Plasticity, permanence, and patient performance: study design and data analysis in the cognitive rehabilitation of acquired communication impairments

Patricia E. Cowell*[1], Sandra P. Whiteside[1], Fay Windsor[2] and Rosemary A. Varley[1]

[1] Department of Human Communication Sciences, University of Sheffield, Sheffield, UK
[2] Health Professions Department, Manchester Metropolitan University, Manchester, UK

Communication impairments such as aphasia and apraxia can follow brain injury and result in limitation of an individual's participation in social interactions, and capacity to convey needs and desires. Our research group developed a computerized treatment program which is based on neuroscientific principles of speech production (Whiteside and Varley, 1998; Varley and Whiteside, 2001; Varley, 2010) and has been shown to improve communication in people with apraxia and aphasia (Dyson et al., 2009; Varley et al., 2009). Investigations of treatment efficacy have presented challenges in study design, effect measurement, and statistical analysis which are likely to be shared by other researchers in the wider field of cognitive neurorehabilitation evaluation. Several key factors define neurocognitively based therapies, and differentiate them and their evaluation from other forms of medical intervention. These include: (1) inability to "blind" patients to the content of the treatment and control procedures; (2) neurocognitive changes that are more permanent than pharmacological treatments on which many medical study designs are based; and (3) the semi-permanence of therapeutic effects means that new baselines are set throughout the course of a given treatment study, against which comparative interventions or long term retention effects must be measured. This article examines key issues in study design, effect measurement, and data analysis in relation to the rehabilitation of patients undergoing treatment for apraxia of speech. Results from our research support a case for the use of multiperiod, multiphase cross-over design with specific computational adjustments and statistical considerations. The paper provides researchers in the field with a methodologically feasible and statistically viable alternative to other designs used in rehabilitation sciences.

Keywords: cross-over design, apraxia of speech, computerized speech therapy, lexical duration, lexical accuracy, rehabilitation

INTRODUCTION

In the field of speech and language rehabilitation research, there has been debate regarding the most appropriate design for evaluating the outcomes of treatment. In general terms, the debate has addressed the relative advantages and disadvantages of randomized controlled trials in comparison to single case or case series designs. A number of randomized controlled trials for treatment of developmental and acquired speech and language impairments have been conducted (e.g., Lincoln et al., 1984; Glogowska et al., 2000), but the methods of such trials are often subject to criticism. One key issue is the difficulty in blinding participants in studies of behavioral interventions to the arm of the trial to which they have been randomized. A second major issue is that participants often did not receive a single standard protocol treatment, but rather individuals were administered interventions that were appropriate to their particular profile of cognitive–behavioral strengths and weaknesses. In the face of such problems, Howard (1986) advocated the use of single case research designs in order to establish the outcomes of interventions. To some degree, the limited generalization from a single case to the larger population was addressed through the use of case series designs where a number of participants were administered an intervention via a standard protocol (e.g., Robertson et al., 1995).

The multiperiod cross-over design is a staple tool for researchers in the clinical sciences, primarily for the study of drug treatment effects. It is a powerful design which allows direct comparison of within- and between-group effects linked to specific treatments and their clinical control or sham conditions. Yet, within the clinical cognitive neurosciences, cross-over designs are not routinely used. The difficulty presented by cross-over designs in neurocognitive applications is that joining of data from the multiple arms of the design has specific statistical requirements that depend on washout effects between study periods. This works well for treatments where washout is a typical and expected event, for example with pharmaceutical interventions (Senn, 2000). However, for cognitive interventions in neurorehabilitation, treatment objectives are often to instate a permanent change to the system which obviates the notion of washout. Thus, the very success of a cognitive treatment brings with it a lack of washout between study periods and, thus, the potential for carry-over effects. This is perhaps the strongest argument against the use of cross-over designs, and perhaps explains why it has not been the design of choice for the field of neurocognitive rehabilitation research. As a result, the field has not routinely applied a study design that is, for many other reasons, well adapted to study functional change in relation to

cognitive interventions and therapies. For instance, cross-over designs are ideal for chronic or incurable medical conditions since the power of the repeated measures element of the design can be harnessed whereby individuals serve as their own controls through the introduction of treatment and sham interventions. These are particularly salient benefits for clinical research into disorders and disease states that are characterized by population heterogeneity, a hallmark of neurocognitive conditions following stroke or brain injury.

Applications are seen in clinical research areas related to behavioral and cognitive neurorehabilitation, such as neuropsychiatry (Borras et al., 2009), stroke recovery (Mount et al., 2007; Ploughman et al., 2008), and cochlear implantation (Petersen et al., 2009). Cross-over studies have primarily been employed in the field of cognitive rehabilitation for the purpose of allowing participants to serve as their own controls. Mount et al. (2007) demonstrated this principle in a study designed to contrast errorless versus trial and error based training of every day motor tasks in recovering stroke patients. Cross-over design has also been employed to progressively include large numbers of participants by offsetting the start point of the study for part of the sample (Borras et al., 2009). In the treatment of acquired communication disorders, use of cross-over designs is rather limited, particularly for group studies. However, there are some published works that showcase the potential for wider application of this design; these are reviewed below.

Cross-over designs have been used successfully in case series studies to track patient-specific responses to therapeutic intervention sequences by contrasting the impact of multiple treatments within each patient. For example, Fillingham et al. (2005) conducted a series of case studies where cross-over design was used to evaluate two therapies for treatment of word finding deficits in seven adults with acquired anomia. More recently, Raymer et al. (2010) applied a cross-over design to explore the efficacy of errorless versus errorful training interventions for spelling in a series of four cases of acquired agraphia. Cross-over design was also used to examine the combined impact of the pharmacological agent dexamphetamine with behaviorally based anomia therapy in two patients with aphasia (Whiting et al., 2007). In a larger scale application, Carlomagno et al. (2001) applied a multiperiod cross-over design to investigate the impact of lexical versus non-lexical therapy on writing abilities in patients with aphasia. The key difference in the two study phases was the timing of the lexical versus non-lexical therapy period in relation to a third non-writing based therapy (**Table 1**). The authors measured for change after each writing-based therapy in relation to performance at the most recent checkpoint and for stability in relation to the non-writing based therapy. Further analysis, using ANOVA, was conducted for data collapsed across phases in order to examine the effects of the two types of writing therapy on three measures of writing ability. Results were then reported as a function of main effects for therapy, task and their interactions, making this study a good example of the potential that cross-over designs have to document the impact of multiple interventions over time. Together, the studies reviewed above highlight the values of (1) having each case serve as his/her own control; and (2) using multiple checkpoints to measure progress along a trajectory of cumulative therapeutic change.

A major challenge for researchers in cognitive rehabilitation is that the treatments and therapies cannot typically be disguised. This precludes cognitive rehabilitation evaluations from some of the more stringent aspects of clinical trial design such as the double-blind experiment available to other clinical scientists (e.g., drug trials). In the field of acquired communication disorders, for example, a patient is often acutely aware of their disability and will also understand whether or not an intervention involves speech and language related content. For this reason, the notion of a sham condition compared to a speech–language intervention cannot be blinded to the patient-participant in the study. It is also the case that distinctions between treatment and control conditions cannot be hidden from researchers at the front line of treatment delivery and data collection. Indeed, the only stages of the data processing pipeline which can be successfully blinded in terms of the allocation of participants to experimental conditions are data scoring and some preliminary aspects of data analysis. The consequences of these limitations for researchers are multifold and must be carefully considered in the design of treatment studies.

The premise of this article is that the multi-period cross-over design can be used to monitor and better understand a range of problems in non-blinded neurocognitive treatment studies. Just as cross-over designs have proved useful in the implementation of case series studies to evaluate communication based therapies, they can serve as useful frameworks for larger scale group studies. The logical development of these arguments will be made with a pilot dataset that was used in the initial stages of our research group's program to investigate the efficacy of computer based treatment for acquired communication disorders. Specifically, the methods and results sections below will consider treatment effects, "sham" or placebo effects, carry-over and other "nuisance" variables in the context of conducting an open (i.e., non-blinded) two-period cross-over design in the experimental evaluation of cognitive neurorehabilitation for acquired apraxia of speech.

It is not the aim of this paper to argue for the superiority of cross-over designs in all cases, or claim that they are always the better choice over parallel designs. Rather, the purpose of the current paper is to describe how data from therapy experiments can be used, with some key considerations in terms of baselines and some essential transformations of the dependent measures of interest, to allow researchers to harness the power of cross-over designs beyond the level of individual case series comparison. It also describes how a "lack of washout effects" can actually be exploited in a two-period cross-over study to make important and systematic observations about functional gains in relation to therapy and its delivery in relation to control or sham procedures. Overall, the objective is to systematically work through, by way of practical example, both the potential benefits and limitations of cross-over designs in the application of cognitive neurorehabilitation of acquired communication disorders.

METHODS
METHODOLOGICAL CONSIDERATIONS
This section provides a detailed methodological rationale for the development of novel solutions to measure changes in speech production from a study of computerized treatment for acquired communication disorders. Factors such as the choice of behavioral

Table 1 | Examples of cross-over study design in the cognitive neurorehabilitation literature. Time points of behavioral or cognitive measurement are schematically represented as T1–Tk.

Study	Phase	T1	T2	T3	T4	T5
Borras et al. (2009)	1	Post self esteem intervention	Post 12 week rest period			
	2	Post 12 week wait period	Post self-esteem intervention	Post 12 week rest period		
Carlomagno et al. (2001)	1	Post traditional therapy	Post writing therapy (lexical)	Post non-writing therapy	Post writing therapy (non-lexical)	One month follow-up
	2	Post traditional therapy	Post writing therapy (non-lexical)	Post non-writing therapy	Post writing therapy (lexical)	One month follow-up
Fillingham et al. (2005)	1	Post naming therapy with counterbalanced treatment order: Errorful→errorless sequence in session 1 of 10	Five week follow-up			
	2	Post naming therapy with counterbalanced treatment order: errorless→errorful sequence in session 1 of 10	Five week follow-up			
Mount et al. (2007)	1	Post errorless learning on wheelchair task	Post trial and error learning on sock donning task			
	2	Post trial and error learning on wheelchair task	Post errorless learning on sock donning task			
Ploughman et al. (2008)	1	Post exercise therapy on cognition; Post control task on motor ability	Post control task on cognition; Post exercise therapy on motor ability			
	2	Post control task on cognition; Post exercise therapy on motor ability	Post exercise therapy on cognition; Post control task on motor ability			
Raymer et al. (2010)	1	During errorful spelling therapy	Post errorful spelling therapy	During errorless spelling therapy	Post errorless spelling therapy	One month follow-up
	2	During errorless spelling therapy	Post errorless spelling therapy	During errorful spelling therapy	Post errorful spelling therapy	One month follow-up
Whiting et al. (2007)	1	During naming therapy with dexamphetamine	Post naming therapy with dexamphetamine	During naming therapy with placebo	Post naming therapy with placebo	
	2	During naming therapy with placebo	Post naming therapy with placebo	During naming therapy with dexamphetamine	Post naming therapy with dexamphetamine	

Studies in the area of communication based therapies are highlighted in bold.

probes that are robust to the potential bias of the open cross-over design and formulation of raw measures into dependent variables sensitive to neurocognitive plasticity will be covered. The inherent challenges of defining indices of change, and statistically analyzing data for evidence of successful rehabilitation, will be considered in light of: (a) the ever-shifting baselines that emerge through the course of successive treatment and control procedures; (b) the issue of whether or not to combine data from multiple phases of a multi-period cross-over design; and (c) individual differences in patient severity represented in acquired communicative and cognitive impairments.

There are several aspects of cross-over designs that lend themselves to in-depth examination of neurocognitive change. First, one is able to follow participants' progression through two or more interventions using within-subjects comparisons that provide the statistical benefits of repeated measures analysis. Second, one can evaluate the presence of carry-over effects. Carry-over effects have traditionally been viewed as the unwanted result of an incomplete washout phase between the administration of a pharmacological intervention in period 1 and the administration of a placebo in period 2. In severe cases, carry-over effects can prevent the joining of the study phases and, as a result, restrict analysis to between group comparisons of the first period results (Senn, 2000). However, in the current study we demonstrate that there is much to be learned about the psychology of cognitive neurorehabilitation in relation to changes in the placebo/ sham phase and how the order of delivery, in terms of treatment or sham first, can inform the actual rehabilitation process. [See Section "Methodological Development" for details of carry-over effect measurement using lambda (λ)]. Finally, the multiple dependent measures that are commonly used in cognitive rehabilitation research can be

strategically analyzed in relation to both experimental therapies and sham/control interventions (Carlomagno et al., 2001; Fillingham et al., 2005; Raymer et al., 2010). Provided a sensitive means of measuring ongoing change can be employed, cross-over designs can be used to examine the differential impact of treatment and sham effects, or multiple treatment effects, in the context of participant expectations and psychological awareness of their own cognitive progress.

PILOT DATA

The exemplar for this article involves data from a pilot study of computerized treatment for apraxia of speech. A subset of data for these six cases will be presented in the current article and the entire study described in more detail in another paper (Varley et al., in preparation). This work provided the basis for a larger scale clinical trial which is currently underway. The pilot sample were six participants (one male, five females), adults with acquired apraxia of speech, aged 53–83 years (mean = 66 years). All were chronic cases with apraxic symptoms ranging from mild to severe (one mild, three moderate, two severe) and accompanying symptoms of aphasia. The study protocol was approved by The North Sheffield NHS Research Ethics Committee (NS200331635) and all participants provided informed consent prior to taking part.

The study followed a two-period, two-phase cross-over design with multiple baselines. The periods consisted of (1) an active treatment phase that involved self-administered interaction with speech-based computer software, and (2) a sham phase that involved self-administered interaction with visuospatial computer software. The phases consisted of a "sham first speech second" condition and a "speech first sham second" condition. Participants were randomly allocated to the conditions such that three participants were studied for each phase. Participants 1–3, allocated to the Speech First condition, were three females (severe AOS age 53, mild AOS age 59, moderate AOS age 69). Participants 4–6, allocated to the Sham First condition, were one male (severe AOS, age 59) and two females (ages 70 and 83 years, moderate AOS). Multiple dependent measures consisted of target and control behaviors designed to respond differentially to the speech based treatment. The design structure is presented in **Table 2**.

The speech program was designed to be consistent with principles that maximize sensory-motor components in speech processing, and minimize abstract tiers of representation and use of generative-computational mechanisms in speech control (Varley and Whiteside, 2001). Participants received multimodality sensory stimulation (auditory, visual orthographic, visual object, somatosensory) of whole words prior to attempts at speech production. The aim was to activate a broad network of sensory-perceptual zones, and mirror neuron mechanisms that integrate input systems with action control (Wilson et al., 2004). In order to facilitate the reacquisition of cohesive motor patterns, errorless learning and error reduction strategies were employed. Before output was required, performance was primed by the multimodality stimulation, and participants were required to imagine production before overt production (Page et al., 2005). Furthermore, the intervention aimed to achieve a high dosage of stimulation and massed practice in order to facilitate reorganization (Pulvermüller and Berthier, 2008). This was achieved by designing software programs that allowed the participant to self-administer intervention. Thus, treatment could be administered at times and locations convenient to the user, and did not require the presence of a clinician. The speech program involved stimulation of 30 treated words. In the assessment phases, the ability to produce these words was evaluated in terms of word accuracy and word duration. Word accuracy represents a functional measure which links to the listener's likely ability to understand a word. Word duration was an acoustic measure that reflected the cohesiveness and fluency of on-target words (Kent and Rosenbek, 1983).

Two types of controls were included within the study. One was a set of dependent measures designed to control for possible treatment effects on untreated word items. The other was the inclusion of a sham intervention to control for behavioral changes due to the introduction of any novel computer task to a patient with communication impairment (i.e., to control for placebo effects). These two distinct control mechanisms are described in detail below.

Although not limited to use in cross-over designs, a useful element that may be incorporated into treatment studies is a set of multiple dependent measures that allow one to evaluate: (a) dissociations in relation to communication therapies and sham interventions; and (b) potential generalizations between behaviors which were specifically targeted to those which were not targeted by a given therapy. Therefore, in addition to evaluating performance on the treated word set, the production of 30 items from each of two untreated word sets was also assessed. In order to explore generalization of behavioral change, participants produced untreated phonetically- and frequency-matched words (e.g., treated word "sick"; matched word "sit"), and frequency-matched but phonetically distinct forms (e.g., treated word "sick"; control word "catch") (Leech et al., 2001). For the purposes of this article, only data from treated items will be considered.

With regard to the sham program, this involved no stimulation of speech production. The program interfaces mimicked those of the speech program, however it contained activities such as short-term visual memory tasks (recognizing patterns) and timed jigsaw completion. There was some attempt to blind participants as to the sham nature of the program in order to motivate them to practice with the sham program with some frequency in order that the "dose" levels of the two programs would be equivalent. Participants were informed that the program would assist them in using a laptop computer and would develop abilities in attention,

Table 2 | Two period, two phase cross-over study design for the apraxia of speech treatment study. Time points of behavioral measurement are represented as B0–B3 for baselines and R1–R4 for reassessment measures after interventions. Speech and sham interventions were 6 weeks each in duration; the rest phase was 4 weeks in duration; follow-up was conducted 8 weeks after completion of the second intervention period.

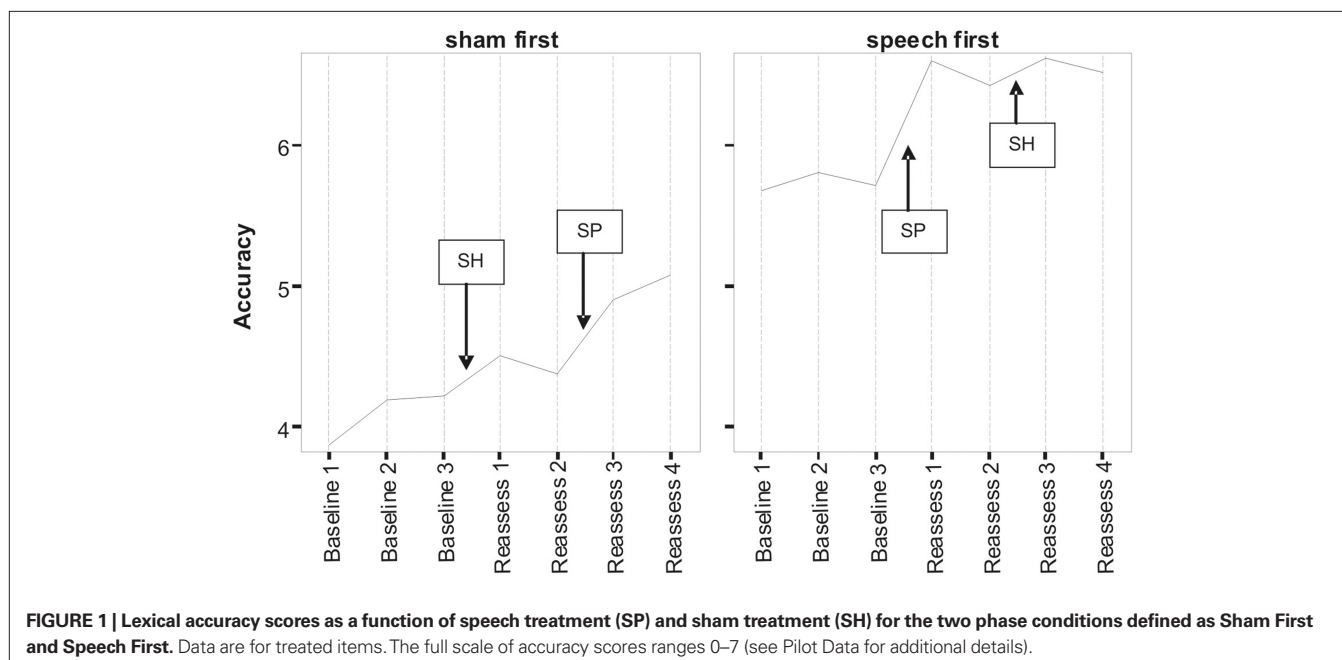| Phase		Period 1		Period 2	
	B0–B3	R1	R2	R3	R4
1 Sham first	Repeated baselines	Post sham	Post rest	Post speech	Follow-up
2 Speech first	Repeated baselines	Post speech	Post rest	Post sham	Follow-up

perception, and memory. However, participants with speech disorders would be aware that the sham program did not target communicative abilities.
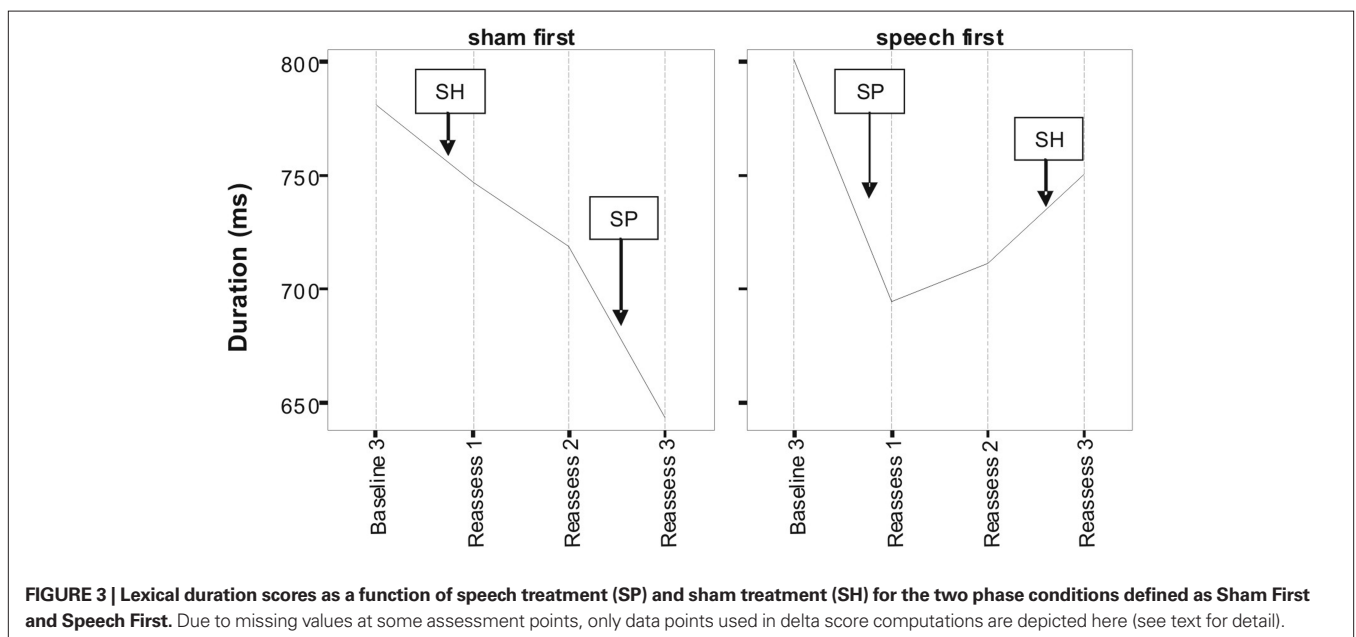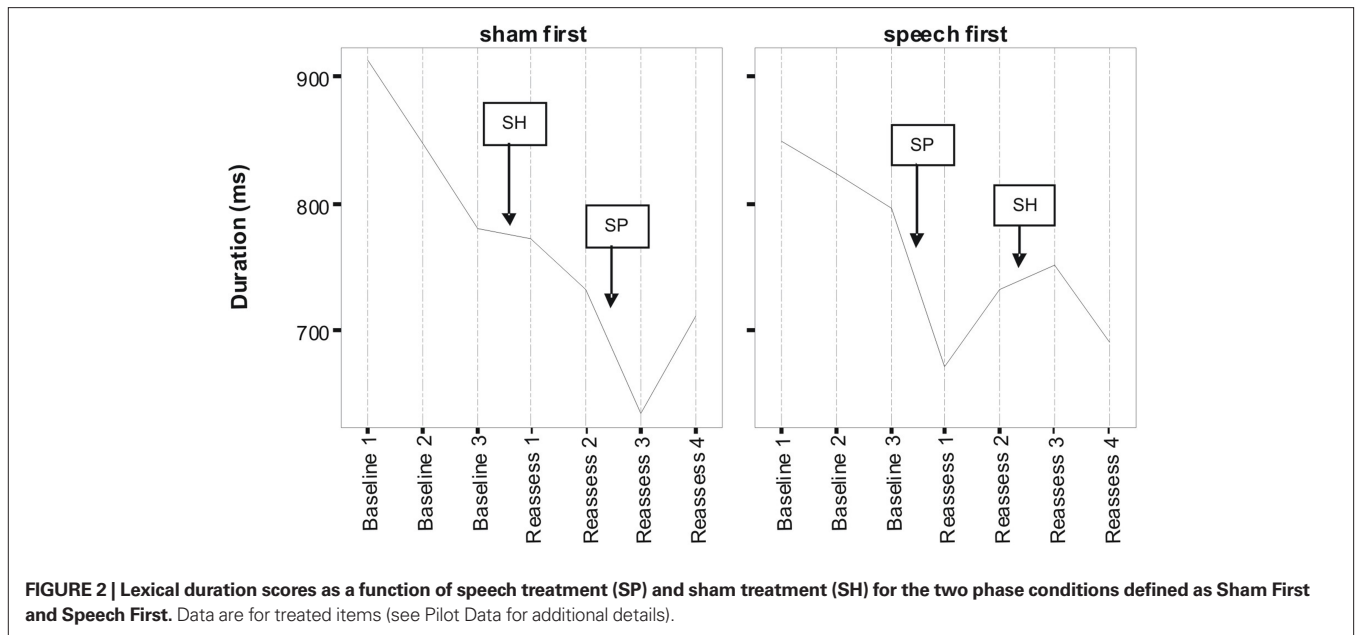
Participants were instructed to use the programs for a period of 30–40 min at least once per day. Both programs recorded the interactions the participant had with the software, and the degree of compliance with the instruction could subsequently be determined. In research with larger samples of patients, the program usage data can be used to explore the relationship between treatment dose and outcome. Self-administered therapy has advantages, such as therapy dose is no longer dependent upon clinician availability. However, the amount of treatment is dependent upon the compliance of the participant with the dose recommendation. As a result, intervention research using this model of service delivery is conducted under "field conditions" rather than laboratory conditions in which dose levels can be tightly controlled. In this way, there is some blurring of the boundary between treatment efficacy research (evaluation of the intervention under ideal conditions) and effectiveness research (the intervention under realistic, "field" conditions). Reducing the translational gap between laboratory and clinic could be seen as advantageous, allowing clinicians to make valid inferences concerning the effectiveness of a form of therapy under usual conditions.

The dependent measures of interest for this article consist of parallel outcomes for the same sets of spoken words which were recorded at each time point for scoring and analysis. Words were elicited through a verbal repetition task. Spoken production of word items were elicited at each assessment and scored in two ways. The first scoring method applied an ordinal scaling of phonetic accuracy which ranged from a possible 0–7 rating (0 = lowest accuracy to 7 = highest accuracy). The second scoring method involved acoustic analysis of the speech waveforms of the lexical items for precise duration measures (ms). Duration measurements were used to assess the speed and fluency of the lexical items which were on target or near to target, in order to allow the paired

comparisons of phonetically similar forms. For each lexical item, the duration was measured from its acoustic onset to its acoustic offset using spectrographic analysis. Two raters performed the speech analysis. Rater 1 was not blind to case, or phase at which the speech sample was recorded. Rater 2 analyzed 10% of all samples and was blind to case, point of data collection (e.g., baseline, pre-, post-speech/sham intervention) and the measurements made by Rater 1. Inter-rater correlations revealed high levels of agreement (accuracy Spearman's $r = 0.892$, $p < 0.001$; word duration Pearson's $r = 0.956$, <0.001).

Figures 1 and 2 respectively, show changes in mean accuracy and duration for words produced over the course of the study by the three participants in phase 1 (Sham First) and the three participants in phase 2 (Speech First). In the case of accuracy scores, increasing values indicate that speech intelligibility had improved and listeners were more likely to accurately understand the word. With regard to the durational measure, reduced duration suggests greater cohesiveness and fluency in articulatory movement. In **Figure 1**, all data were available for the accuracy measures. This is because the ordinal scale allows a score (i.e., zero) even in cases where the participant produces no speech response. In contrast, for the duration data, if there is no utterance or a response that is unrelated or distant from the target, it is not feasible to make an acoustic measurement of duration. Hence, there are missing data in the duration data sets especially for more severe patients. For this reason, the duration data are plotted twice. **Figure 2** displays means derived for all available data points. **Figure 3** shows the same data set as presented in **Figure 2**: but (a) reduced by the cases where computations of paired time points involved one missing value; and as a result (b) only showing the study time points used in the direct evaluation of treatment and sham effects. This issue is discussed in more detail below in relation to the computation of delta scores (see Results) and specific effects of missing data on the repeated measures components of cross-over designs (see Discussion).

FIGURE 1 | Lexical accuracy scores as a function of speech treatment (SP) and sham treatment (SH) for the two phase conditions defined as Sham First and Speech First. Data are for treated items. The full scale of accuracy scores ranges 0–7 (see Pilot Data for additional details).

FIGURE 2 | Lexical duration scores as a function of speech treatment (SP) and sham treatment (SH) for the two phase conditions defined as Sham First and Speech First. Data are for treated items (see Pilot Data for additional details).

FIGURE 3 | Lexical duration scores as a function of speech treatment (SP) and sham treatment (SH) for the two phase conditions defined as Sham First and Speech First. Due to missing values at some assessment points, only data points used in delta score computations are depicted here (see text for detail).

Despite randomized allocation to phase conditions, accuracy scores across all three baseline assessments were lower in the Sham First group. However, the final baseline (B3) for word duration was not markedly different between the groups, but did show an approximately 20 ms advantage for participants in the Sham First phase. This was due to the factors outlined in the previous paragraph in relation to loss of utterances suitable for scoring duration in the more severely impaired patients. A complex pattern of similarities and differences in "baseline stability" was also present. Although lower in overall accuracy, the Sham First group had comparable levels of fluctuation among the three baseline measures to the Speech First group. In contrast for duration, the two groups showed a comparable range of measures during the experimental periods, but the Sham First demonstrated a more marked shift across the three baseline measurements. The effects above were probably due in part to the small sample size of this pilot which made it vulnerable to these types of phase allocation bias.

A striking commonality in the two sets of measures (**Figures 1 and 2**) is the relative impact of speech treatment in comparison to sham treatment. For both accuracy and duration, the net change from the start to the end of the speech treatment was greater than the net change from the start to the end of the sham intervention (statistical analysis of effects is covered in the Results section below). However, there were also subtle differences in the impact of speech and sham interventions as a function of the order in which they were delivered; these effects were observed for both accuracy and

duration. Such details expose important aspects of the cognitive neurorehabilitation process that are showcased particularly well in the framework provided by the multiple baseline cross-over design. Several such patterns of interest are described below.

For accuracy in the Sham First phase, the improvement in relation to speech treatment was greater than the change in relation to the sham intervention. However, improvement in relation to speech treatment in the Sham First phase was modest in comparison to the speech treatment effect in the Speech First phase. A similar trend was observed for the decreases in word duration (a sign of improved fluency) after speech treatment which were greater in the Speech First phase. There are two explanations for this disparity in the impact of treatment for the two phases. One may be that the patients in the Sham First phase were less able to make improvements due to the greater severity of word production deficits. A second reason could be related to participants' psychological experience of the two study periods. Patients are acutely aware of their communicative impairments and despite attempts to obscure the sham nature of the visuospatial program, will be able to differentiate the likely impact of the two computer programs on their impairment. The psychosocial effect of receiving a speech related experience early in the study progression may have benefits for the participants who were randomly assigned to that phase. Evidence to support this claim comes from the data on patients' interaction with the study software. The amount of total time spent actively using the speech treatment software differed among patients (range: 18–60 mean minutes per day) with greater use in the three participants in the Speech First condition (Speech First: 43.3 mean minutes per day; Sham First: 30.67 mean minutes per day). Total time using the sham software was lower than time spent on the speech treatment software for the Speech First group (20.67 mean minutes per day), but higher for the Sham First group (41 mean minutes per day). Both groups interacted for longer with the program they were exposed to in the first compared to the second study period. However, a clear differentiation in response occurred in relation to program use in the second study period where participants in the Sham first phase showed more interaction than participants in the Speech first phase. Thus, participants' motivation may have been affected by a combination of factors such as knowledge of program content and the order of program delivery. A further level of interaction between clinical and psychological response to the study protocol is also possible, such that the more severe a patient's symptoms, the more susceptible they are to temporal factors in treatment delivery.

For duration, the effect of speech treatment was also greater than the changes observed for the sham intervention for participants in both the Speech First and Sham First study phases. The groups had comparable baselines at the start of the first intervention period, but different endpoints with respect to the assessment directly following the speech treatment. At baseline 3, the Sham First group had an average duration which was ~20 ms shorter than that observed in the Speech First group. However, after speech treatment, the difference had widened to ~50 ms. Difference in the direction of sham intervention is also noteworthy and its impact on statistical analysis of treatment effects is discussed in the Results section. Although both accuracy and duration showed evidence of treatment related improvements, microanalysis reveals subtle but

important differences between the two measures. These phenomena emphasize the importance of selecting behavioral measurements which tap into different modes of change across the course of treatment for the same set of speech items.

In summary, there was evidence for similar patterns of change with accuracy and duration measures. For both, there were greater improvements after speech treatment than sham despite the fact that there were different patterns of baseline starting points across the measures for participants in the Speech First compared to Sham First phases. This highlights the need for a technique that allows comparable statistical analysis of treatment effects for measures such as accuracy (which are non-parametric; more closely linked to the whole word experience of participants and "non-blinded data collection researchers") and duration (which are parametric; more remote from patient whole word experience, and objectively/procedurally separable from "non-blinded data collection researcher" experience). Such a method would need to account for participant differences in multiple key baseline measures across the course of the study including the start of the interventions given in both period 1 and period 2.

METHODOLOGICAL DEVELOPMENT

In the traditional medical cross-over model, the data from the two phases of the design are evaluated prior to combining them for statistical analysis of treatment effects. One important consideration is the estimate of any residual carry-over effects from the treatment (or possibly sham intervention) that have the capacity to impact on the subsequent period of the study. This is done by evaluating the magnitude of lambda (λ; Eq. 1), a statistic which compares the difference in treatment effect of each study phase to the sham effect of the other study phase. Confirmation or rejection of the null hypothesis H_0: $\lambda = 0$ determines whether data can be pooled across the two phases of the design (Wang and Hung, 1997).

Equation 1: $\lambda = (\mu_{21} - \mu_{11}) - (\mu_{12} - \mu_{22})$ where μ_{ik} represents the measurement parameter for phase or condition i (1 = sham first; 2 = treatment first) and period k (1 = period 1; 2 = period 2).

This statistical approach to two period cross-over models is largely based on pharmacological studies where treatment effects are physiologically diminished or eradicated during a wash-out phase. The post wash-out baseline, in theory, is assumed to be equivalent to the initial untreated baseline in the sham-first sequence. However, in successful behavioral treatment studies that instate more permanent changes to the system, an assumption of equivalence between these two baselines is not appropriate. This is because performance levels at the start of the sham treatment are quite different in the sham-first condition compared to the treatment-first sequence. In the sham-first phase, the purpose of measurement after period 1 is to show that the sham condition does not radically alter the pre-treatment baseline for behaviors hypothesized to respond to treatment. This contrasts with the conceptual significance of the sham treatment in the speech-first phase, where the purpose of measurement after period 2 is to document that there has been no additional change subsequent to the initial therapeutic improvements following treatment (i.e., to measure maintenance effects). These relationships are depicted schematically in **Table 3**, where a key point of comparison is the post-rest measurement after period 1 in the treatment first study

Table 3 | Schematic data comparing a pharmacological treatment study (Pharm) to a neurocognitive treatment study (Cog). Both are two period cross-over designs with a Sham First (ShTr) and a Treatment First (TrSh) phase.

Study phase	Baseline	Post period 1	Post rest	Post period 2	Computation	21–11	12–22	λ
1 Pharm_ShTr	0	2	1	12	μ	8	9	−1
2 Pharm_TrSh	0	10	1	3				
1 Cog_ShTr	0	3	4	13	μ	7	−1	8
2 Cog_TrSh	0	10	11	14				
1 Cog_ShTr	0	3; $\Delta = -3$	4	13; $\Delta = -9$	Δ	−7	−6	−1
2 Cog_TrSh	0	10; $\Delta = -10$	11	14; $\Delta = -3$				

Data for the cognitive study are shown twice. In the middle rows, data show computation of effects based on the mean measurements after each period of interest. This is contrasted to the same data but with computation of effects based on difference (delta) measures of change between locally relevant "baselines." (Top row: 21 = phase 2, period 1; 11 = phase 1, period 1; 12 = phase 1, period 2; 22 = phase 2, period 2; Δ = delta; μ = estimate of population mean; λ = lambda).

phases. In the pharmacological example, the measurement drops back to baseline (from 10 to 1), but in the neurocognitive example, change instated in period 1 is maintained (from 10 to 11). This leads to a numerical anomaly in computation of change during the second period of the cognitive treatment study which seems at first glance to represent an increase from 0 at baseline to 14 after the sham intervention. However, there is a straightforward solution for this, as shown in the last two rows of **Table 3**. Clearly, in a cognitive study where neurorehabilitation is taking place, a sustained response for repeated measure of the target behaviors would be expected, perhaps with some minor fluctuations. One must make computations for analysis of any further change, or lack thereof, from the end rather than the start of period 1. This can be achieved through the use of delta scores.

To accurately model the continuous pattern of change for a given set of behaviors over the course of cognitive neurorehabilitation, a modified statistical approach is required when estimating both carry-over effects and treatment effects. As described above, the pilot data for computerized word learning in apraxia of speech showed similar rates of increase in performance after the speech treatment condition, and a similar lack of change for the sham condition, across the two study phases. In this context, one can still apply the formula for λ, but must replace the parameter μ with Δ, where Δ signifies the change in performance that occurs between the start and end of each key evaluation point (Eq. 2). The normalizing effect on the computations is shown in the lower two rows of **Table 3**.

Equation 2: $\lambda = (\Delta_{21} - \Delta_{11}) - (\Delta_{12} - \Delta_{22})$ where Δ_{ik} signifies the change in performance that occurs during each of the four key stages of the study defined by phase i (1 = sham first; 2 = treatment first) and period k (1 = period 1; 2 = period 2).

This standardizes the data across the two sham conditions for starting performance (i.e., intercept) and allows one to focus instead on the presence or absence of change (i.e., slope). In turn, it provides the appropriate conceptual and numeric basis for computing (1) estimates of carry-over effects; (2) adjustments for variations in starting baseline (e.g., due to natural variations in symptom severity across participants); and (3) treatment effects using the full power of the cross-over design. Moreover, in parametric analyses that involve general linear models, use of delta scores provides an additional means to probe for carry-over effects and assess their impact on study results through consideration of the Phase × Period

interaction term. This modified approach to the cross-over design represents a modeling innovation and is detailed in the statistical analysis of results from the pilot study data in the next section.

RESULTS

This section reports on the computation of delta scores for accuracy and then duration of treated words for the six participants described above. Computations follow those described by Senn (2000), but with the appropriate delta scores used as the index of change.

Table 4 (columns 3 and 4) shows data for accuracy of treated items as a function of phase (Sham First or Speech First) and period. Delta scores for period 1 were computed as the difference in accuracy between the final baseline (B3) and the end of the first period intervention (R1). Delta scores for period 2 were computed as the difference in accuracy between the end of the rest phase (R2) and the end of the second period of intervention (R3). Delta scores in the negative direction indicate improvement in accuracy. For a patient with low accuracy of zero before treatment and higher accuracy of 5 after treatment, delta would be −5.

Based on these data, $\lambda = -0.27$. This deviation from zero was due to the fact, as described above in relation to **Figure 1**, that the average sham effects were similar in both study phases (Phase 1 = −0.28; Phase 2 = −0.20), but the treatment effect was larger in the Speech First phase (Phase 1 = −0.53; Phase 2 = −0.88). In summary, the speech treatment effect, over and above the sham effect, was present in both periods. However, this effect was greater in period 1 (−0.60) than in period 2 (−0.33). The issue of how to evaluate the source of this difference and manage it in the course of data analysis is not straightforward (Putt and Chinchilli, 1999; Wang and Hung, 1997). In the current data set, given the equivalent sham effects, it is possible that the differences in treatment effects reflected inherent differences in severity of patients' clinical symptoms between the two study phases (see **Figure 1**). Also, a case by case evaluation of the data shows that for five out of six participants, delta scores for the speech treatment were greater than delta scores for the sham intervention (scores for participant "P6" were equivalent for both). Moreover, the magnitude of λ was equivalent to the sham effect. For these reasons, the full cross-over design was employed for the next stage of statistical analysis so that within-subject analysis could be used. This was deemed to be more appropriate than resorting to a between group comparison

of the two first study periods since it would allow the treatment results from participants at all levels of severity to be included in the analysis.

Table 4 (columns 5 and 6) displays the same data in a manner that elucidates the values in relation to the speech and sham treatments for both phases. Because accuracy ratings were derived using an ordinal scale, Wilcoxon signed ranks test was used to compare the delta scores from the speech treatment to those from the sham intervention. A significant difference was observed ($z = -2.21$, $p < 0.03$).

Table 5 (columns 3 and 4) shows data for word duration of treated items as a function of phase (Sham First or Speech First) and period. As with accuracy, delta scores for period 1 were computed as the difference in duration between the final baseline (B3) and the end of the first period intervention (R1). Likewise, delta scores for period 2 were computed as the difference in duration between the end of the rest phase (R2) and the end of the second period intervention (R3). Delta scores in the positive direction indicate improvement in spoken word fluency. For a patient with long duration of 900 ms at baseline and shorter duration of 807 ms after treatment, delta would be 93 ms.

Based on these data, λ = −57.42. This deviation from zero was due to the fact, as described above in relation to **Figure 2**, that the sham effect in the Speech First phase was of comparable order of magnitude to that in the Sham First sequence, but was of the opposite sign (i.e., −39.20 signifies an increase in duration rather than a decrease). In summary, the speech treatment effect, over and above the sham effect, was present in both periods. However, this effect was greater in period 2 (118.34) than in period 1 (60.92). The matter of how to adjust for the possible differences in the two phases of the design can be handled as part of the statistical analysis in this case.

Table 5 (columns 5 and 6) displays the duration data in a manner that elucidates the values in relation to the speech and sham treatments for both phases. ANOVA was based on unweighted mean delta scores from the six participants. Phase (Sham First/Speech First) was the between group factor and Intervention (Sham intervention/Speech treatment) was the repeated measure. The Phase × Intervention effect partitions the variance due to any differences in the degree to which change in each phase is affected by the two interventions; also, the significance of this effect can be determined. Results showed a marginally significant effect of Intervention ($F(1,4) = 6.11$, $p = 0.069$) but no significant effects of Phase or Intervention × Phase. Together, the presence of an Intervention effect and absence of Intervention × Phase interaction indicate that the speech treatment led to decreases in duration for both the Speech First and Sham First groups.

DISCUSSION

A primary advantage of the multiperiod, multiphase cross-over design in cognitive neurorehabilitative research is exemplified in the current study of computerized treatment for apraxia of speech. Specifically, it takes account of the fact that patients cannot usually be blinded to the content of their treatment in relation to their disorder. In the current study, for example, it was not possible to disguise the speech based content of the treatment or the non-speech based content of the sham software. Provided no harmful effects of treatment are detected in the first study period, the research includes built-in access to the potential benefits of treatment to participants who initially received the sham intervention. Moreover, the delivery of both interventions within all patients allows researchers to "subtract out" the effects of the treatment compared to the sham using each participant as his or her own control. The current analyses showed that the design can be used not only to document the significance of treatment effects, but can be used in a detailed temporal analysis to investigate the treatment process as it unfolds over the course of the study. The data from the current study suggest that there is a tendency for participants to spend more time on the intervention delivered in the first period of the study, and possibly also, a subtle interaction between severity and temporal sequencing of treatment delivery.

Table 4 | Accuracy delta scores presented by Phase and for each participant (P1–P6).

Phase	Participant	Period 1	Period 2	Speech	Sham
Sham first	P4	0.00	−0.37	−0.37	0.00
	P5	0.03	−0.33	−0.33	0.03
	P6	−0.87	−0.90	−0.90	−0.87
	Mean	−0.28	−0.53	−0.53	−0.28
Speech first	P1	−1.23	−0.33	−1.23	−0.33
	P2	−0.30	−0.07	−0.30	−0.07
	P3	−1.10	−0.20	−1.10	−0.20
	Mean	−0.88	−0.20	−0.88	−0.20

Data in columns 3 and 4 are presented by study period. The same data are presented in columns 5 and 6 as a function of Speech versus Sham treatment. Data are for treated items.

Table 5 | Duration delta scores presented by Phase and for each participant (P1–P6).

Phase	Participant	Period 1	Period 2	Speech	Sham
Sham first	P4	140.14	78.67	78.67	140.14
	P5	−14.38	68.75	68.75	−14.38
	P6	45.00	90.00	90.00	45.00
	Unweighted mean	56.92	79.14	79.14	56.92
	Weighted mean	34.08	75.32	75.32	34.08
Speech first	P1	96.68	−28.54	96.68	−28.54
	P2	40.08	−35.96	40.08	−35.96
	P3	216.76	−53.12	216.76	−53.12
	Unweighted mean	117.84	−39.20	117.84	−39.20
	Weighted mean	106.47	−39.30	106.47	−39.30

*Data in columns 3 and 4 are presented by study period. The same data are presented in columns 5 and 6 as a function of Speech versus Sham treatment. Data are for treated items. Due to different patterns of missing data in terms of participants and periods, both weighted (as shown in **Figure 3**) and unweighted (as used in statistical analyses) means are shown.*

The use of delta, or "change scores," in conjunction with cross-over designs is infrequently reported in the rehabilitation literature (e.g., Ploughman et al., 2008), particularly in relation to the treatment of acquired communication disorders. The current paper demonstrated that delta scores provide a sensitive means by which to track the ongoing treatment process in relation to the changing baseline and potentially permanent neurocognitive changes over the course of the study. By normalizing indices of change in relation to local baselines, one can focus on the net change at each critical juncture in the design. These values are useful in statistical comparisons of treatment compared to sham interventions, and provide data for computing effect sizes and power estimates for future experiments. For the current study in apraxia of speech, delta scores revealed relatively large effect sizes for both first period and full cross-over repeated measures. However, for research into measures which yield small to medium effect sizes, the full cross-over design affords additional power provided by correlations across the repeated measures (i.e., periods) of the study. When effect sizes are large and repeated measures comparisons can be fully powered, even with smaller sample sizes, cross-over designs offer potential benefits not only in terms of research outcomes, but also for the resourcing of clinical research activity.

We have shown that with a few key computational adjustments, cross-over designs can be used to full effect. However, the design also has potential disadvantages which must be fully considered. These include the dependence on repeated measures such that participants must make a relatively long term commitment to remain in the study. This could present particular challenges for study populations who may be prone to withdraw part way through the study by virtue of change in their medical or personal circumstances. Other effects that need to be considered include the different experiences of participants randomly assigned to the two study phases as a function of whether they receive a treatment or sham intervention first. The intervention delivered during the first period could affect participant perceptions about its importance (i.e., the first intervention may be viewed as the more important one), and possibly also their motivations about participation (i.e., receiving treatment first may stimulate more software usage). Thus, in any study design that includes a multiperiod repeated measures component, researchers must be alert to the potential for treatment order effects (see also Raymer et al., 2010).

Another difficulty is the management of carry-over effects, or any form of order effect in the comparisons between study phases. Detection of factors that contribute to the non-zero lambda is one challenge for researchers. Removal of the effects is considered a fairly controversial statistical issue as it may lead to over-corrections or distortions to the data (Wang and Hung, 1997; Senn, 2000). Thus, one must ensure that studies are designed with sufficiently large sample sizes to provide adequate statistical power for the first period between group comparison in the event that large carry-over effects or a high number of missing values in period 2 of the study preclude using repeated measures. Another recommendation is to apply multiple dependent measures, and where theoretically possible, include at least one set of parametric measures that allow for the full factorial investigation of period × phase interactions.

Delta scores proved useful in adjusting the data for differences in baseline performance in order to evaluate the specific effects of speech treatment and sham intervention in the current pilot study.

However, there remain important questions in communication therapy research as to the importance of baseline performance as a prognostic indicator of therapeutic response. In a larger scale study where random allocation should insure that baseline differences are evenly distributed between phases, it is possible to statistically measure and adjust for potential correlations between baseline indicators (both in absolute level and degree of stability) and patient responses. An advantage of cross-over designs is that potential baseline effects would be assessed and addressed for *both* treatment and sham interventions in all participants.

The use of delta scores will require systematic evaluation in larger scale studies. Prior work in cognitive neurorehabilitation has noted normal distribution of delta scores (Ploughman et al., 2008). However, the statistical properties of delta scores, in relation to word accuracy and duration scores, require further evaluation. Larger scale studies will allow investigation as to the impact of the particular features of individual patient groups and treatment response on delta computations and lambda estimates. For example, variation in patient dosage via their use of computer based interventions (both treatment and sham programs) may impact speech performance. If so, the statistical properties of residual delta scores from which dose effects have been removed will also need to be studied. Additional work is also necessary to determine whether it is more appropriate to use weighted or unweighted means in relation to missing data in analyses based on composite delta scores. Analysis of a larger scale version of the pilot study in this paper, involving 50 participants with apraxia of speech, is currently underway and will be used to address the issues outlined above.

The current study examined the impact of a computerized treatment for AOS, an acquired disorder of communication for which comparable computerized therapies are not widely available. For this reason, and for the specific purpose of controlling for the effects of regular computer based interaction, a single treatment compared to a sham intervention was the optimal study design. In this context, the cross-over design offered clear benefits which are discussed above. However, it may be the case for other types of neurocognitive treatment that innovations in the field must be compared with previously established therapies. Indeed, this may be the case for the currently investigated computer based therapy at a point when therapies are developed that give rise to the need for multiple-intervention comparisons. Cross-over designs have been successfully applied in case series of patients with acquired communication disorders involving multiple treatments or the same behavioral treatment delivered under different pharmacological test conditions (Fillingham et al., 2005; Whiting et al., 2007; Raymer et al., 2010). Results of these studies indicate that by designing measures of change through carefully controlled, well-matched sets of test items, the cross-over design may be adapted for testing multiple treatments with larger samples of patients. For example, Raymer et al. (2010) compared therapies for acquired spelling impairments using 20 words trained with an errorful procedure and 20 psycholinguistically matched words trained with an errorless procedure. Although some patients reached ceiling performance on one set of treated items in earlier periods of the study, comparison to the alternative treatment in later study periods was accomplished through the use of different, but matched sets of test items. Similar approaches have been used in cross-over designs of

naming (Fillingham et al., 2005; Whiting et al., 2007) and more widely in the cognitive neurorehabilitation literature (e.g., Mount et al., 2007).

In a traditional cross-over design application such as the pharmaceutical context, positive clinical outcomes would be temporary, and expected to wash-out prior to subsequent interventions. However, as raised in the Introduction, attainment of excellent clinical outcomes presents cognitive neurorehabilitation research with a particular challenge given the semi-permanent nature of the desired effects. As emphasized in the previous paragraph, this is a key issue for studies where clinical progress is made in an early period of a study sequence, particularly if ceiling effects are attained. Indeed, in the current study, patients in the Speech First condition showed near-ceiling effects on accuracy measures after the first period of the study. This raises the question of how one interprets lack of change in the subsequent study period after patients have used the sham computer program. This is an issue of particular interest in studies where the same test items (lexical items in the current study) are compared after both study periods. On the one hand, the treatment shows evidence of positive outcomes, and the effect is maintained through the course of the second period when the patients interact with the sham intervention. However, a critical analysis of the data raises the possibility that the ceiling effect of treatment precluded the possibility of further sham-related improvements, thus reducing claims of comparable sham effects from the speech-first and sham-first study phases. For this reason, researchers interested in implementation of similar designs are urged to consider multiple means of measuring behavioral change. In the current study we examined production for each spoken item in terms of both scaled accuracy and acoustically measured duration. In contrast to the 0–7 ratings used for lexical accuracy, duration measures are not subject to the constraints of a defined ceiling effect and therefore allowed us to confirm the comparability of sham-based changes across study phases for word duration. In the context of the research on which the AOS therapy was based (Whiteside and Varley, 1998; Varley and Whiteside, 2001), the study of word duration was of primary theoretical interest. Moreover, it serves as a dependent measure robust to the potential problems of ceiling effects in multi-period comparisons.

Ultimately, there is a complex set of trade-offs to consider when selecting a study design for cognitive rehabilitation research. For example, parallel studies typically require more participants than cross-over studies in order to attain equivalent statistical power. However, cross-over studies typically need more time points to include multiple intervention periods for each participant. Both sets of limitations present the clinical researcher with dilemmas to overcome. Cross-over study designs offer a combination of benefits which until now have been associated mainly with patient case series experiments. These include the ability to track responses to multiple interventions and to document events that correlate with periods of behavioral stability and change – all *within* a given patient. In addition, with cross-over designs, the potential for benefits of a particular treatment to be delivered to all study participants is "built in" to the experimental process. This may have benefits for participant morale and motivation. In contrast, there may be elements of resentful demoralization (Pocock, 1983) for participants in parallel studies where knowledge of receiving a control or treatment intervention is difficult to hide, and where pre-study documentation will typically mention that allocation is to only one of the two groups. Bias in participant motivation poses specific problems for research into computer based treatments or any intervention that relies on independent participant co-operation outside the clinic or laboratory. The factors discussed above would ideally be tested through small scale pilot work. If early pilot work indicated that the logistics of large scale longitudinal research were experimentally infeasible or financially prohibitive, then cross-over designs would not be the most suitable choice. If however, the particular subfield of research was theoretically poised to support such work and resourcing was available, then the benefits of conducting research with fully powered multi-phase multi-period cross-over designs would be worth serious consideration.

In conclusion, the multiphase, multiperiod cross-over design is ideally suited for cognitive neurorehabilitation research. It provides the basis for microanalysis of change sequences for individual participants, and participant groups, as a function of experimental and control interventions. The elegance of the repeated measure design is that it allows within-subject comparisons of multiple cognitive interventions. When used in conjunction with treatment-sham delivery procedures that are based on valid neurofunctional dissociations and delta score computations, the cross-over design functions in a manner which is conceptually analogous to task-based comparison techniques used in functional neuroimaging where brain activation is compared across task or stimulus types with counterbalanced order of delivery. Thus, well designed behavioral studies in cognitive neurorehabilitation research have the potential to be adapted as the basis for, and to provide convergent evidence with, parallel cognitive neuroactivational experiments. From this, a better integrated cognitive neuroscience for the study of treatment and recovery in acquired communication disorders could be achieved.

REFERENCES

Borras, L., Boucherie, M., Lecomte, T., Perroud, N., and Huguelet, Ph. (2009). Increasing self-esteem: efficacy of a group intervention for individuals with severe mental disorders. *Eur. Psychiatry* 24, 307–316.

Carlomagno, S., Pandolfi, M., Labruna, L., Colombo, A., and Razzano, C. (2001). Recovery from moderate aphasia in the first year poststroke: effect of type of therapy. *Arch. Phys. Med. Rehabil.* 82, 1073–1080.

Dyson, L., Roper, A. H., Inglis, A. L., Cowell, P. E., Whiteside, S. P., and Varley, R. A. (2009). Self-administered therapy for word production impairments in aphasia and apraxia. *Paper presented at the British Aphasiology Society*, Sheffield, UK, September 2009.

Fillingham, J. K., Sage, K., and Lambon Ralph, M. A. (2005). Treatment of anomia using errorless versus errorful learning: are frontal executive skills and feedback important? *Int. J. Lang. Commun. Disord.* 40, 505–523.

Glogowska, M., Roulstone, S., Enderby, P., and Peters, T. (2000). Randomised controlled trial of community based

speech and language therapy in pre-school children. *BMJ* 321, 923.

Howard, D. (1986). Beyond randomised controlled trials: the case for effective case studies of the effects of treatment in aphasia. *Br. J. Disord. Commun.* 21, 89–103.

Kent, R. D., and Rosenbek, J. C. (1983). Acoustic patterns of apraxia of speech. *J. Speech Hear. Res.* 26, 231–249.

Leech, G., Rayson, P., and Wilson, A. (2001). *Word Frequencies in Written and Spoken English.* Harlow: Longman.

Lincoln, N. B., McGuirk, E., Mulley, G. P., Lendrum, W., Jones, A. C., and Mitchell, J. R. A. (1984). Effectiveness of speech therapy for aphasic stroke patients: a randomized controlled trial. *Lancet* 1, 1197–1200.

Mount, J., Pierce, S. R., Parker, J., DiEgidio, R., Woessner, R., and Spiegel, L. (2007). Trial and error versus errorless learning of functional skills in patients with acute stroke. *NeuroRehabilitation* 22, 123–132.

Page, S. J., Levine, P., and Leonard, A. C. (2005). Effects of mental practice on affected limb use and function in chronic stroke. *Arch. Phys. Med. Rehabil.* 86, 399–402.

Petersen, B., Mortensen, M. V., Gjedde, A., and Vuust, P. (2009). Reestablishing speech understanding through musical ear training after cochlear implantation. *Ann. N. Y. Acad. Sci.* 1169, 437–440.

Ploughman, M., McCarthy, J., Bosse, M., Sullivan, H. J., and Corbett, D. (2008). Does treadmill exercise improve performance of cognitive or upper-extremity tasks in people with chronic stroke? A randomized cross-over trial. *Arch. Phys. Med. Rehabil.* 89, 2041–2047.

Pocock, S. J. (1983). *Clinical Trials: A Practical Approach.* Chichester: Wiley.

Pulvermüller, F., and Berthier, M. L. (2008). Aphasia therapy on a neuroscience basis. *Aphasiology* 22, 563–599.

Putt, M., and Chinchilli, V. M. (1999). A mixed effects model for the analysis of repeated measures cross-over studies. *Stat. Med.* 18, 3037–3058.

Raymer, A., Strobel, J., Prokup, T., Thomason, B., and Reff, K.-L. (2010). Errorless versus errorful training of spelling in individuals with acquired dysgraphia. *Neuropsychol. Rehabil.* 20, 1–15.

Robertson, I. H., Tegnér, R., Tham, K., Lo, A., and Nimmo-Smith, I. (1995). Sustained attention training for unilateral neglect: theoretical and rehabilitative implications. *J. Clin. Exp. Neuropsychol.* 17, 416–430.

Senn, S. (2000). *Cross-Over Trials in Clinical Research.* Chichester: John Wiley & Sons.

Varley, R. (2010). "Apraxia of speech: From psycholinguistic theory to the conceptualization and management of an impairment," in *The Handbook of Psycholinguistic and Cognitive Processes: Perspectives in Communication Disorder*, eds J. Guendouzi, F. Loncke, and M. J. Williams (New York: Psychology Press), 535–550.

Varley, R. A., Cowell, P. E., Dyson, L., Roper, A. H., Inglis, A. L., and Whiteside, S. P. (2009). Lexical therapy for apraxia of speech. *Paper Presented at The Science of Aphasia.* Antalya, Turkey, September 2009.

Varley, R. A., and Whiteside, S. P. (2001). What is the underlying impairment in acquired apraxia of speech? *Aphasiology* 15, 39–49.

Wang, S. J., and Hung, H. M. J. (1997). Use-of two-stage test statistic in the two-period cross-over trials. *Biometrics* 53, 1081–1091.

Whiteside, S. P., and Varley, R. A. (1998). A reconceptualisation of apraxia of speech: a synthesis of evidence. *Cortex* 34, 221–231.

Whiting, E., Chenery, H. J., Chalk, J., and Copland, D. A. (2007). Dexamphetamine boosts naming treatment effects in chronic aphasia. *J. Int. Neuropsychol. Soc.* 13, 972–979.

Wilson, S. M., Saygin, A. P., Sereno, M. I., and Iacoboni, M. (2004). Listening to speech activates motor areas involved in speech production. *Nat. Neurosci.* 7, 701–702.