such as a word retrieval treatment involving repetition of a word in the presence of a picture) but for others the potential impact is obvious (such as practising the use of scripts for ordering a drink in the clinic versus in a cafe setting).

Specification of the characteristics of the individual

It is important that the characteristics of the individual participating in the intervention are fully described (ROBinT 8: Baseline characteristics; Roberts, Code & McNeil, 2003). If the intervention is effective, future attempts at replication will need to use these characteristics as a reference point. Unfortunately, however, it is not easy to specify which characteristics are important. It is clear that the particular impairment under consideration should be defined and defined in some detail. For example, if the treatment is for a spoken word retrieval (naming) impairment then it is not sufficient to say that the participant had impaired picture naming, but to also specify the level of impairment (e.g. semantic impairment, access to phonological output lexicon). For example, Nickels (1992) aimed to carry out a replication of treatment to improve sublexical reading skills (measured by nonword reading) that had previously been successfully used by De Partz (1986). Both TC (Nickels, 1992) and SP (De Partz, 1986) had impaired nonword reading and an inability to convert graphemes to phonemes. However, following treatment to reteach grapheme phoneme correspondences, SP's

nonword reading improved but TCs did not. It became apparent that TC also had an impairment to another component of the sublexical reading system (the ability to blend phonemes), which we have to assume was intact in SP. This was why, despite relearning grapheme-phoneme correspondences, he remained unable to read nonwords aloud. This case illustrates how the absence of a comprehensive analysis of the processing components of the target behaviour can severely limit the ability to replicate intervention studies.

However, even a detailed analysis of the impairment underlying the target behaviour may be insufficient to ensure replicability of results. Nickels and Best (1996), for example, performed the same intervention (word-picture matching) with three individuals who seemed to have the same level of impairment in the target behaviour (naming). However, all three responded differently to the treatment. Consequently, it is important that background testing is more comprehensive: Other aspects of language processing need to be assessed, in particular those that may be involved in the therapy task. Best and Nickels (2000) recommend a 'microanalysis' of deficits, strengths and therapy tasks. For example, in a treatment that involves word repetition, the extent to which repetition is impaired or intact is relevant, and perhaps the extent to which the word that is provided for repetition can also be comprehended. Similarly, it has been hypothesised that non-linguistic cognitive processing influences the success of treatment (e.g., Fillingham, Sage & Lambon Ralph, 2005). In which case it is also essential to test cognitive processing that may influence treatment outcomes and also that may be required for a particular treatment task to be carried out successfully (e.g. working memory in a word-picture matching task).

Indeed, Rapp, Caplan, Edwards, Visch-Brink and Thompson (2013) suggest that broad and comprehensive measurement is required not only *prior* to treatment but also there is a need to measure widely *after* treatment. Their argument is that if one is going to draw conclusions about the impact of a therapy on the language system, then a comprehensive analysis is also required after treatment to identify *how* it has changed. For studies which aim to relate changes in neural activity to effects of treatment, it is particularly important to demonstrate that there is change only in those processes claimed to have changed as a result of treatment.

We turn now to the three key areas over which there is continuing controversy.

Design

It is in the area of the design of single case studies of intervention that there is the strongest debate and this is reflected in the commentaries. We first reiterate our concern that by restricting the use of the term 'single case experimental design' (of intervention) to a particular class of designs some authors are being overly restrictive, and indeed, Goldstein (2015, this issue) particularly objects to our more liberal use of this term. Nevertheless, rather than there being a single defensible approach, there are a variety different kinds of single case experimental designs for intervention each with strengths and weaknesses (see also, e.g., Beeson, 2015, this issue; Hillis, 2015, this issue; Johnson & Kiran, 2015, this issue; Laganaro, 2015, this issue). As Beeson (2015, this issue) suggests, exactly which design is chosen will be affected by experiment-specific factors, including the phase of research, participant characteristics, and the anticipated response to treatment.

All designs examine single cases and use a variety of methods in an effort to maximise experimental control. Of these different approaches, those we labelled Approach A and B are a subset. Approach A is that advocated by, for example, Thompson (2006; 2015, this issue) and Goldstein (2015, this issue). Approach B is that advocated by our target article (Howard et al., 2015, this issue) and many from within the cognitive neuropsychological approach (e.g., Byng & Coltheart, 1986; Franklin, 1997; Nickels, 2002b).

Goldstein (2015, this issue) is concerned that we were confusing single case experimental designs with case studies. We were (and are) not. We agree with Goldstein that this would be misguided: In this use of the term, 'case studies' refers to descriptive studies that do not have any experimental control to allow determination of whether intervention is the source of any change in behaviour (see also Tate et al, 2013). Tate et al. (2013), in their Figure 1, clearly set out those designs which they believe have experimental control and hence can be called single case study experimental designs and those that do not. Designs which are excluded include biphase (A-B) designs (where, as is standard, A refers to a baseline phase and B to a treatment phase) and 'pre-post' designs. We have concerns with these exclusions, in the text, Tate et al further specify that by 'pre-post' they mean "pre-test/post-test type of studies of an individual (where data are not collected *during* the treatment phase)" (p623). We argue that both biphase and pre-post designs can indeed have experimental control and moreover that experimental control is not all or none. What differs across (and within) designs is the confidence with which one can attribute the results to treatment-specific effects: the more restricted the effects of treatment (e.g., item specific effects) the fewer methodological controls are required to be able to unambiguously attribute these effects to treatment (i.e. to maintain experimental control).

Given the debate regarding the issue of experimental control, it is worthwhile to take some time clarifying our position (see also, Byng & Coltheart 1986; Franklin, 1997; Nickels, 2002b). To be specific, if an intervention study has experimental control it means that it has the ability to attribute any change to the specific effects of intervention and to exclude the possibility that they could instead be due to factors other than the intervention.

Pre-test – post-test design

The most simple pre-post design is one where there is a single pre-test and a single post-test with a treatment period intervening (see Figure 1)[2]. Figure 1a illustrates the case where assessment is of items that have been treated (such as the names of a set of pictures) or items sampling processes that have been treated (such as nonwords for reading that include treated letter-sound correspondences). It is clear there has been a change between Pre and Post assessment points, however, it is not possible to attribute this to treatment – there is no experimental control. The improvement could be due to spontaneous recovery, practice effects, placebo effects, Hawthorne

[2] We believe that pre-post designs represent a minimal ABA design: performance is assessed once during a no treatment phase (A), then treatment is applied (B) but performance is not probed, followed by assessment once at the beginning of the second no treatment phase (A). There could be an argument, however, that they represent an AB design: the post-test could be the last (and only) point in the treatment phase (see e.g., Laganaro, 2015, this volume). However, the terminology itself is not of importance, what is vital is that there is an understanding of the strengths and weaknesses of a particular design.

(charm) effects, or any other 'extraneous' variable. Greater experimental control can be gained in a number of ways. In Figure 1b, there is additional sampling of another set of items which are not treated (labelled 'untreated'). These items show no improvement. This provides some experimental control – if the improvement on the treated set had been the result of spontaneous recovery, practice, placebo or Hawthorne effects or any other factor, it would have applied equally to the treated and untreated items tested on the same task and randomly interleaved. The fact that the untreated items did not improve suggests that it was indeed the treatment that improved performance for the treated items[3].

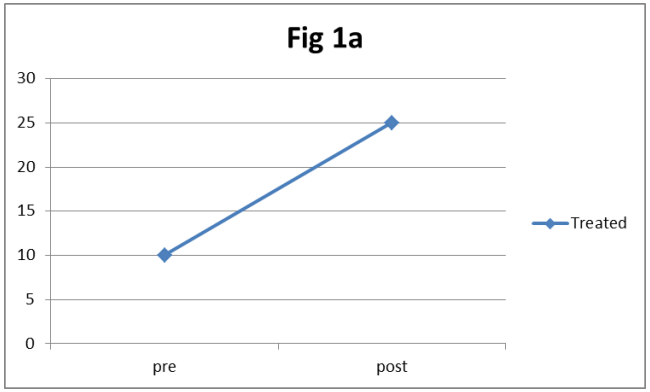
While we would not advocate the use of such a simple design when embarking on intervention research, we would argue that it is NOT the case that this design has no experimental control, contrary to the claim of Tate et al. (2013).

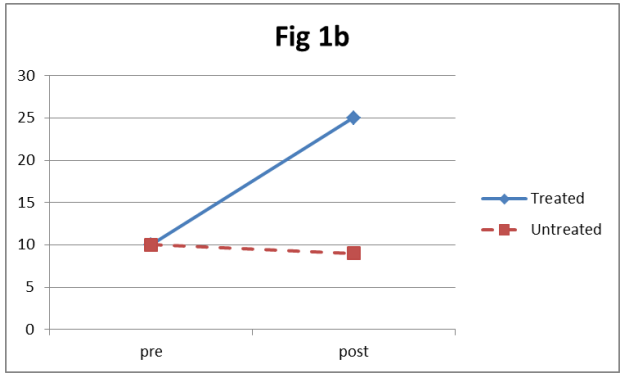
Thompson (2015, this issue) raises the concern that the untrained sets may not show improvement because they are in some way 'harder' than the treated sets. As we noted, to avoid this possibility, it is important that there is matching of (randomly allocated) items within sets for both accuracy and relevant psycholinguistic factors (e.g. frequency, imageability, number of phonemes, semantic category) and then the sets are randomly allocated to treatment or no treatment conditions. Kratchowill & Levin (2010, page 127) note that randomisation is an important feature that can enhance the scientific credibility of even the most basic design. Hence, if sets are not only matched, but also randomly allocated to treated and untreated conditions, it is extremely unlikely that one set will be 'harder'. The use of statistics with randomly allocated sets necessarily takes into account the possibility of some (unanticipated) imbalance making one set harder. Nevertheless, it can, of course, be empirically demonstrated by including a second phase where the previously untreated items are treated (a cross-over phase; Figure 1c). If the untreated items show improvement, when treated, then this provides stronger evidence that the treatment has been effective, through a replication of the treatment effect (see also Laganaro, 2015, this issue).

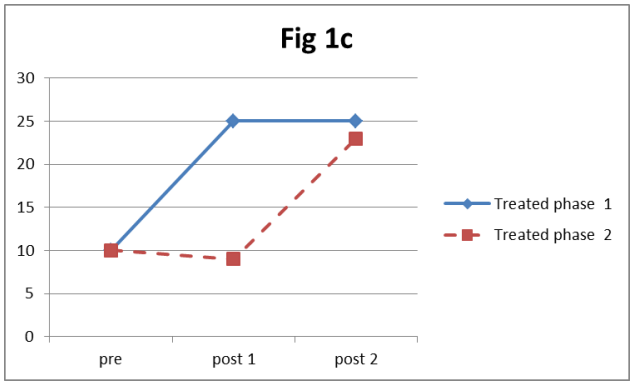
Generalisation across sets

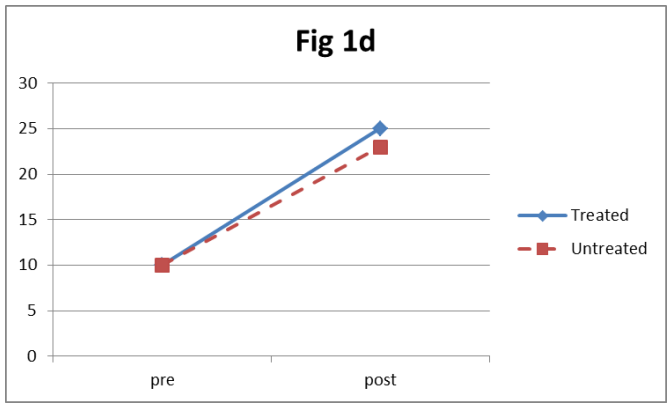
There are conditions under which this simple pre-post design loses experimental control: when both the treated and the untreated items (or items sampling treated and untreated processes) improve, as illustrated in Figure 1d. Here we are, once again, in the position that it is impossible to determine whether the change is due to treatment, which has generalised across items (perhaps by enabling a strategy, or by improving general underlying processes), or due to spontaneous recovery, Hawthorne effects etc. .

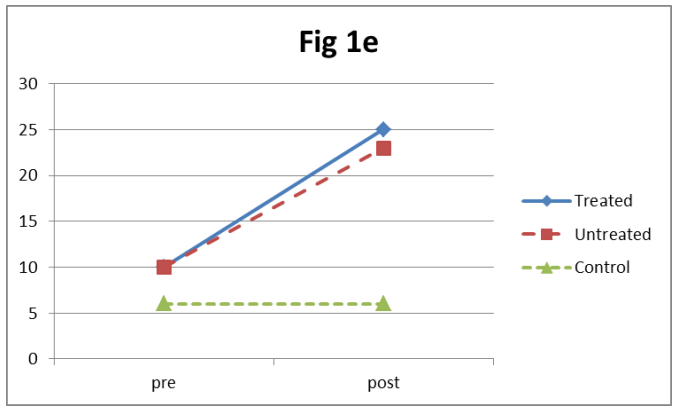
[3] For the sake of illustration we have not used statistical analysis here, however, we are committed to the view that these conclusions could only be drawn if 1) the treated items showed significant improvement and 2) the difference in any change was significantly greater for the treated than the untreated items.











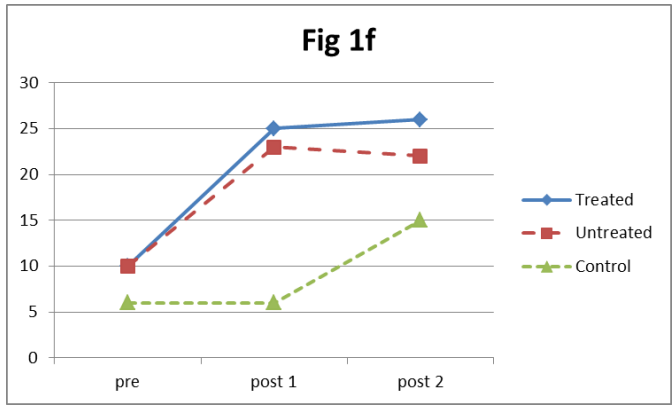


Figure 1: Illustration of different possible patterns of improvement within an intervention design with a single pre-therapy probe and post-therapy probe. Figure 1a illustrates a design with only treated item probes, Figure 1b illustrates improved experimental control by including an untreated set (which does not improve) and Figure 1c demonstrates that this untreated set improves when treatment is applied to it. Figure 1d illustrates improvement of both treated and untreated items in the treatment period. In Figure 1e addition to the design of a control task, which does not improve, suggests that the effects on treated and untreated items are a result of treatment. This conclusion is strengthened in the design in Figure 1f which demonstrates that the control task improves when treatment is directed to it.

Generalisation across sets, while desirable for participants and clinicians, is a challenge for all experimental designs not just for this minimalist pre-post design. Thompson (2015, this issue, page XX) raises this issue, noting that "in Approach B, generalization across sets also renders the results as internally invalid". This can be true (but see below for design features commonly applied in Approach B to circumvent this problem). However, the same applies to Approach A – if, when the first phase of treatment is applied to the first set, there is change in the within treatment probes not only for the treated set but also for other sets, then there is a reduction in experimental control. In other words, if when treatment is applied, all the probed sets improve, Approach A also cannot distinguish between 'true' treatment effects and more general non-specific treatment effects (Hawthorne effects or placebo effects).

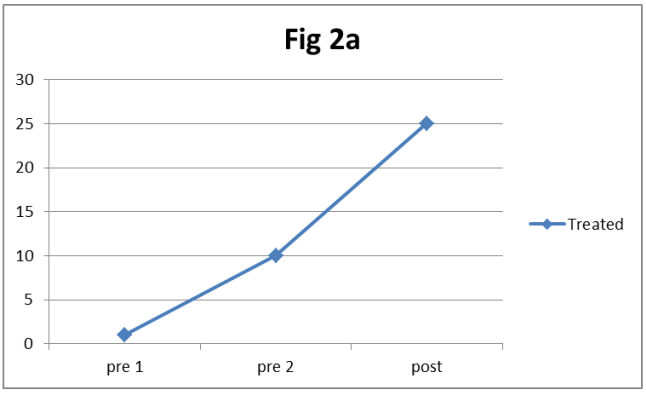
Across both approaches generalisation, which is clinically desirable, can lead to difficulties in attributing change to intervention. Nevertheless, by incorporating additional design features, one can determine whether there really are treatment-specific effects. One possibility is to include the assessment of an additional task, process or set of items that is not predicted to improve as a result of treatment: a control task or item set. If this task/process/item set does not improve then it is more likely any change shown on the treated and untreated subsets of the treated task is due to treatment-specific effects. However,it is possible that the control task may simply be less susceptible to spontaneous recovery/practice or 'extraneous' effects. Hence a demonstration that this task improves when treatment is directed to it would provide greater strength of evidence (i.e., use of a cross-over design, see Figure 1f; e.g., Byng and Coltheart, 1986).

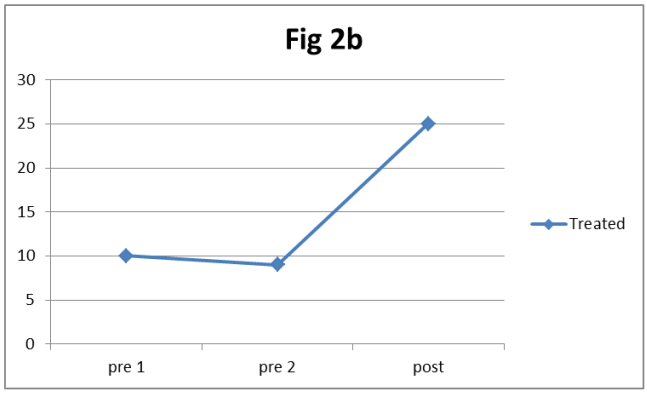
In sum, we contest the view that pre-test post-test designs never have experimental control. Under certain conditions, they have sufficient experimental control to conclude that improvements to treated stimuli are a specific effect of the treatment - specifically, when stimuli are randomly assigned to conditions and there is no improvement to untreated stimuli,.

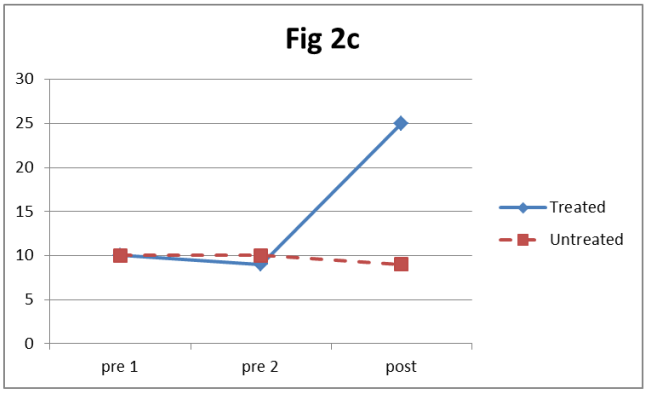
However, we do not advocate the widespread use of pre-post designs with a single assessment at each point. Amongst other things, it is important to probe more than once after treatment in order to determine whether, and for how long, treatment effects maintain (Johnson & Kiran, 2015, this issue). Similarly designs where different treatment techniques are compared within an individual allow for refinement of hypotheses regarding treatment mechanisms (Laganaro, 2015, this issue). Finally, as we will demonstrate in the next section, probing more than once before treatment begins allows for a much more robust experimental design.

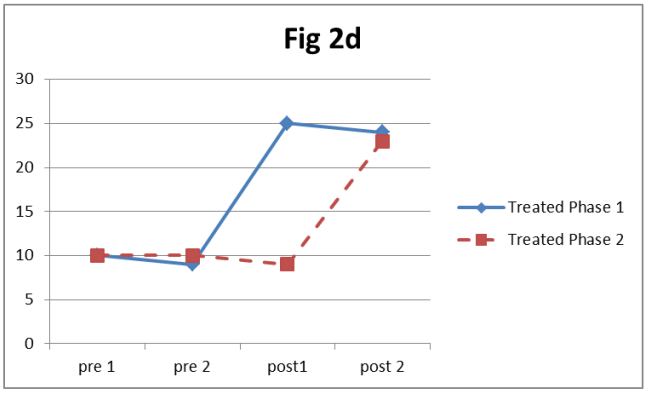
Multiple baseline design

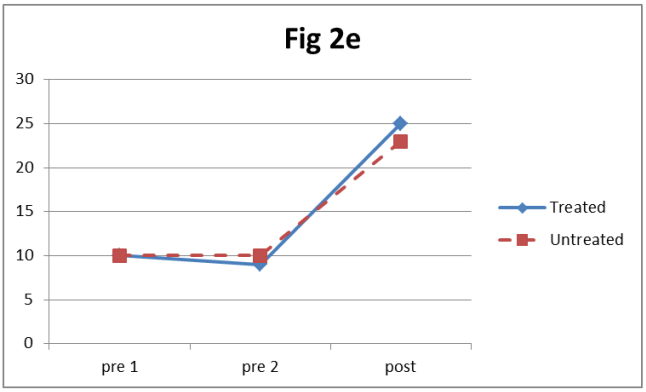
If we return to the situation depicted in Figure 1a, where items improve when treatment is applied, another method by which further experimental control can be achieved is the use of more than one pretreatment probe – multiple baselines. The vast majority of single case intervention research incorporates more than a single pretreatment assessment reflecting the agreement across approaches regarding the importance of this design element (Beeson, 2015, this issue; Kearns, 2015, this issue).

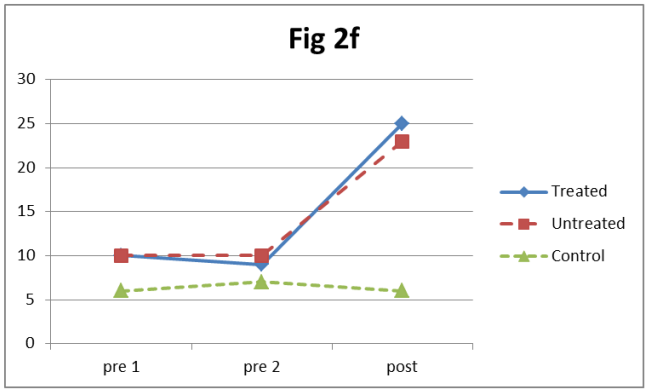


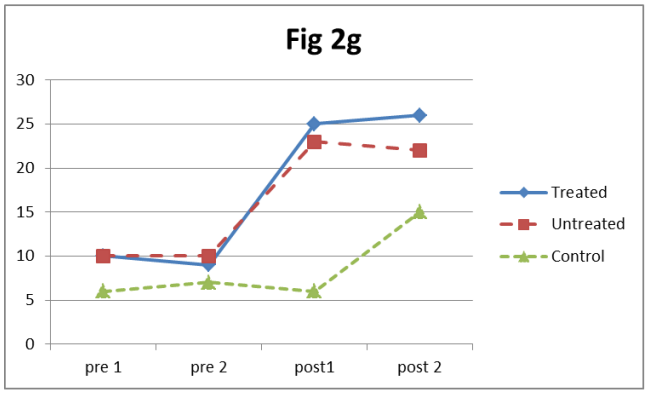












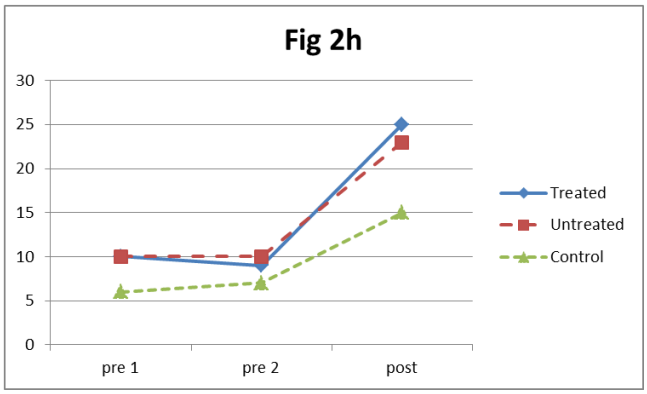


Figure 2: Illustration of experimental control within a multiple baseline design with two pretreatment baselines. Figure 2a shows a situation where the trend is similar across baseline and treatment phases, whereas in Figure 2b, there is a clear distinction between the two phases with greater improvement over the treatment phase. Figure 2c shows the addition of an untreated condition, which does not improve, with 2d adding the additional control of this, previously untreated, set showing improvement when treatment is targeted to it. Figure 2e shows the situation where the untreated set also improves in the treatment phase, and 2f adds a control task to the design allowing dissociation of generalisation from other factors (e.g. Placebo effects), and 2g uses a

cross-over design to demonstrate that the control task can also improve when treatment is directed at it. 2h shows the condition where the control task also improves.

These pre-treatment probes allow an estimate of the rate of change without intervention. This change could reflect spontaneous recovery and/or practice effects. If we compare the pattern across the two pretreatment probes (Pre1 and Pre2) in Figures 2a and 2b, suitable statistical analysis comparing the rate of change across the phases (e.g. WEST-ROC, Howard et al., 2015, this issue) would enable us to infer that in 2a, there is no effect of treatment, but rather the improvement in the treatment phase simply reflects a continuation of steady improvement over time. In contrast, in 2b there seems to be an effect of treatment as there is a clear difference in the rate of change over the baseline (pre1-pre2) and treatment phases (pre2-post). However, we still cannot distinguish whether this difference is a 'true' effect of therapy or due to Hawthorne/placebo effects. One option is to include regular contact with the therapist over the baseline phase (e.g. Best et al., 2011) that matches the intensity and duration of contact provided during the intervention. If there is greater change during the treatment phase within this design we can be sure it was not simply due to the Hawthorne effect. In this case a difference in improvement between the phases reflects an intervention effect surpassing any therapist charm. It is also possible that this contact with therapist could be perceived as 'therapeutic' and, particularly if participants are blinded, or otherwise unaware of the distinction between assessment and intervention sessions, then this manipulation may also provide control for potential placebo effects.

Alternatively, as before, we can start to distinguish between effects of treatment and other factors by incorporating assessment of another, untreated, set of stimuli. Figures 2c and 2e again contrast the outcomes. In Figure 2c the untreated set shows no improvement, and as in Figure 1b, this provides evidence for specificity of treatment effect (and exclusion of Hawthorne and placebo effects). The inclusion of the two pretreatment baselines in Figure 2c simply increases the evidence compared to Figure 1b. Similarly, by including a replication of the effects by treating the previously untreated set (cross-over phase; see Figure 2d), the weight of evidence (or strength of experimental control) is further increased (as also suggested by Laganaro, 2015, this issue). In the scenario of Figure 2e, where both treated and untreated sets improve when treatment is given, we once again require a further methodological manipulation – addition of a control task to distinguish between Hawthorne/placebo effects and 'true' effects of treatment (see Figure 2f). The weight of evidence can be increased by demonstrating that this control task also improves when treatment is directed to it (i.e. a crossover phase between Post 1 and Post 2 in Figure 2g).

Finally, a scenario like that of 2h, where following the treatment phases both treated and untreated items and the unrelated control task improve, makes it most likely that the improvements are not specific to the treatment or that the control task contains processing components shared by the treated task and influenced by the intervention. This would be the pattern one might expect not only with placebo or Hawthorne effects, but also in the scenario suggested by Thompson (2015, this issue): the participant was unwell during the course of the pretreatment baseline but is much improved at posttest.

In sum, as noted above (*generalisation across sets*), there are some scenarios where no approach can distinguish 'real' effects from artefacts - the important point is that, whichever approach is used, design features are incorporated to maximise the chance of being able to do so, and, of course, statistics used to evaluate the results. As Johnson & Kiran (2015, this issue) note the significant advantage of single case studies is that threats to the validity of research can be navigated with flexibility and that strict adherence to a single design strips away much of that flexibility. Furthermore, while not discussed here, if maintenance of treatment effects beyond the treatment period is of interest (as it usually is), then multiple assessments points should be included at posttest (Johnson & Kiran, 2015). Similarly, should the purpose be to compare the effectiveness of two different treatments (Laganaro, 2015, this issue) then subsequent phases can be incorporated to enable this comparison.

Number and characteristics of pre-treatment baselines.

Both we and the advocates of Approach A promote the use of multiple baselines, but there are differences in the recommendations and requirements of each approach (Howard et al. , 2015, this issue).

Number of baselines

As McDonald (2015, this issue) notes, Tate et al (2013; ROBiNT 3: Sampling of behaviour) follow Horner, Carr, Halle, McGee, Odom & Wolery (2005) and Kratochwill, Hitchcock, Horner, Levin, Odom, Rindskopf & Shadish (2010) in setting a criterion of at least 5 data points in every phase (we will return to within therapy probing later). Reduced credit is given for 3 data points. It is unclear from these sources what the justification is for 'at least 5'. Thompson (2015, this issue) differs in suggesting that in Approach A there is no requirement to test many times before treatment and suggests a minimum of two to three measures, but should performance be fluctuating or increasing then baseline probing is continued.

Role of Baselines

Howard et al (2015, this issue) discuss the problems with continuing probing until stability is achieved as advocated by Approach A[4], providing simulations to support these assertions[5]. In addition, the requirement for steady-state baselines (Goldstein, 2015, this issue) means that Approach A has limited applicability for populations where this is not achievable (e.g. acute aphasia, developmental populations). While disagreeing with a requirement for steady state baselines, we agree that the key function of the baseline probes is to provide information regarding trend and variability (Kearns, 2015, this issue; McDonald, 2015, this issue). A minimum of two baselines is required to provide confidence intervals on the level and trend at baseline, however, of course, confidence will increase with greater numbers of observations (and larger set sizes). Finally, we wish to make it clear that we have no in principle objection to the use of many baseline probes. Indeed,

[4] However, as noted by McDonald (2015, this volume), not all those who in general use Approach A design principles use the criterion of stability of baseline (e.g. Kiran & Johnson, 2008).

[5] Thompson (2015, this volume) suggests that in Approach B stable pre-treatment is desired. This is not the case. While a stable pre-treatment performance provides the most straightforward pattern to evaluate change as a result of treatment, the use of statistics comparing rate of change across treated and untreated phases (WEST-ROC) means that the effectiveness of treatment can be evaluated even when there is an upward trend over baseline.

when dependent variables are extremely variable from session to session, this is crucial. Critically, the benefit of many baselines for statistical reliability needs balancing against the practical and ethical considerations of extended baseline testing (e.g. McDonald, 2015, this issue).

Baselines and Regression to the Mean

Kratochwill et al. (2010, p.9-10) note the problems that ensue when cases (e.g., items, participants, schools) are selected on the basis of extreme scores. In this case, their score when remeasured will typically be less extreme. This effect is regression to the mean and Howard et al. (2015, this issue) discuss the issue in detail. Kratochwill et al. note that it is a particular problem when baseline assessments are performed and then the to-be-treated stimuli are selected on the basis of poor performance on the baseline. This point is also stressed in our target article (Howard et al., 2015, this issue) and supported by Willmes (2015, this issue). In this case, regression to the mean is likely to result in improved performance at the next probe which can be confused with an effect of treatment, and more commonly with generalisation of treatment effects to untreated items. As discussed in our target article (Howard et al., 2015, this issue) there are a number of ways to avoid regression to the mean, for example matching treated and untreated sets or tasks for baseline performance (without excluding any items). This has the added benefit of including 'easier' items as part of the intervention. We strongly recommend that researchers do not select items or tasks on the basis of poor performance at baseline. If 'difficult' items are to be the target of treatment, they must be chosen in a 'selection phase' before the start of the baseline testing phase. The baseline assessment phase will then identify the extent of any regression to the mean. The potential trap of regression to the mean is one that affects all researchers, regardless of the approach they use.

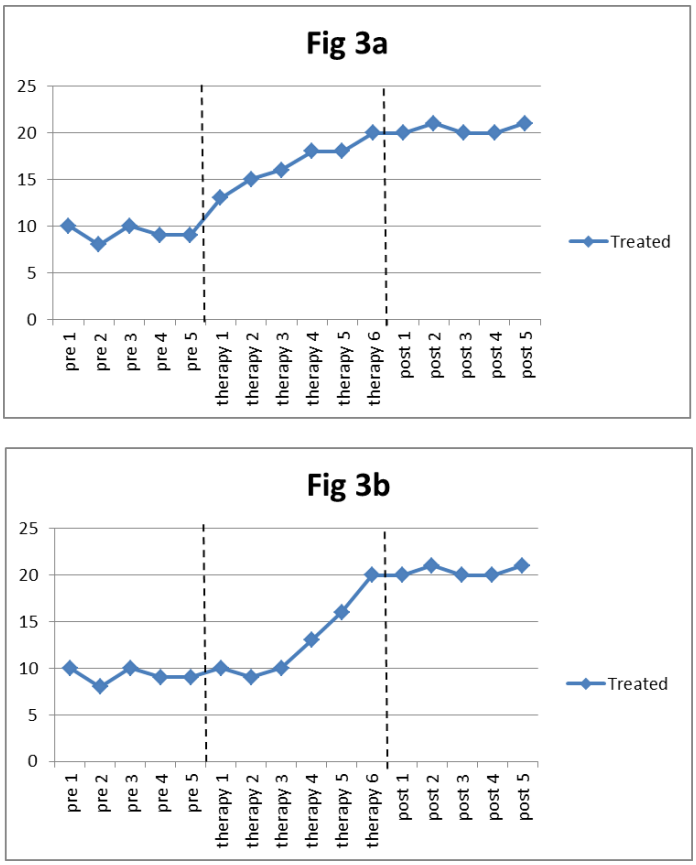
Within treatment sampling of behaviour

As noted above, Tate et al. (2013) require sampling within every phase (ROBiNT 3: Sampling of behaviour) on at least five occasions. The requirement to sample within the treatment phase is a key distinction between the two approaches. Advocates of Approach A argue that this probing is necessary for experimental control, specifically, in order to determine that changes in the rate and trend of change *coincides* with the onset of treatment (e.g., Kearns, 2015, this issue). For example, Kratochwill and Levin (2010, Table 1) state that "the manipulated variable is introduced and concomitant changes in the outcome measure(s) are assessed in the level, trend, and variability between phases of the series." However, consider the scenarios in Figure 3. In both Figures 3a and 3b the baseline levels and rate of change are equivalent, and so are the final levels of performance and rate of change at the post test phase.

However, critically it is only in 3a that the onset of the change in trend occurs simultaneously with the onset of treatment. As Kratochwill et al. (2010, p18) note, for Approach A "The more rapid (or immediate) the effect, the more convincing the inference that change in the outcome measure was due to manipulation of the independent variable. Delayed effects might actually compromise the internal validity of the study." Yet one can imagine many scenarios where the onset of treatment might not result in immediate change in probe measures despite the treatment being the cause of the effects when they do occur. This could be the result, for example, of a severe naming impairment resulting in the need for several exposures during the treatment task (e.g. semantic feature analysis) before benefit is observed in unaided naming, or in a task requiring learning (such as reacquisition of letter-sound correspondence rules) where such learning takes time, or when

learning to use a new strategy in conversation where awareness of the need for repair is likely to precede change in behaviour entailing use of the new strategy. Hence, the interpretation of within treatment probing as required by Approach A can be less that straightforward. In contrast, in Approach B and using WEST statistics as advocated by Howard et al., both scenarios shown in 3a and 3b would be interpreted as indicating a significant treatment effect.

Figure 3: Illustration of when changes in trend do (3a) and do not (3b) coincide with the onset of treatment.



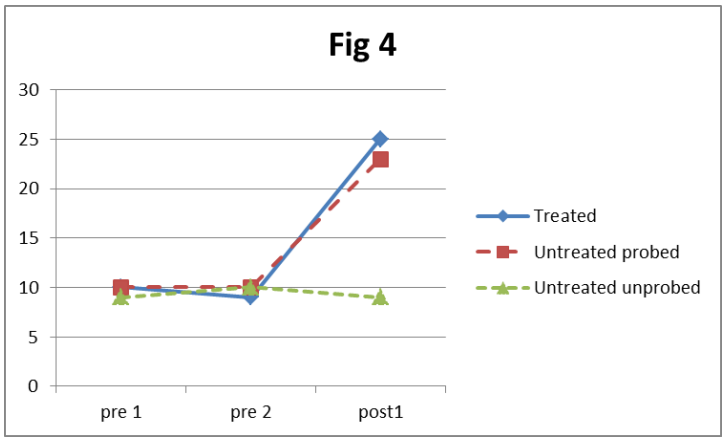
Howard et al. (2015, this issue) discuss three potential problems with the requirement for repeated probing:

- Repeated probing may influence performance

Goldstein (2015, this issue) asserts that any effects of repeated probing are 'small at best'. That may be true, but they are present (e.g., Abel, Schultz, Radermacher, Willmes, & Huber, 2005; Howard, Patterson, Franklin, Orchard-Lisle, & Morton, 1985; Martin, Fink, Renvall, & Laine, 2006 Nickels, 2002a), and to unpredictable extents across individuals. Hence, given that treatment effects are also not guaranteed, and may be small, there is a very real risk of repeated probing effects being interpreted as treatment effects. Indeed, it has been acknowledged that effects that have been

interpreted in the past as reflecting generalisation may have in fact been effects of repeated probing (e.g. Boyle 2011 on semantic feature analysis; Howard 2000, on Howard et al., 1985)

Figure 4: The pattern of results consistent with effects of repeated probing on performance



Paradoxically, the potential effects of repeated probing on performance actually biases *towards* the need for probing on an untreated set of items. If for example, treatment first requires the individual to attempt to name the picture before the therapeutic intervention (e.g., cueing, repeating, semantic feature analysis), then repeated probing is required to be sure that it is the therapeutic intervention rather than the attempts at naming that have produced the effect[6]. Hence, comparison is required between a set that are probed as often as the treated items are named ('untreated probed' or 'naming controls') and a set that are only probed before and after treatment ('untreated unprobed' or 'unseen') (Johnson & Kiran, 2015, this issue). Figure 4 illustrates the pattern consistent with an effect of repeated probing with no additional benefit from intervention.

- Repeated probing may not be possible in practice

Howard et al. note that under some conditions repeated probing may not be possible, including in the case of home programmes. Furthermore, as McDonald (2015, this issue) notes, it is also the case that some participants cannot or will not tolerate repeated probing, particularly when this is on a task that is difficult for them to perform. The added time constraint of probing on top of treatment can also be unfeasible for those individuals with limited attention or stamina (even when a reduced probing set is used as suggested by Thompson (2015, this issue)).

- Repeated probing biases towards small-n therapy sets

Howard et al. discuss in depth the benefits of larger sets of items for the evaluation of the effects of treatment. However, with repeated probing, to reduce how onerous this is, there is a bias towards using smaller items sets. While we appreciate Thompson's (2015, this issue) point that sampling of a

[6] Of course, in some situations this may not be a question of interest: what may be important is that the treatment as a whole has been effective, and whether that is the naming component or the other aspects of the intervention may be unimportant.

subset of the stimuli to probe on each occasion reduces the strength of this bias, unfortunately sampling brings in an additional source of noise and presents difficulty with statistical evaluation when different items are tested at each time point.

Several commentators raised the issue of the practicality and feasibility of large item sets (Hillis, 2015, this issue; Goldstein, 2015, this issue; Johnson & Kiran, 2015, this issue; McDonald, 2015, this issue). We reiterate that 30 is not a magic number - it may need to be less and it may be able to be more: more items mean more power for statistical analysis, hence when more items are available and practical we would recommend they are employed. Moreover, at least for anomia treatment, it seems that the more items that are treated the greater the number of items that improve (Laganaro, Di Pietro & Schnider, 2006; Snell, Sage & Lambon Ralph, 2010).

However, clearly statistical power is not the only factor. First, there is the difficulty of the task for the person with aphasia. If they can only tolerate a small number of items in the assessment task at any one point in time, then the task can be split over a number of sessions (with these sessions close in time and allocation of items to sessions being randomly determined). Similarly, if the treatment task is one that is time consuming for each item (e.g. semantic feature analysis) then not every item need be treated at every point of time and the length of the treatment programme extended to ensure sufficient dosage (Johnson & Kiran, 2015, this issue). Second, there is the issue of naturally restricted sets of items. Hillis (2015, this issue) provides the example of training grapheme-phoneme conversion, where each of three sets were necessarily restricted to 10 items each. While it is not possible to 'create' more of a restricted set, it might be possible to test also this skill in a different way, which may allow greater number of items to be included. For example, by examining reading of words (or nonwords) that contain trained graphemes versus those that contain untrained graphemes. Similarly, while there may be only a limited range of facial emotions (6 basic facial expressions of emotion; McDonald, 2015, this issue), it is possible to sample each emotion several times across different faces, thereby providing a larger pool of stimuli for analysis. McDonald also gives examples of behavioural treatments where a limited set of behaviours are defined and monitored and suggests that this would be hard to accommodate with the need for larger item sets. Once again, while acknowledging that it is less straightforward than with a set of words, it will often be possible. For example, if "meal preparation" is a behaviour that is to be monitored (and presumably intervention provided), this is not a single indivisible entity, but rather comprises a set of specific sub-behaviours, for example, retrieve ingredient 1, ingredient 2, ingredient 3, retrieve utensils 1, 2, 3, cut ingredient 1,cut ingredient 2 etc. At each monitoring point success in carrying out each of these subcomponents is noted and these comprise the data for statistical analysis.

Summary - within treatment probing

In sum, we have no in principle objection to the use of within treatment probing. However, we do not consider it a requirement for experimental control and it is clear that it entails practical constraints. As Hillis (2015, this issue) notes, whether or not you choose to probe frequently depends on the hypothesis you want to test.

Replication

Several commentators point out that replication is central to Approach A (e.g., Goldstein, 2015, this issue; McDonald, 2015, this issue; Thompson, 2015, this issue). Correspondingly this is one of the

elements in the ROBiNT scale: Item 1 (Design) requires at least 3 replications of the treatment effect for full marks and Item 14 (Replication) refers to the repetition of the entire experiment across subjects (again 3 times for full marks). Replication was not addressed in Howard et al. (2015, this issue) and we wish to make clear that we absolutely agree that the weight of evidence for an effect is increased by replication within an individual. As illustrated in the examples in Figures 1 and 2 above, incorporating replication of treatment using a second set of previously untreated stimuli in the same individual can be extremely valuable. However, we disagree with Thompson (2015, this issue) and as argued above, believe there are conditions under which a change in performance from baseline to treatment *can* constitute evidence for an effect of treatment (i.e. experimental control; e.g. Figure 2c, earlier) without replication.

Moreover, we would note that direct replication of the treatment effect within subjects is not always possible. For example, if one is teaching a strategy (e.g., for use in reading, or naming, or conversation) once the subject has acquired this strategy, it will not be possible to replicate this therapy phase. Similarly, if, for example, a particular syntactic structure is acquired, while there may be the potential to teach another syntactic structure, direct replication will not be possible. Moreover, whether or not the subject acquires the second syntactic structure has no impact on the strength of evidence for the effect of treatment on the first syntactic structure.

Replication across individuals and across settings is also informative and important in building a evidence base. As Thompson (2015, this issue) notes the more replications there are, the greater the confidence we can have that the treatment effect is generalizable to other people with aphasia with similar language deficits. However, there is a crucial difference between type A and B approaches with regard to replication across participants. In type A, replication across participants is a design feature which can help establish the effectiveness of an intervention. In type B, successful replication across individuals provides evidence that an intervention can be effective in those different individuals. However, should a treatment fail to replicate in another individual with aphasia, this in no way decreases the strength of evidence regarding a positive effect of treatment in the first individual. Instead, the reasons for the different effects of treatment should be sought (be they differences in impairments, retained language abilities or slight differences in how the treatment was enacted; e.g., Best & Nickels, 2000).

In sum, replication of treatment effects *within* an individual increases the strength of evidence for the original treatment effect. Replication *across* individuals (or across settings) does not impact on the strength of evidence of an original treatment effect in a first individual, but increases evidence regarding the conditions under which that treatment is effective.

Randomisation

On the RoBiNT scale (Tate et al, 2013), item 2 refers to randomisation. We have already stressed the need for randomisation of items into sets and sets to conditions. It is important that there is random selection of which sets are to be treated, or which are to be treated first, or which are to receive treatment A and which treatment B (Willmes, 2015, this issue). However, while randomisation of sequence or treatment is encompassed by this item on the ROBiNT scale, it also refers to randomisation of the specific time points at which each of the different phases commence (A and B). Note, however, that randomisation of when the second A phase begins, also implies randomisation of when treatment phase B ends. This seems to be at odds with either a criterion-referenced start to

treatment (dependent on stability), end to treatment (e.g. finish when 80% accuracy is achieved) or a fixed number of treatment sessions (e.g. finish after 20 sessions). The rationale behind this is closely related to the need to examine when the trend changes in the data (i.e. when improvement starts). We have no strong views on this aspect of randomisation, but remain to be convinced that it adds greatly to the strength of evidence over and above a pre-chosen length of phase.

Data analysis

In Howard et al (2015, this issue) we argued strongly that statistical analysis is required in order to determine whether or not a treatment has been effective, to compare the effectiveness of one kind of treatment over another or the effectiveness of the same treatment on two different kinds of stimuli. Having considered all the commentaries in detail, we hold firm to this view. Tate et al. (2008) in their original Single Case Experimental Design Scale also required statistical analysis (Point 9, Appendix, p401 "To demonstrate the effectiveness of the treatment of interest by statistically comparing the results over the study phases"). In Howard et al., we also provided some new statistical methods (Weighted Statistics - WEST) that we believed would be useful to researchers and clinicians in evaluating their treatment. In this section, we address issues raised by the commentators regarding data analysis.

Visual Analysis

We should note that it is good practice as part of all statistical analysis to examine the patterns in the data - in that sense we support visual inspection of the data. Statistical analysis without visual inspection can lead to mistaken interpretation of the meaning of the statistics.

Our target article included a number of reasons for scepticism about 'visual analysis' as a reliable method for detecting real therapy effects. While some commentators agree that visual analysis is 'flawed' (Kearns, 2015, this issue), other commentators (e.g., Martin & Kalinyak-Fliszar, 2015, this issue) note that there are techniques that may assist in improving its reliability. Moreover, since we wrote the target article there have been articles that promote a very disciplined approach to visual analysis including a variety of quasi-statistical evaluation methods (e.g., Brossart, Vanest, Davis and Patience, 2014; Lane & Gast, 2014,as part of a special issue of Neuropsychological Rehabilitation (volume 24, issue 3-4, 2014) devoted to single case experimental design for rehabilitation). These build on proposals by Kratchowill et al (2010, 2013) but are not used widely in practice and their validity has yet to be tested.

Interestingly, while Tate et al. (2008) required statistical analysis, Tate et al. (2013) do not (ROBiNT item 13: Data Analysis), stating, "Controversy remains about whether the appropriate method of analysis in single-case reports is visual or statistical. Nonetheless, 2 points are awarded if systematic visual analysis is used according to steps specified by Kratochwill et al. (2010; 2013), or visual analysis is aided by quasi-statistical techniques, or statistical methods are used where a rationale is provided for their suitability. One point is awarded if systematic/aided visual analysis is incomplete/not conducted for every phase change or no rationale is provided for statistical analysis".

Unfortunately, therefore, it seems that there is not a consensus that "Visual inspection of data is not an adequate analysis" (Beeson, 2015, this issue, Table 1), although, like Beeson, we hold this view. Many of the other commentators also acknowledge that visual analysis has shortcomings and/or they are generally supportive of the need for statistical analysis or the use of statistical analysis to supplement visual analysis (e.g., Hillis, 2015, this issue; Kearns, 2015, this issue; Martin & Kalinyak-Fliszar, 2015, this issue; McDonald, 2015, this issue; Thompson, 2015, this issue).

Effect Sizes

In Howard et al. (2015, this issue) we discussed several concerns with the use of effect size statistics. (See below for discussion regarding other concerns about interpretation of effect sizes raised by Fischer-Baum (this issue)). Beeson (2015, this issue) discusses effect size statistics in some detail and makes an important and insightful distinction between primary and secondary statistical analyses: the weighted statistics that we propose, along with other non-parametric analyses used in the past (e.g. McNemars, Cochran's Q, Wilcoxon Matched pairs) are examples of primary analyses which answer the question of whether a treatment is effective. In contrast, measures of effect size, including Busk and Serlin's d, are an example of secondary statistical analysis. Secondary statistical analyses aim to provide a standard quantification of treatment effects so that findings can be combined and synthesised across studies in meta-analysis of effects. As Beeson notes, in this sense primary and secondary analysis procedures are complementary rather than competing approaches. Beeson also correctly surmises, that while we do have concerns with the use of Busk and Serlin's d_1 statistic, our primary frustration is that this is often the only analysis that is reported. We are glad that this frustration is not ours alone! Moreover, if item level data are reported routinely, secondary analysis can also be performed from the weighted statistics proposed here.

Weighted Statistics

In the target paper we introduced methods for using λ coefficients to evaluate treatment effects independently from any overall linear change over the investigation period. We argued for a simple method that involved using t tests to investigate treatment effects. We have conducted extensive simulations of these methods varying serial dependency between items (i.e. autocorrelation) and are confident that they can give reliable results including effect sizes and their associated confidence intervals, over a wide range of the number of trials and the number of items.

Three of the commentators (Fischer-Baum, 2015, this issue; Laganaro, 2015, this issue; Willmes, 2015, this issue) while being broadly supportive of our desire to introduce statistical rigour to therapy studies of single subjects, criticise some aspects of our suggested method.

Laganaro (2015, this issue) argues that because the underlying distribution in our example is not normal by a Shapiro-Wilk test, we cannot use a t test to evaluate results. She has a widely held, but incorrect, understanding of when t tests can/should be used. The t test tests just the *mean* of a distribution and assumes only that the *mean* is normally distributed. The Central Limit Theorem guarantees that the mean from a distribution of any shape will be normally distributed as the number of items tends to infinity. How rapidly it converges to normal is largely determined by the degree of skew in the underlying distribution. Our simulations suggest that even with small sample sizes and high item-level serial dependency, the assumption of normality for the means doesn't lead to a raised type I error. In other words, t-tests are entirely appropriate for this kind of data.

Laganaro suggests instead that the data should be analysed using growth curve analysis with binomial outcomes using R. As far as we know R is the only statistical package that can be used to do this. Mirman (2014) provides a very helpful handbook on how one might use R to address these analytic issues. The λ coefficients, that we argue for, apply equally to this approach. Mirman (2014) notes that one can use specific coefficients in such analysis in R but this is 'rather challenging' (p. 45). Laganaro correctly points out that this method of analysis will allow the investigation of the effects of a covariate and its interactions with treatment and recovery effects.There is no doubt that logistic/binomial modelling of growth cuves using R would be possible and sensible with the data from treatment studies. We think it is likely that under almost all circumstances this would produce very similar results to our method; but any conclusive comparison of these approaches will have to be based on a systematic analysis.

Fischer-Baum (this issue) raises a very important question about how we should interpret improvement. His point is that one should not assume that a change of 20% from say 40 to 60% is the same as a change from 70 to 90%. We would agree, and not only because of the shape of the logistic curve – which is Fischer-Baum's focus. The problem is that the assessments used as outcome measures in most therapy studies do not have the psychometric properties to make these comparisons possible.

Any such comparison has to make some kind of assumption about the predicted *shape* of change. Most analyses of therapy studies implicitly or explicitly assume that changes should be broadly linear within and across phases: in our Newtonian universe we believe that data, just like objects, will continue in a straight line except where they are acted on by some external force, in this case therapy intervention. While at first sight sensible, there are reasons for doubting this belief.

For example, an implicit assumption is often made that steady changes in a patient's ability will be reflected in steady changes in probe test performance. What is less often realised is that this is only true when a test's items are equally spaced for difficulty. We simply do not know this for any assessment used in language testing, and given that people with language impairments differ in what affects their performance, the difficulty gradients will almost certainly differ for different people. It is not at all clear in principle or in practice how one might account for this.

We have to assume, then, that the assessments used to evaluate therapies do not necessarily have a linear relationship between ability and score. They may be 'lumpy', in the sense that the same change in ability at one point in the scale may result in a much larger change than at another point in the scale. Once we realise this, we know that we cannot be sure of the shape of change in any test or assessment. Comparisons based on differences in slope between phases or differences in the amount of improvement for different sets need to be viewed with caution because equal changes in underlying ability do not necessarily result in equal changes in score.

However, in the target article, we were arguing for random assignment of items to sets. Random assignment avoids, essentially, all arguments based on the shape of relationship between ability and score. When items are randomly assigned at baseline, significance tests for differences in change between treated and untreated sets make no assumptions about the shape of the relationship between ability and score.

While, as Laganaro (2015, this volume) and Fischer-Baum (2015, this volume) suggest, the ability-score relationship for individual items can be plausibly modelled by a logistic function (with a different intercept and slope for each item) the sum/average of a set of logistic functions will be approximately normally distributed (see Audley & Jonckheere, 1956 for helpful discussion). How rapidly it converges to the normal distribution will depend on the skew of the expected distribution under the null hypothesis. Where there is any doubt we suggested, in our target paper, using randomisation tests (e.g. "randomise" in SPSS) , and Fischer-Baum endorses this approach. However, we think these different statistical approaches should nearly always produce similar results. We note, though, that the randomisation tests necessarily assume random assignment of items to sets (as we recommend).

Laganaro (2015, this issue) also raises the point that, in the target article, we did not address what hypotheses should be addressed in the statistical analysis of a treatment study. She points out absolutely correctly that we report 10 different comparisons in table 5.1. Our purpose there was simply to point out possible comparisons and how they can be carried out. As in any experimental investigation one has to think about the hypotheses and use the appropriate statistics that allow one to address them. For example, were we to have a therapy experiment with three phases (pre-treatment baseline, treatment, and post-treatment follow-up) and two randomly-allocated sets of items one of which was treated in the intervention phase, the following hypotheses would be relevant:

(i) Is there overall improvement over all items over the course of the study?
(ii) Is there greater improvement during the treatment period than the untreated periods?
(iii) Is the overall improvement different for treated and untreated sets? (i.e. Is there an interaction between set and overall improvement?).
(iv) Is the amount of greater improvement during treatment greater for the treated set than the untreated set? (i.e. Is there an interaction between set and greater improvement during therapy?)

Were any of the interactions here significant, one would sensibly use post hoc comparisons to try to understand better the meaning of the results. Table 5.1 was meant to facilitate appropriate comparisons. This is ordinary experimental analysis, with nothing controversial or innovative.

Willmes (2015, this issue) writes a generally supportive commentary, but takes issue with three topics. First, he argues that we have misinterpreted Matthews, Altman, Campbell and Royston (1990). Matthews et al argued (although they were not the first to do so) that reducing an effect of interest to a single number (e.g. 'greater during therapy') would always avoid any problem with autocorrelation in serial data. Willmes points out correctly that Matthews et al only made that argument for the case where units of (statistical) analysis were subjects, and we are extending this argument to the case where the units of analysis are items. We have two responses, both of which suggest that Willmes is incorrect. The first is that Matthews et al.'s argument is not about items or subjects but, in essence, about units of analysis; the statistics don't know anything about the units of

analysis. The second is that we have conducted many simulations that show that whatever the item-level dependency [7](i.e. lag 1 autocorrelation) our methods work.

Second, Willmes suggests that non-parametric procedures might be better. This is a question we have also explored in detail using simulations. It turns out that (i) where the λ coefficients have no skew, the Wilcoxon one-sample test performs well but with lower power than the t test; (ii) where there is skew in the λ coefficients the Wilcoxon one-sample test performs very poorly (because the null hypothesis is based on the assumption that the mean of the differences is zero and the differences are symmetrically distributed). In short, our simulations have demonstrated that non-parametric methods always perform worse than the simple t test, so we recommend use of the t-test.

Third, Willmes argues that we should consider instead statistical methods that are based on the assumption, for the null hypothesis, of permutability of observations from individual items across testing sessions. While these methods are very attractive, unfortunately they should not be used. The reason is that, in intervention data, there can be item-level between-session dependency (i.e. lag 1 autocorrelation), and it would be unwise to make the assumption that there is permutability. We currently have data showing that there can be substantial item-level between session dependency, in naming, for at least one person with aphasia.

To summarise, we argued in the target paper (Howard et al, 2015, this issue) that we could and should use appropriate statistical methods in analysing treatment studies with a single participant and case series. We offered a simple method that is easily implemented, that we have referred to as WEighted STatistics (WEST). Laganaro (2015, this issue) and Fischer-Baum (2015, this issue) correctly pointed out that there are other ways of statistically modelling such studies. There may be other approaches, but all will fall under the umbrella of 'growth curve analysis' (e.g. Mirman 2014). We consider that our approach using λ coefficients to focus on the effects/hypotheses of interest (e.g. is there overall improvement? Is there greater improvement during periods of treatment than when there is no treatment?) is valuable and should be applied whatever the statistical model.

Conclusion.

As Martin and Kalinyak-Fliszar (2015, this issue, pXX) note the "unique dynamic of matching treatment to impairment profile and goals of the person with aphasia ... has been central to the science and 'art' of aphasia rehabilitation, but at the same time presents challenges in testing the efficacy of behavioral treatments for aphasia."

Far from " abandoning single subject controlled experimental research" (Thompson, 2015, this issue) we wish to ensure that all clinical scientists and scientific clinicians embarking on single subject intervention research use designs which enable experimental control to be maintained. There is no 'one size fits all' design. Each individual or research team must choose those elements that are required to answer their research or clinical question and fit within the needs and practical constraints of the particular treatment and particular participant (Hillis, 2015, this issue; Johnson & Kiran, 2015, this issue).

[7] By item-level dependency we mean that for any item the probability correct on trial *n* depends on the performance for that item on trial *n-1*

Finally, as Tate, Perdices, Mcdonald, Togher and Rosenkoetter (2014) note, it is not enough for researchers to design and carry out single case studies of intervention adequately, they also need to be adequately reported so that their quality can be assessed by the reader and others can replicate the design. Tate et al. (2014) report that a set of guidelines are currently under development: Single-Case Reporting guideline In BEhavioural interventions (SCRIBE). The aim of these guidelines is to provide a set of essential features that should be included when reporting single cases, but are not prescriptive regarding design and implementation. We hope that with the publication of these guidelines reporting will improve as was observed when the CONSORT (Consolidated Standards of Reporting Trials) guidelines for randomised controlled trials came into use (Begg et al., 1996).

In sum, we agree wholeheartedly with Goldstein (2015, this issue, pXX) that properly controlled experimental investigations of intervention effects "are not merely means of evaluating treatments; they offer a scientific tool for refining our understanding of behavioral phenomenon [sic]. Understanding variables that inhibit or enhance variability in performance is critical to an iterative process of developing robust treatments."

References

Abel, S., Schultz, A., Radermacher, I., Willmes, K., & Huber, W. (2005). Decreasing and increasing cues in naming therapy for aphasia. *Aphasiology*, 19(9), 831-848**.**

Audley, R. J. and Jonkheere, A. R. (1956), the statistical analysis of the learning process. *British Journal of Statistical Psychology*, 9: 87–94. doi: 10.1111/j.2044-8317.1956.tb00176.x

Beeke, S., Beckley, F., Johnson, F., Heilemann, C., Edwards, S., Maxim, J. & Best, W.(in press). Conversation focused aphasia therapy: Investigating the adoption of strategies by people with agrammatism. *Aphasiology*.

Beeson, P.M. (2015) Primary and Secondary Analyses of Single-Subject Data have Complementary Value. Commentary in response to "Optimising the design of intervention studies: critiques and ways forward" *Aphasiology,* this issue.

Begg, C. B., Cho, M. K., Eastwood, S., Horton, R., Moher, D., Olkin, I., . . . Simel, D. (1996). Improving the quality of reporting of randomized controlled trials: The CONSORT statement. *Journal of the American Medical Association*, 276, 637–639.

Best, W. (2005). Investigation of a new intervention for children with word-finding problems. *International Journal of Language and Communication Disorders*, 40(3), 279-318.

Best, W., Grassly, J., Greenwood, A., Herbert ,R., Hickin, J. & Howard, D. (2011). A controlled study of changes in conversation following aphasia therapy for anomia. *Disability and Rehabilitation*, 33(3), 229-242.

Best, W.M., & Nickels, L.A. (2000). From theory to therapy in aphasia: Where are we now and where to next? *Neuropsychological Rehabilitation*, 10, 231-247.

Brossart, D. F., Vannest, K., Davis, J., & Patience, M. (2014). Incorporating nonoverlap indices with visual analysis for quantifying intervention effectiveness in single-case experimental designs. *Neuropsychological Rehabilitation*, 24. doi: 10.1080/09602011.2013.868361

Bruce, C. & Howard, D. (1987). Computer-generated phonemic cues: an effective aid for naming in aphasia. *International Journal of Language & Communication Disorders*, 22, 3, p 191–201.

Boutron, I., Moher, D., Altman, D., Schulz, K. & Ravaud, P. (2008).Extending the CONSORT statement to randomised trials of nonpharmacologic treatment: explanation and elaboration. *Annals of Internal Medicine*, 148, 295-310.

Boyle, M. (2010). Semantic feature analysis treatment for aphasic word retrieval impairments: what's in a name? *Topics in Stroke Rehabilitation*, 17(6), 411-22.

Byng, S., & Coltheart, M. (1986). Aphasia therapy research: Methodology requirement s and illustrative results. In E. Hjelmquist & L.-G. Nilsson (Eds.), *Communication and handicap: Aspects of psychological compensation and technical aids*. New York: Elsevier Science.

Coltheart, M. (1983). Aphasia therapy research: a single case study approach. In Code, C., and Muller, D.C. (Eds.), *Aphasia Therapy* (pp. 193-202). London: Edward Arnold.

de Partz, M.-P. (1986). Re-education of a deep dyslexic patient: Rationale of the method and results. *Cognitive Neuropsychology*, 3, 149-177.

Fillingham, J. K., Sage, K., & Lambon Ralph, M. A. (2005b). The treatment of anomia using errorless learning vs. errorful learning: Are frontal executive skills and feedback important? *International Journal of Language and Communication Disorders,* 40, 505–524.

Fischer-Baum, S. (2015). How much better? The challenge of interpreting interactions in intervention studies, *Aphasiology,* this issue. DOI: 10.1080/02687038.2014.987050

Franklin, S.E. (1997). Designing single case treatment studies for aphasic patients. *Neuropsychological Rehabilitation*, 7, 401–418.

Goldstein, H. (2015). Throwing the Baby out with the Bathwater: Pitfalls of Misrepresenting Single-Case Experimental Designs. *Aphasiology,* this issue.

Hillis, A.E., (2015). Steam, Broil, or Bake: Good Recipes for Language Treatment Studies. *Aphasiology,* this issue

Hoffmann, T.C., Glasziou, P.P., Boutron, I., Milne, R., Perera, R., Moher, D., Altman, D.G., Barbour, V., Macdonald, H., Johnston, M., Lamb, S.E., Dixon-Woods, M., McCulloch, P., Wyatt, J.C., Chan, A.W. & Michie, S.(2014). Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *British Medical Journal*, 348:g1687 doi: 10.1136/bmj.g1687

Howard, D. (1986) . Beyond randomised controlled trials: The case for effective case studies of the effects of treatment in aphasia. *British Journal of Disorders of Communication,* 21, 89–102.

Howard, D. (2000). Cognitive neuropsychology and aphasia therapy: The case of word retrieval. In I. Papathanasiou (Ed.), *Acquired neurogenic communication disorders: A clinical perspective*. London: Whurr.

Howard, D., Best, N., Nickels, L. (2015). Optimising the design of intervention studies: critiques and ways forward. *Aphasiology*, this issue.

Howard, D., Patterson, K. E., Franklin, S. E., Orchard-Lisle, V. M., & Morton, J. (1985). The treatment of word retrieval deficits in aphasia: a comparison of two therapy methods. *Brain,* 108, 817-829.

Johnson, J.P. & Kiran, S. (2015). Preserving the flexibility of single-subject experimental design: a commentary on "Optimising the design of intervention studies: critiques and ways forward". *Aphasiology,* this issue.

Kearns, K.P., (2015). Mythology and Shape shifters in Clinical Research: False Taxonomies and Erroneous Conclusions: Commentary on 'Optimizing the Design of Intervention Studies: Critiques and Ways Forward.' *Aphasiology,* this issue. DOI:10.1080/02687038.2014.987046

Kiran, S., & Johnson, L. (2008). Semantic Complexity in Treatment of Naming Deficits in Aphasia: Evidence From Well-Defined Categories. *Am J Speech Lang Pathol, 17*, 389-400.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2010). *Single-case designs technical documentation*. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.

Kratochwill, T. R., Hitchcock, J., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial and Special Education,* 34(1), 26–38.

Kratochwill, T. R., & Levin, J. R. (2010). Enhancing the scientific credibility of single-case intervention research: Randomization to the rescue. *Psychological Methods*, 15(2), 124–144.

Laganaro, M. (2015). Optimizing the design of intervention studies: should we test if people require treatment or which intervention has the best outcome? *Aphasiology,* this issue.

Laganaro, M., Di Pietro, M, & Schnider, A. (2006) Computerised treatment of anomia in acute aphasia: issue of treatment intensity and training size. *Neuropsychological Rehabilitation*, 16, 630-640.

Lane, J., & Gast, D. L. (2014). Visual analysis in single case experimental design studies: Brief review and guidelines. *Neuropsychological Rehabilitation*, 24. doi:10.1080/09602011.2013.815636

Martin, N., Fink, R. B., Renvall, K., & Laine, M. (2006). Effectiveness of contextual repetition priming treatments for anomia depends on intact access to semantics. *Journal of the International Neuropsychological Society, 12*, 853-866.

Martin, N. & Kalinyak-Fliszar, M. (2015). The case for single case studies in treatment research: Comments on Howard, Best and Nickels "Optimising the design of intervention studies: Critiques and ways forward" *Aphasiology,* this issue.

Mason, C., Nickels, L., McDonald, B., Moses, M., Makin, K., & Taylor, C. (2011). Treatment of word retrieval impairments in aphasia: Evaluation of a self-administered home programme using personally chosen words. *Aphasiology*, 25(2), 245-268. doi:10.1080/02687038.2010.489258

Matthews, J. N., Altman, D. G., Campbell, M. J., & Royston, P. (1990). Analysis of serial measurements in medical research. *British Medical Journal*, 300(6719), 230-235.

McDonald, S. (2015). Facing the challenges of single case experimental methodology. *Aphasiology,* this issue.

Mirman, D. (2014) *Growth curve analysis and visualization using R*. CRC Press; Boca Raton.

Nickels, L. A. (1992). The autocue? Self-generated phonemic cues in the treatment of a disorder of reading and naming. *Cognitive Neuropsychology*, 9, 155-182.

Nickels, L. A. (2002a). Improving word finding: Practice makes (closer to) perfect? *Aphasiology,* 16(10-11), 1047-1060.

Nickels, L. A. (2002b). Therapy for naming disorders: Revisiting, revising, and reviewing. *Aphasiology*, 16(10-11), 935-979.

Nickels, L.A. and Best, W.M. (1996). Therapy for naming disorders (Part II): Specifics, surprises and suggestions. *Aphasiology*, 10, 109-136

Rapp, B., Caplan, D., Edwards, S., Visch-Brink, E. & Thompson, C. (2013) Neuroimaging in aphasia treatment research: Issues of experimental design for relating cognitive to neural changes. *NeuroImage*, 73, 200–207.

Roberts, P.M., Code, C., & McNeil, M.R. (2003). Describing participants in aphasia research: Part 1: Audit of current practice. *Aphasiology*,17, 911- 932.

Snell, C., Sage, K. and Lambon Ralph, M. A. (2010) How many words should we provide in anomia therapy? A meta-analysis and a case series study. *Aphasiology*, 24 (9).

Tate, R., Mcdonald, S., Perdices, M., Togher, L., Schultz, R., & Savage, S. (2008). Rating the methodological quality of single-subject designs and n-of-1 trials: Introducing the Single-Case Experimental Design (SCED) Scale. *Neuropsychological Rehabilitation*, 18(4), 385-401

Tate, R. L., Perdices, M., McDonald, S., Togher, L., & Rosenkoetter, U. (2014). The conduct and report of single-case research: Strategies to improve the quality of the neurorehabilitation literature. *Neuropsychological Rehabilitation*, 24. doi: 10.1080/09602011.2013.875043

Tate, R. L., Perdices, M., Rosenkoetter, U., Wakim, D., Godbee, K., Togher, L., & McDonald, S. (2013). Revision of a method quality rating scale for single-case experimental designs and n-of-1 trials: The 15-item Risk of Bias in N-of-1 Trials (RoBiNT) Scale. *Neuropsychological Rehabilitation, 23*, 619-638. doi: 10.1080/09602011.2013.824383.

Thompson, C. K. (2006). Single subject controlled experiments in aphasia: The science and the state of the science. *Journal of Communication Disorders, 39*, 266-291.

Thompson, C. K. (2015). Establishing the effects of treatment for aphasia using single subject controlled experimental designs. *Aphasiology*, this issue.

Willmes, K. (2015). The curse of serial dependency in single-case data- Commentary on Howard, Best, & Nickels „Optimising the design of intervention studies: critiques and ways forward“. *Aphasiology*, this issue.