Treatment Integrity of Elaborated Semantic Feature Analysis Aphasia Therapy Delivered in Individual and Group Settings

R2

Vasiliki Kladouchou[1], Ilias Papathanasiou[2], Eva A. Efstratiadou[1], Vasiliki Christaki[3], Katerina Hilari[1]

[1] Division of Language and Communication Science, School of Health Sciences, City, University of London

[2] Dept. of Speech and Language Therapy, Technological Educational Institute of Western Greece

[3] Private Practice, Athens, Greece

Corresponding author

Professor Katerina Hilari
Division of Language and Communication Science
School of Health Sciences
City, University of London
Northampton Square
London EC1V 0HB
UK
k.hilari@city.ac.uk
Tel: +44 (0) 207 040 4660

Declaration of Interest

This study evaluated the Speech and Language Therapy treatment delivered within the Thales Aphasia Project. The Thales Aphasia Project was co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: THALES – UOA - "Levels of impairment in Greek aphasia: Relationship with processing deficits, brain region, and therapeutic implications", Principal Investigator: Spyridoula Varlokosta.

Abstract

Aims: This study ran within the framework of the Thales aphasia project that investigated the efficacy of Elaborated Semantic Feature Analysis (ESFA). We evaluated the treatment integrity (TI) of ESFA, i.e. the degree to which therapists implemented treatment as intended by the treatment protocol, in two different formats: individual and group therapy.

Methods & Procedures: Based on the ESFA manual, observation of therapy videos and TI literature, we developed two ESFA integrity checklists, for individual and group therapy, and used them to rate 15 videos of therapy sessions, delivered by three speech-language therapists (SLTs). Thirteen PwA were involved in this study. Reliability of the checklists was checked, using Kappa statistics. Each session's TI was calculated. Differences in TI scores between the two therapy approaches were calculated, using independent sample t-tests. Treating SLTs' views on what facilitates TI were also explored through a survey.

Outcomes & Results: Inter- and intra-rater reliability were excellent ($.75 \leq \kappa \leq 1.00$) for all but one video ($\kappa = .63$). Overall, a high TI level (91.4%) was achieved. Although both approaches' TI was high, TI for individual therapy sessions was significantly higher than for group sessions (94.6% and 86.7% respectively), $t(13) = 2.68$, $p = .019$. SLTs found training, use of the treatment manual, supervision, and peer support useful in implementing ESFA therapy accurately.

Conclusions & Implications: ESFA therapy as delivered in Thales is well described and therapists can implement it as intended. The high TI scores found enhance the internal validity of the main research project and facilitate its replication. The need for more emphasis on the methodological quality of TI studies is discussed.

Key Words: Elaborated Semantic Feature Analysis, Aphasia, Treatment Integrity, Treatment

Fidelity, Thales Aphasia Project

What this paper adds to existing knowledge

What is already known on this subject

- **Treatment Integrity (TI) is the extent to which core components of a treatment are implemented in clinical testing as intended by treatment protocols. TI data facilitate the implementation of evidence-based practice by allowing researchers to come to valid conclusions on the effectiveness of different treatments.**
- **Despite its importance, TI is infrequently reported: in a review of aphasia therapy studies (n=149), only 14% reported on some aspect of TI.**

What this study adds

- **This study provides evidence on the TI of an aphasia therapy: Elaborated Semantic Feature Analysis (ESFA). It shows that ESFA as delivered in this project was well described and therapists could effectively follow the manual and deliver the therapy as intended (TI level = 91.4%)**

Clinical Implications

- **Offering training, providing clinicians with a treatment manual and ongoing supervision and peer support, can help them deliver ESFA aphasia therapy as intended in order to improve the word finding difficulties of people with aphasia.**
- **The integrity checklists developed for this study can help clinicians monitor how closely they follow the treatment protocol.**

Treatment Integrity of Elaborated Semantic Feature Analysis Aphasia Therapy Delivered Individual and Group Settings

When testing the efficacy of a treatment, like Elaborated Semantic Feature Analysis (ESFA) (Papathanasiou and Mihou 2006), it is important to ensure that the treatment is delivered by therapists as planned. Treatment fidelity (TF), which includes treatment integrity (TI), refers to the methodological strategies used to monitor and enhance the reliability and validity of an intervention (Bellg et al. 2004).

The present study focuses on delivery of treatment and TI in particular. Although different terms have been used in the literature to describe TI, including procedural reliability, implementation fidelity and treatment fidelity, the term TI will be used consistently here, defined as the extent to which core components of a treatment are implemented in clinical testing as intended by treatment protocols (Yeaton and Sechrest 1981, Dusenbury et al. 2003, McIntyre et al. 2007), in other words, therapists' adherence to the treatment protocol.

Treatment integrity is a key feature of TF, and refers to whether the treatment was delivered as intended (Kazdin 1986). TI as well as *treatment differentiation*, which refers to whether the treatment conditions differed from one another in the intended manner (Kazdin 1986), focus on delivery of treatment aspects and were the first two concepts used in the literature by Moncher and Prinz (1991) to define TF. The concept of TF was expanded to include *treatment receipt,* which involves checking that the participant understands and can use treatment skills, and *treatment enactment,* which includes optimising the degree to which the participant is using skills learned in treatment in daily life (Lichstein et al. 1994). Further, in the Treatment Fidelity Workgroup of the National Institutes of Health Behavior Change Consortium (BCC) two more concepts were purposed: *study design,* i.e. the establishment of

procedures that ensure that a study can adequately test its hypotheses, and *training provider*, which involves procedures that standardise training of therapists (Bellg et al. 2004).

Treatment integrity has received attention in the literature because it has important implications. Firstly, TI is necessary to maintain internal validity (Moncher and Prinz 1991). It plays a key role in the interpretation of treatment results, as it allows researchers to establish whether the results of a study are attributable to the planned treatment or to the treatment that was actually implemented (Linnan and Steckler 2002, Perepletchikova and Kazdin 2005).

Treatment integrity also promotes external validity in terms of intervention replication and therefore comparisons across studies (Moncher and Prinz 1991). Treatments that can be measured for adherence to the protocol are likely to be sufficiently well described to be replicated (Mowbray et al. 2003, Hinckley and Douglas 2013). Despite the broad understanding of the importance for a study to be replicable, many studies do not meet the criteria for replication. In a literature review aiming to describe the reporting of TI data among aphasia treatment studies from 2002 to 2011, Hinckley and Douglas (2013) reported that only half of studies provided sufficient treatment description to allow replication.

The issue of TI pertains also to evidence-based practices (EBP). A critical bridge between the accumulated evidence for a treatment and its implementation in clinical practice is the understanding of its core components, which typically begins with the establishment of integrity and the measure with which it has been assessed (Fixsen et al. 2005). Moreover, without TI data, intervention effectiveness cannot be evaluated with accuracy (Lane et al. 2004). Researchers should use a therapy protocol for training and supervising clinicians, but also for checking programme quality and performance, ensuring fidelity of the trialed intervention (Mowbray et al. 2003).

Treatment Integrity in Aphasia Therapy Studies

Despite its importance, TI is not routinely reported in speech and language therapy studies of treatment effectiveness. In aphasia therapy, which is the focus of this study, recent reviews of the literature suggest that the measurement of TI is uncommon (Cherney et al. 2008, Cherney et al. 2013, Faroqi-Shah et al. 2010, Hinckley and Douglas 2013, Rose et al. 2013). Cherney et al. (2008), focusing on constraint-induced language therapy for individuals with aphasia, included 10 studies in their review; only two of them reported data on TI. In a systematic review of 14 studies on treatment effects for bilingual individuals with aphasia (Faroqi-Shah et al. 2010), only 14% of studies checked TI. Similar findings were reported in a review of the methodological quality of 23 studies on communication partner training for people with aphasia; only 13% of them included TI data, which led the authors to conclude that one widely failed criterion of methodological quality across studies was TI (Cherney et al. 2013). In their review of gesture treatments for people with aphasia (PwA), Rose et al. (2013) found that 22% of the included studies reported on TI.

One could argue that the above reviews included a relatively small number of studies in specific areas and thus the notion that TI data are lacking is exaggerated. However, recently Hinckley and Douglas (2013) published the first review on the importance of TI and the frequency with which it is reported in aphasia treatment studies. After reviewing 149 papers published between 2002 and 2011, they confirmed the results of the studies above: only 14% of studies stated clearly some aspect of TI.

Integrity Measures

In terms of methods that can be used to evaluate TI, both direct and indirect approaches exist. In direct integrity measures the researcher observes sessions, either video-recorded or live, and integrity is evaluated with the use of any sort of objective observational

measure (Kaderavek and Justice 2010, Schoenwald et al. 2011), such as checklists specifically developed or treatment protocols themselves. Indirect methods of integrity assessment, on the other hand, mostly include self-reports of therapists who are asked to indicate after sessions whether they included all the required components of the treatment; or self-reports of clients who are asked to report whether they received all of the components of the assigned treatment (Hinckley and Douglas 2013, Kaderavek and Justice 2010, Schoenwald et al. 2011).

In aphasia studies, the vast majority of researchers that incorporated TI measures have adopted direct methods. In particular, an independent rater checked a randomly selected sample of treatment sessions either live (Edmonds and Babb 2011, Edmonds et al. 2009, Kiran 2008, Kiran and Johnson 2008) or videotaped (Dietz et al. 2014a, Dietz et al. 2014b, Edmonds and Kiran 2006, Goff 2013, Heilemann et al. 2014, Hickey et al. 2004, Hinckley and Carr 2005, Kiran and Thompson 2003, Leonard et al. 2008, Wright et al. 2008). They used a list of core therapy components or the protocol itself to check whether each component of the treatment was implemented. To calculate adherence to the protocol, the number of components implemented by the therapists was divided by the total number of components planned (i.e. the components that would be rated for TI) and the result multiplied by 100 (Dietz et al. 2014a, Dietz et al. 2014b, Edmonds and Babb 2011, Hickey et al. 2004). An example of a study that describes in detail the procedures followed for checking TI is that of Hickey et al. (2004).

Yet many aphasia studies that checked TI directly do not specify clearly all strategies and methods followed, for example whether they used live or videotaped observation, how they calculated the percentage of TI, or what types of scales they used to check TI (e.g. present/absent or Likert-type) (Goff 2013, Griffith et al. 2014, Kiran et al. 2011, Rider et al.

2008, Rose and Douglas 2006, Rose et al. 2002, Rose and Sussmilch 2008, Schneider and Frens 2005, Wambaugh and Wright 2007). The lack of such information creates uncertainty regarding the quality of the procedures followed and the data generated.

In terms of indirect methods for measuring TI in aphasia therapy, studies have employed supervision of clinicians in conjunction with other methods, such as discussions about the treatment and its protocol as well as observations (Kempler and Goral 2011, Peach and Reuter 2010). In addition, training of providers (Goff 2013) and completion of questionnaires (Egan et al. 2004) or surveys (Heilemann et al. 2014) have also been used.

ESFA therapy is a modified version of Semantic Feature Analysis (SFA) therapy, as it is based on the SFA approach, but also prompts the individual, after word retrieval, to elaborate the features of the word elicited on the SFA chart into a sentence. It also includes provision of elaborate cueing hierarchies to elicit features when participants cannot produce them. Moreover, during ESFA therapy, participants are encouraged to write the features on the chart, as writing can be developed into a self-cueing strategy. When PwA are not able to write though, the therapist writes down the word for them. The additional purpose of this elaborated approach is to enable the individual to transfer their naming abilities to connected speech. Given that ESFA is a new therapy, no studies that check TI of ESFA were found. Thus, methods used in SFA treatments will be described below.

In the evaluation of TI of SFA treatments for PwA (Boyle and Coelho 1995, Coelho et al. 2000), both direct and indirect methods have been adopted. The study of Peach and Reuter (2010) is an example of using indirect methods. They examined the utility of SFA for improving verb and noun retrieval in aphasic discourse and reducing the frequency of word retrieval deficits in discourse. Their methods comprised review of the published principles for SFA therapy, discussion about them before treatment, and the presence of investigators in all

treatment sessions to ensure adherence to SFA guidelines during programme implementation. In studies using direct methods to explore the TI of SFA, adherence to protocol was measured by an independent observer viewing videotaped or live sessions, as described above (Rider et al. 2008). Rider et al. (2008) aiming to ensure that the examiner followed the therapy procedures appropriately, employed an individual trained in SFA to watch 10% of randomly selected trials from each participant's therapy, and found a TI score of 99.7%.

Although a combination of several indirect measures for checking TI make integrity data more robust, these methods have low correlations with objective measures and are less reliable (Gresham et al. 2000). Direct observation is considered the gold standard in the literature as it results in more thorough and objective data. Yet, this approach also has limitations, such as staff and time requirement as well as the fact that direct observation may not represent a "natural" implementation due to the treating therapist's awareness of observation (Cochrane and Laux 2008). Indirect data can be used to supplement objective data derived from direct methods (Heilemann et al. 2014, Hickey et al. 2004). This approach is supported by the BCC too (Bellg et al. 2004).

Research Aims

The evaluation of TI is an important part of the methodological quality of a treatment study. The present study ran within the framework of the Thales Aphasia project (http://thales-aphasia.phil.uoa.gr), which aimed among other factors to investigate the efficacy of Elaborated Semantic Feature Analysis (ESFA) therapy (Papathanasiou and Mihou, 2006), delivered through two different approaches: individual therapy vs. a combination of individual and group therapy.

We investigated the TI of the ESFA aphasia therapy in individual and group therapy sessions. We focused on programme adherence, by checking therapists' consistency in the

delivery of the therapy.

The research questions were:

i What is the degree of therapists' adherence to the ESFA protocol, in individual therapy sessions, group sessions and overall in all sessions?

ii Is there a significant difference in protocol adherence between individual and group therapy sessions?

In order to facilitate the interpretation of the findings and enhance our understanding of TI, an exploration of the therapists' views on different aspects of the therapy related to TI was additionally undertaken, via an e-mail survey.

Methods

Participants

Participants in this study were the three research speech and language therapists (SLTs) who were trained in ESFA and delivered the treatment in the Thales aphasia project. All three participants had a Master's degree and had worked with PwA from two to seven years. Their ages ranged from 28 to 38 years, with a mean (SD) age of 31.7 (5.5), their education was between 19 to 20 years [mean (SD)=19.3 (0.6)] and their clinical experience ranged from 2 to 10 years with a mean (SD) of 5.7 (4.0) years.

People with aphasia were recruited for the main Thales aphasia project from Neurologists and SLTs working in state hospitals and private rehabilitation centers in Athens, Greece. Thirteen PwA were involved in this study. They had to meet the following eligibility criteria: were > 18 years old and native Greek speakers; had aphasia due to a stroke, as reported by their referring clinician; were at least four months post stroke and medically

stable; had no history of other neurological or psychiatric problem and no considerable cognitive impairment according to their score (score ≥ 32 out of 38) on the Brief Cognitive Screening Test. Brief Cognitive Screening Test is a Greek test for cognition specifically targeted to PwA, based on items from the Dementia Rating Scale (Mattis 1988) and the Raven's Coloured Progressive Matrices (Raven 2004). People were excluded if they received other speech and language therapy services during the Thales project. Also PwA were excluded if they did not live independently at home prior to the stroke, to ensure they did not suffer from substantial comorbidities that could affect response to aphasia therapy.

Therapy videos of 13 PwA were used in this study. Aphasia was assessed with the Greek version of the Boston Diagnostic Aphasia Examination (Papathanasiou et al. 2008). Five participants had global aphasia, four Broca's aphasia, two anomic aphasia, one conduction aphasia and one transcortical motor aphasia. Of them, 61.5% were men (n=8) and the remaining 38.5% women (n=5). The participants' ages ranged from 40 to 79 years, with a mean (SD) age of 59.5 (12.1) years. Regarding their education, it ranged between 6 and 19 years [mean (SD)= 13.3 (3.8)]. In terms of their time post-stroke, PwA had a median (IQR) time post-stroke of 10 (7.0–67.5) months, with five of them being over 12 months post stroke.

Materials and Procedures

Data and sampling procedure. All participants gave their written informed consent to take part. Within the timeframe of 10 months leading to the data analysis of this study, each of the three SLTs had to provide three individual and two group therapy videos, recorded during the main research project. These videos had to meet the following criteria: the full therapy session had to be recorded, and both therapist and client(s) had to be clearly visible on the recording. The videos were recorded with a Panasonic VC-H110 video camera.

They were analysed from the beginning to the end, in order for all the important components of the ESFA therapy to be checked, for each session.

Therapy Procedure. The ESFA therapy, including the stimuli selection procedure, is fully reported according to the TIDieR guidelines (Hoffmann et al. 2014) in Appendix.

Therapy Overview. ESFA is a variant of SFA. SFA therapy aims to improve word retrieval, by focusing on strengthening the connections between the target word and its semantic network (Boyle 2004, Boyle and Coelho 1995, Coelho et al. 2000, Conley and Coelho 2003, Lowell et al. 1995). ESFA takes this approach a step further, aiming to enable the individual to transfer their naming abilities to connected speech. Like in SFA, during ESFA treatment, individuals with word retrieval difficulties are encouraged to generate words that are semantically related to the target word (i.e., semantic features), by completing a feature analysis chart. Unlike SFA, in ESFA, specific cueing hierarchies are employed to elicit features when participants cannot produce them (see integrity checklists in supplemental material). Moreover, participants are encouraged to write the features on the chart, as writing can be developed into a self-cueing strategy. The ESFA therapy also prompts the individual, after word retrieval, to elaborate the features of the word elicited on the SFA chart into a phrase and then a sentence (see individual and group therapy below for more info).

In the Thales aphasia project participants were randomised to receive either 36 hours of individual therapy (three one-hour sessions per week for 12 weeks) or 36 hours of a combination of individual and group therapy (two 45-minute individual therapy sessions and one 1½ hour group session per week for 12 weeks). The sessions took place mainly in the participants' home and some in hospital settings.

Individual therapy. The therapy process is detailed in Appendix. In summary, during

the therapy session, the client chose a picture from the stimuli set and the therapist asked them to name it. Then, presenting a semantic feature chart [same to that shown in Boyle (2004), but translated in Greek language], the therapist prompted the client to think of and say words related semantically with the target word (semantic features). The chart included 6 categories: *superordinate category*, *use*, *action*, *physical properties*, *location*, and *association*. To elicit features, the therapist asked questions or provided the client with sentence-completion cues, while prompting them to write down the features generated. If needed, the therapist used an alphabet table to help clients write; and if they were unable to write, the therapist filled in the chart.

After the chart completion and the retrieval of the word by the client, the therapist prompted the client to produce phrases with the target word and each of its features; and then to make a sentence of their choice with the target word and at least one of its features. There was no specific number of pictures to be worked on during each therapy session. The number of pictures worked on depended on the client's abilities.

Group Therapy. During the group therapy sessions the same principles and criteria as in the individual therapy were followed. The clients were asked in turn to answer the therapist's questions to find the target word, to complete the chart, to produce phrases with therapist's cues, and finally, to produce a sentence including the target word and at least one of its features. In addition, while during the initial therapy sessions the therapist provided phonological or semantic cues as needed, over time, the therapist gave participants the opportunity to interact and provide appropriate cues to each other. The therapist controlled turn taking to ensure individuals got similar amounts of exposure to targets and cues, whilst being mindful of disturbing peer-to-peer interactions as little as necessary.

Integrity checklists

In order to evaluate treatment integrity of the ESFA therapy, we developed two checklists (one for individual and one for group therapy) outlining the therapy process against which TI to be checked.

Development. The development of the ESFA integrity checklists was based on guidelines suggested by Stufflebeam (2000) and Stein et al. (2007). They were developed as a measure to be completed by an assessor, who was independent from the therapy process, but familiar with it (Heilemann et al. 2014). The checklists aimed to cover the critical therapist-oriented components of the intervention (therapists' strategies and responsibilities) in order to check their adherence to the protocol (Hogue et al. 2005).

The ESFA integrity checklists were developed by the first author who undertook two observations of live therapy sessions and had two meetings with the first author of the manual and trainer on ESFA, (EE), to ensure good understanding of the therapy and its important components. The construction of the checklists began with the creation of a list of potential items, after identification of the primary components of the therapy, by reviewing the ESFA protocol. As the active ingredients of ESFA therapy - components that are expected to create therapeutic change - are not known, each therapy component that could feasibly be checked through videos and was related to therapists' responsibilities was examined (Carroll et al. 2007). Then, the potential items for the checklists were put together according to the time point of the session that they should occur.

The initial set of components was assembled as a checklist (review version) and submitted to the manual developers and therapy experts for further review and critique, in terms of relevance and comprehensiveness of the content of items, as suggested in the literature (Netemeyer et al. 2003). In this way, content validity of the ESFA integrity checklists was established. Based on the experts' suggestions, the checklists' content was

revised, by adding, deleting or modifying the components on the list. Consensus was reached on the content and format of the checklists, through an iterative process of consultation between the developers of the ESFA therapy in this study and the authors. Two different checklists were developed this way, one for individual (ESFA integrity checklist) and one for group therapy sessions (ESFA integrity checklist-G) (Supplemental Materials [insert link]).

The ESFA integrity checklists were piloted by being applied to their intended use: the first author rated four ESFA therapy session videos that were not included in the data analysis of the present study. The ratings were discussed with the last author and further changes were made to formatting and the rating method used (see below).

The final version of the ESFA Integrity checklists included three main columns labeled: (1) components, where all the therapy aspects needed to be rated for TI are listed, (2) target word, where the name of the target word worked on would be indicated, so that ratings would take place for all the words targeted during the therapy session. The inclusion of all target words for analysis was considered crucial for TI as it would allow all therapists' behaviours to be captured, which differ according to clients' performance. Moreover, adherence could be affected by the time point during the session, e.g. therapists' fatigue at the end of the session could lead to lower TI results, and (3) comments, where notes on the nature of possible deviations and troubleshooting procedures or explanation of some ambiguous ratings could be made.

Rating method of the ESFA Integrity checklists. Both Likert-type scales (Clarke 1998, Heilemann et al. 2014) and scales that capture the presence or absence of a behavior (Hinckley and Carr 2005, Schneider and Frens 2005) have been used in the literature to check TI. As TI is perceived as the *degree* to which core components of a treatment are implemented as intended, and therapy components may be implemented but not fully, a

Likert-type scale was considered the most appropriate rating method. For instance, when a therapist has to follow a specific cueing hierarchy and they follow it but they omit some of its steps, it is more accurate to rate them 0.5 than 0. To this end, a three-point scale was used as the rating method for the ESFA integrity checklists, where the rater was asked to use one of the following ratings: *0* (not implemented as planned), *0.5* (partly implemented as planned) and *1* (fully implemented as planned), for each component of each target word. A component could also be marked as *NA* (not applicable). Further explanation of the rating system used, with some relevant examples, is given on the checklists.

Reliability of the integrity checklists. To check inter-rater reliability of the ESFA integrity checklists, an independent rater (R2) observed and rated a randomly selected sample of three of the nine individual therapy sessions (33%) and two of the six group sessions (33%), a total percentage (33%) that is within suggested guidelines (15- 40%) (Heilemann et al. 2014). Their ratings were then compared to those of the first author (R1), who rated all videos (n=15). For intra-rater reliability, a randomly selected sample of three of the nine individual therapy (33%) and two of the six group therapy sessions (33%) were re- rated by the R1 after an interval of three weeks. After the randomisation procedure, there was an overlap between the videos checked for inter- and intra- rater reliability; 2 out of 5 (40%) videos checked from each rater were common (videos 6 & 14).

SLT participants' views on ESFA therapy survey

To facilitate the interpretation of the findings a survey was developed (Supplemental Materials [insert link]), to explore the therapists' views on different therapy aspects which are related to TI. Using email was considered the most feasible way to conduct this survey given the small number of SLT participants (n=3), the fact that the first author was independent of the Thales team and the geographical distance between the first author and SLT participants.

The development of the survey's questions was based on the Implementation Fidelity Framework (Carroll et al. 2007). It aimed to explore some of the so-called *moderating factors* that may influence the degree of TI. The survey consisted of seven questions, which were categorised under three parts / possible moderating factors. The first part covered *facilitation strategies* used to support the implementation of the ESFA therapy programme. The second part elicited the therapists' views on the *ESFA manual*. As the checklists were based on the manual, therapists' views on manual properties could enhance conclusions about the face and content validity of the checklists. The third part was about *intervention complexity*, where therapists were asked to rate the complexity of ESFA therapy as low, moderate or high, based on given descriptors for the ratings. This question was added because complex interventions have greater scope for variation in their delivery (Carroll et al. 2007), and therefore some components may be more likely not to be implemented as they should.

Data analysis

For the calculation of inter- and intra-rater reliability of the integrity checklists, Kappa statistics were used (the Kappa coefficient of Cohen) (Cohen 1960). A Kappa coefficient of .75 - 1.00 is excellent, .60 - .74 is good, .40 -.59 is fair, and below .40 is poor (Cicchetti and Sparrow 1981; as cited in Cicchetti 1994, p. 286). These guidelines are in line with benchmarks that have suggested a level of 70% and above to be regarded as an acceptable level of agreement (Heilemann et al. 2014). For TI, the first authors' ratings were used in the analyses. The TI score for each session was calculated by summing up the ratings for all the components 'implemented' (components rated as 0.5 and 1) and dividing the results by the total number of the applicable components 'planned' (referred to as maximum score). All scores were converted to percentage scores for comparability. The TI scores for (a) individual therapy sessions (n=9), (b) group sessions (n=6) and (c) overall (n=15) were calculated by

summing up the scores for each session and dividing by the number of sessions (n). Different authors have considered different degrees of integrity as high (Carroll et al. 2007, Clarke 1998). However, because the level of TI that can be 'tolerated' in clinical implementation is not yet known (Kaderavek and Justice 2010), for the purposes of the present study the benchmarks suggested by Heilemann et al. (2014), which are based on a literature review, were adopted. Thus, a percentage of 80% and above was accepted as a high level of TI. Differences between individual vs group sessions on adherence percentage were analysed with an independent samples t-test, as data were normally distributed (Shapiro Wilks $p = .115$). All analyses were carried out on IBM SPSS v.22.

Results

Integrity Checklists' Reliability

Inter-rater reliability. Table 1 details the inter-rater reliability values separately for each of the five sessions, including TI scores given by the two raters. There was an excellent level of agreement between the two raters for four videos out of five, and a good level of agreement (.63) for the remaining video (video 3). The average Kappa was .82 ($p < .001$), indicating an excellent agreement between the two observers' ratings.

[table 1 about here]

Intra-rater reliability. Table 2 presents intra-rater reliability values separately for each of the five sessions, including TI scores given by rater 1 at two different time points. There was an excellent level of agreement between time 1 and time 2 (three weeks later) ratings, for all sessions. The average Kappa across the five sessions was .98 ($p < .001$), indicating excellent intra-rater reliability.

[table 2 about here]

Treatment Integrity (TI)

Fifteen videos of ESFA therapy sessions were rated using the ESFA integrity checklists, in order to examine the degree to which therapists adhered to the ESFA protocol (TI score), in individual sessions (n=9), group sessions (n=6), and overall in all sessions.

Treatment integrity for individual therapy approach. Table 3 details the number of components planned and implemented per session and the TI scores. The overall number of planned components across the sample of the nine individual therapy sessions was 450, with the number of components per session varying as the number of target words presented to each session was dependent on the clients' performance. The mean number of components planned per session was 50 (SD=16.5) with a range between 21 and 69. Concerning the components implemented by the therapists, they were 424 (out of 450) in total, ranging across sessions from 21 to 67, with a mean (SD) of 47.3 (15.8). In terms of the session-specific TI scores for the individual therapy approach they ranged 87% - 100%, with a mean TI score across all sessions of 94.6% (SD=4.6), showing a high level of TI.

[table 3 about here]

Treatment integrity for group therapy approach. Table 4 details the number of components planned and implemented per group therapy session and the TI scores. Across the six group therapy sessions, the overall number of planned components was 386, with the number of components per session varying as the number of target words presented to each session was dependent on the clients' performance. The mean number of components planned per session was 64.3 (SD=25.9), with a range between 28 and 98. Concerning the components implemented by the therapists, they were 334 in total, ranging across sessions from 25 to 77 [mean (SD) = 55.7 (22.3)]. In terms of the session-specific TI scores for the group therapy, they ranged from 77.2% to 92.6%, with a mean (SD) TI score of 86.7% (6.9).

This shows a high level of TI.

[table 4 about here]

Overall TI score. As can be seen from the results of the ratings of the individual therapy and group therapy sessions, the components planned for the whole sample of 15 sessions were 836, while the components implemented by the therapists were 758, representing an overall (SD) TI score of 91.4% (6.7), with scores ranging from 77.2% to 100%. Relating the TI scores to the cut-off value of 80%: 13 of the 15 videos had TI scores > 80%, with 10 of them > 90%; two sessions (videos: 10, 11) had TI scores below 80% (78.6% and 77.2% respectively).

In summary, therapists showed a high level of TI for individual therapy sessions (94.6%), for group therapy sessions (86.7%) and overall for all therapy sessions (91.4%).

Difference in treatment integrity between individual and group therapy sessions. The TI scores of all individual sessions (n=9) were compared with the TI scores of all group therapy sessions (n=6). The TI score for group therapy was significantly lower [mean (SD)=86.7% (6.9)] than the one for individual therapy [mean (SD) = 94.6% (4.6)], ($t(13)= 2.68$, $p=.019$).

SLT participants' views on ESFA therapy - Survey

Facilitation Strategies. The first part of the survey was related to facilitation strategies used to support the implementation of the ESFA therapy programme (see figure 1). Therapists indicated training, use of the treatment manual, supervision and support by developers, and peer support as useful strategies to facilitate an accurate implementation of the ESFA programme. One therapist also found role-playing useful. On average, they rated these strategies as being of a very good to excellent quality.

[figure 1 about here]

ESFA Manual. All respondents found the ESFA manual adequate in terms of its content and rated its properties, including ease of use, clarity and comprehensiveness, as very good to excellent (see figure 2). When therapists were asked if there were any therapy components included in the manual that the therapist should be more flexible on how to implement, rather than just following the manual, responses varied. While one of the SLTs believed that a therapist should be flexible with the manual in some cases, the other two indicated that the manual instructions should be followed without deviations. The former justified her opinion by stating that not all clients are able to strictly follow the manual's instructions and thus some therapy components should be adjusted to suit clients' strengths and weaknesses. All three participants rated their adherence to the treatment manual as high (4; on a scale 1-5).

[figure 2 about here]

Treatment complexity. One of the therapists (therapist 1) indicated that ESFA therapy has a high-level of complexity (all complexity dimensions applied), while therapists 2 and 3 found ESFA moderately complex (some of the complexity dimensions applied), as detailed in Table 5.

[table 5 about here]

Discussion

Integrity Checklists' Reliability

The majority of aphasia studies that adopted direct methods to examine TI do not provide evidence of inter-rater reliability of their instruments. Most did not check for

reliability between raters (Dietz et al. 2014a, Dietz et al. 2014b, Edmonds and Babb 2011, Edmonds and Kiran 2006, Edmonds et al. 2009, Goff 2013, Griffith et al. 2014, Hickey et al. 2004, Kiran and Thompson 2003, Kiran 2008, Kiran et al. 2011, Rider et al. 2008, Rose and Douglas 2006, Rose et al. 2002, Rose and Sussmilch 2008, Schneider and Frens 2005, Wambaugh and Wright 2007). Furthermore, no aphasia TI studies that checked intra-rater reliability were found; some researchers set it as a future goal though (Heilemann 2013). Given that the evaluation of TI is dependent on the psychometric soundness of the TI tool used, the lack of information about reliability or the use of inadequate methods for checking it (such as point-to-point agreement) creates uncertainty for the tools that have been used in some aphasia studies and in turn for the TI scores reported. This constitutes a gap in the TI literature that needs to be addressed.

Both TI checklists developed for this study had high inter- (κ= .82) and intra-rater (κ= .98) reliability, suggesting they are reliable measures for checking the therapists' adherence to the ESFA protocol and stable measures for TI evaluation. Other aphasia researchers who checked the reliability of their TI tools using statistical coefficients had similar findings. Heilemann (2013) for instance, who used a tool similar to this study to examine TI, tested inter-rater reliability with an intra-class correlation coefficient (ICC) and found an excellent level of agreement between the two raters for all but one session, where ICC was fair (ICC= .57). The small sample of videos included in the inter-rater investigation (n=3), however, should be kept in mind when interpreting these reliability results.

Other TI studies have reported point-to-point agreement as an inter-rater reliability measure (Yoder and Symons 2010) and found a high level of agreement too (96% and above) (Kiran and Johnson 2008, Leonard et al. 2008, Wright et al. 2008). These findings however should be interpreted with caution, as they are likely to be inflated due to the fact that

percentages of agreement do not correct for agreement expected by chance. This is why reporting percentages of agreement, without including statistical coefficients, has received criticism as a measure for inter-rater reliability (Hallgren 2012).

Some attention should be given to video 3 in the current study, which showed a lower level of agreement between rater 1 and 2, with a good but not excellent inter-rater reliability (κ= .63). A closer look on the ratings indicated a systematic pattern in the non-agreed components: the majority of differences between raters regarded the type of paraphasia produced by the client; rater 1 considered most of the client's paraphasias as circumlocutions, while rater 2 tended to consider them as semantic. The specific client was able to produce two- or three-word phrases/ structures, e.g., "turn-on, turn-off, button" for the target word 'light switch'. While for rater 1 such productions were an attempt for description of the target word (circumlocution), rater 2 considered them as semantic paraphasias mainly because of the brevity of the responses. This pattern implies that more specific rating instructions for such cases are probably needed.

Treatment Integrity

TI degree across the therapy sessions observed was high for individual (94.6%), group (86.7%), and overall all sessions (91.4%). This illustrates that the therapists implemented components of the ESFA therapy as intended by the treatment protocol with high integrity. These results were consistent with SLT participants' survey replies, as all of them indicated that they implemented the therapy with a high level of integrity (4/5).

Similar findings have been reported in other aphasia studies. In particular, the majority have reported a high TI score (92% and above) (Dietz et al. 2014a, Dietz et al. 2014b, Edmonds and Babb 2011, Edmonds et al. 2009, Griffith et al. 2014, Heilemann et al. 2014, Hickey et al. 2004, J. Hinckley and Carr 2005, Rider et al. 2008, Rose and Douglas

2006, Rose et al. 2002, Rose and Sussmilch 2008, Wambaugh and Wright 2007). When more than one rater was used for checking therapists' adherence, TI score was reported in the form of point-to-point agreement between the raters, and was high as well, varying between 96%-100% (Kiran and Johnson 2008, Leonard et al. 2008, Wright et al. 2008).

Facilitation strategies used to enhance therapy implementation have probably contributed to the high TI scores found in the present study. According to SLTs' replies to the survey, training, use of the treatment manual, supervision and support by developers and peer support were used to facilitate an accurate implementation of the ESFA programme. Such strategies have been found to optimise and standardise TI: 'the more that is done to help implementation, through monitoring, feedback, and training, the higher the potential level of implementation fidelity achieved' (Carroll et al. 2007, p. 7). In addition, the fact that therapists rated the manual properties, including ease of use, clarity and comprehensiveness, as very good to excellent, could have also optimised therapy implementation. Specificity enhances adherence and the comprehensiveness of a therapy's nature can influence how far the therapists successfully adhere to its prescribed components when implemented (Carroll et al. 2007).

Though TI checking for the ESFA therapy focused on therapists' behaviour, TI scores may also have been influenced by client's performance. Some of the treatment components were more prone to deviations (0.5 and 0 ratings) than others, and these appeared to be mostly components dependent on client's performance: a client with a less severe aphasia, for example, would produce fewer paraphasias, needing in turn fewer cues by the therapist. In such cases, the paraphasia components were coded as NA and thus there were fewer instances of 0.5 or 0 ratings.

Two of the videos (10 & 11) scored below 80% (78.6% & 77.2% respectively),

showing lower adherence. Both these videos were group therapy sessions, which by nature required more therapy components to be implemented by the therapists, making the treatment more complex and therefore more susceptible to variation in its application compared to the individual approach. Moreover, both these sessions were carried out by the same SLT participant (therapist 1). Therapist 1 was the only SLT who felt in the survey that the therapist could deviate from the manual. She found that not all clients are able to strictly follow the manual's instructions and thus some therapy components should be adjusted to suit client's strengths and weaknesses. This shows that according to the SLT's views therapist's drift is justifiable. Therapist drift refers to the modification of a treatment protocol in small and gradual ways, unintentionally or unknowingly, so that a clinician amends the original protocol in an attempt to respond to a client's specific needs and behaviors (Hinckley and Douglas 2013). Although therapist-drift threatens TI, it is acknowledged that there is a conflict in situations where a researcher, who is also a therapist, feels the obligation to comply with the protocol, but at the same time believes that a deviation from the prescribed treatment would be more helpful to their clients, and thus faces an ethical dilemma (Aradi and Piercy 1985, Sweifach and Linzer 2015). In such instances, the researcher's belief in conjunction with the fact that 'trialists may struggle to exchange their role of providers of individualised care with that of researchers required to follow standardised trial procedures' (Lawton et al. 2011, p. 7) could make the SLT researcher more prone to deviations from the manual.

Treatment integrity scores for group therapy sessions were significantly lower [mean (SD) = 86.7% (6.9)] than for individual therapy sessions [mean (SD) = 94.6% (4.6)], ($t(13)= 2.68$, $p=.019$). This finding is not surprising. Findings from other fields, e.g. mental health, are in line with this, as protocol adherence was significantly higher in individual than in group therapy sessions (Long et al. 2010). It is reasonable to expect a treatment protocol to

be easier to follow when there is only one client in a session. Moreover, ESFA group therapy is more complex in nature, as more therapists' behaviours are anticipated, such as prompts for interaction between the clients and turn-taking control, and it includes more interacting and interconnecting components (Craig et al. 2008); this was also evident by SLT participants' survey replies in terms of therapy complexity. Both of these factors could explain group therapy's lower TI, as it is easier to reach high integrity in simple than complex interventions (Dusenbury et al. 2003). Trying to indicate possible sources of heterogeneity in implementation of group ESFA therapy and address them in a next step could be a useful strategy for achieving even higher TI scores for this approach (Carroll et al. 2007).

Limitations and Directions for Future Research

Limitations of the study include that the checklists comprised all core components rather than only active ingredients of ESFA therapy, and the video sampling method. The active ingredients of a therapy are those which act as catalysts for change and which can be linked to targeted outcomes. They are distinguished from non or less essential components by identifying core components and examining them in relation to measured outcomes (Abry et al. 2015). When the active ingredients of a therapy, like ESFA, are not known, all therapy components need to be examined (Carroll et al. 2007). This approach was followed in the present study. Yet, establishment of the active ingredients of ESFA therapy would facilitate the creation of meaningful thresholds of TI for the ESFA therapy and the identification of the relative importance of each component, which is crucial when guidelines for evaluating integrity are developed (Gresham et al. 2000). It can also provide guidance to practitioners on what to prioritise to make the most of the therapy. Sensitivity or component analysis needs to be conducted using TI information and performance outcomes from a number of ESFA therapy studies to determine which components or combination of them are essential (i.e.,

they are prerequisite if the therapy is to have its expected effect) (Carroll et al. 2007). When the active ingredients will have been identified, the current TI checklists can be modified into more precise tools.

In this project, each of the three SLTs had to record a specific number of videos meeting specific criteria, and these videos were used for the TI evaluation. However, a randomly selected sample should preferably be analysed (Heilemann et al. 2014) in order for the videos sample to be as representative as possible and eliminate sampling bias. Due to lack of resources, this was not feasible for the present study and should be kept in mind when interpreting the findings.

Evaluation of other aspects of TI such as *quality of delivery* and *participant responsiveness* could be targeted in future TI studies of the EFSA therapy, as the degree to which full adherence is achieved may be moderated by these two factors (Carroll et al. 2007). *Quality of delivery* refers to the manner in which a provider delivers a programme, while *participant responsiveness* focuses on the clients, and measures how far they respond to, or they are involved in a therapy, including their judgments about the outcomes and relevance of an intervention (Carroll et al. 2007). Kaderavek and Justice (2010) recommend that quality delivery evaluation is important as ‘a treatment can be implemented badly even when adherence to the procedure is high’ (p. 372). To this end, it is important first to explore which therapist skills are connected with the delivery of the ESFA and then to check the degree to which these skills reflect desired ESFA therapy principles, by including them in the ESFA integrity checklists, as a qualitative section, and applying the same methods as for adherence evaluation (Heilemann et al. 2014). Questionnaires and interviews could be useful methods for addressing the above. Such measures (interviews with therapy stakeholders, patient surveys and document analyses) could also be used in the future, in addition to direct

observation, to make TI findings more robust (Bellg et al. 2004).

As the TI concept gains ground, a conventional criterion for the adequate level of integrity is of paramount importance; until then, decision rules can be seen as arbitrary, with inconsistency on TI score interpretation among researchers. Moreover, the TI terminology needs to be unified for accurate interpretation of findings. Finally, more speech and language therapy studies need to include TI data as an essential component and for those who do so to include precise information about the methods adopted to achieve TI.

Clinical and Research Implications

This study contributes to the outcomes of the Thales aphasia project that investigated the efficacy of ESFA aphasia therapy. The high TI scores enhance the internal validity of the main research project, i.e. confidence that treatment outcomes relate to the treatment as originally planned, given that the protocol was implemented as planned to a high degree (Linnan and Steckler 2002). Moreover, the TI evaluation of the ESFA therapy facilitates the replication of the main study. Treatments that can be measured for adherence to protocol are likely to be sufficiently well described to be replicated (Mowbray et al. 2003, Hinckley and Douglas 2013), permitting future comparison across studies. In this way, external validity is also enhanced (Moncher and Prinz 1991). In addition, this study shows that ESFA as delivered in Thales is well described and therapists can effectively follow the manual and deliver the therapy as intended. Should ESFA prove to be an efficacious approach in Thales, then the first step to implementing it in clinical practice has been taken. Furthermore, the ESFA integrity checklists developed constitute the basis for a follow-up more specific TI tool that could be consistently used for future TI testing of the ESFA therapy.

Last but not least, the present study provides information about current trends in methodology for TI evaluation, while it identifies weaknesses in TI literature, especially in

the aphasia field. Overall, it contributes to the growing prominence of TI in speech and language therapy. While many researchers highlight the need for inclusion of TI data as an essential component in future speech and language therapy studies, the present study highlights the need for more emphasis on the methodological quality of TI reports, to ensure the accurate interpretation of treatment findings.

Supplemental materials

1. **ESFA Integrity checklist for individual therapy**
2. **ESFA Integrity checklist for group therapy**
3. **SLT participants' views on ESFA therapy survey**

Table 1. Session-specific inter-rater reliability values and TI scores

Video Number	TI score R1	TI score R2	Inter-rater reliability (κ, $p < .001$)	Level of agreement (Cicchetti and Sparrow 1981)
3	87.0%	78.7%	.63	good
6	94.8%	94.4%	.94	excellent
8	96.7%	94.3%	.88	excellent
10	78.6%	75.3%	.81	excellent
14	92.6%	97.2%	.78	excellent

R1= rater 1 (first author); R2= rater 2 (independent rater)

Table 2. Session-specific intra-rater reliability values and TI scores

Video Number	TI score T1	TI score T2	Intra-rater reliability (κ, $p < .001$)	Level of agreement (Cicchetti and Sparrow 1981)
6	94.8%	93.5%	.95	excellent
7	97.2%	97.2%	1.00	excellent
9	100.0%	97.8%	.93	excellent
11	77.2%	77.2%	1.00	excellent
14	92.6%	93.7%	.99	excellent

T1= time 1; T2= time 2

Table 3. Session-specific TI scores and overall TI score for individual therapy approach

Session No (Therapist No)	Components Planned**, Maximum TI score**	Components Implemented**, Actual TI score**	**TI score (%)**
1 (1)	51	50	98%
2 (1)	61	53.5	87.7%
3 (1)	50	43.5	87%
4 (2)	34	31.5	92.6%
5 (2)	69	67	97.1%
6 (2)	67	63.5	94.8%
7 (3)	36	35	97.2%
8 (3)	61	59	96.7%
9 (3)	21	21	100%
Overall	450	424	-
Mean (SD)	50 (16.5)	47.3 (15.8)	94.6% (4.6)

Table 4. **Session-specific TI scores and overall TI score for group therapy approach**

Session No (Therapist No)	Components Planned**, Maximum TI score**	Components Implemented**, Actual TI score**	**TI score (%)**
10 (1)	98	77	78.6%
11 (1)	46	35.5	77.2%
12 (2)	55	49.5	90.0%
13 (2)	28	25	89.3%
14 (3)	81	75	92.6%
15 (3)	78	72	92.3%
Overall	386	334	-
Mean (SD)	64.3 (25.9)	55.7 (22.3)	86.7% (6.9)

Table 5. Therapists' views on complexity of ESFA

Complexity Dimensions	Therapist 1	Therapist 2	Therapist 3
Large number of (complex) behaviours required by those delivering or receiving the intervention	☑		☑
Different groups targeted by the intervention	☑		☑
There is a variability in therapy outcomes	☑		
High level of flexibility or tailoring of the intervention is permitted	☑	☑	

Figure 1. Therapists' ratings for facilitation strategies used

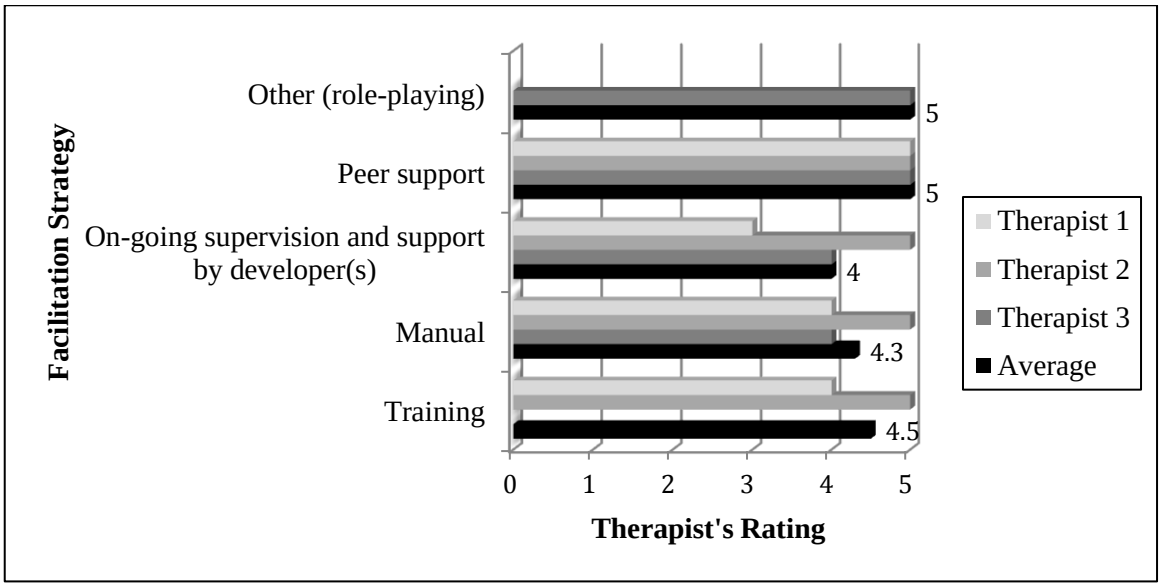
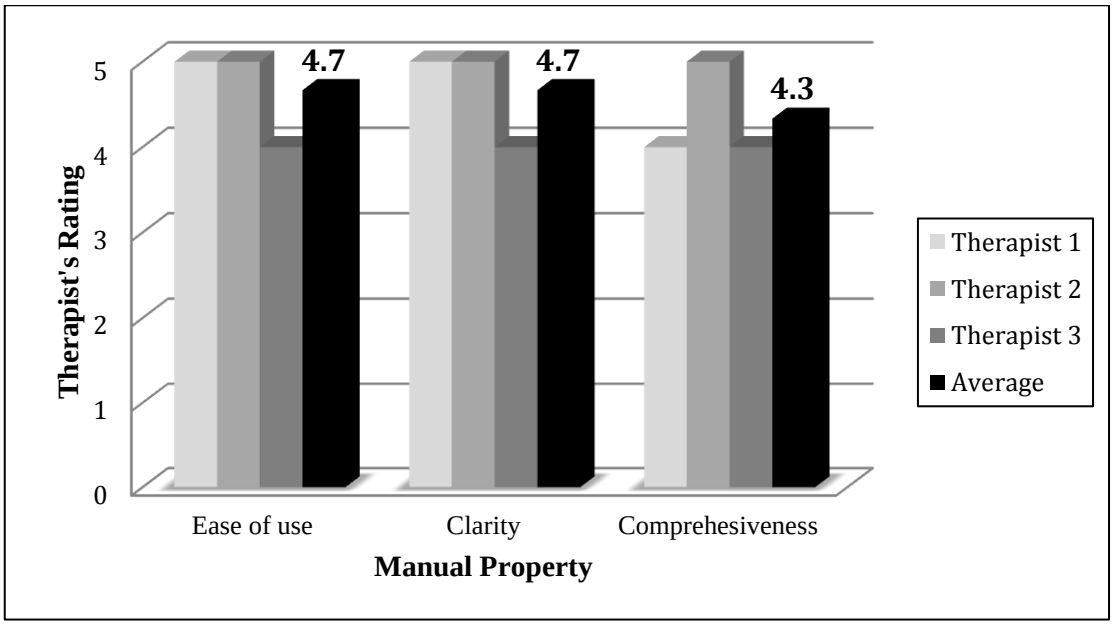


Figure 2. Therapists' ratings of the manual's properties



References

ABRY, T., HULLEMAN, C. S., and RIMM-KAUFMAN, S. E., 2015, Using indices of fidelity to intervention core components to identify program active ingredients. American Journal of Evaluation, 36(3), 320-338.

ARADI, N. S., and PIERCY, F. P., 1985, Ethical and legal guidelines related to adherence to treatment protocols in family therapy outcome research. The American Journal of Family Therapy, 13(3), 60-65.

BELLG, A. J., BORRELLI, B., RESNICK, B., HECHT, J., MINICUCCI, D. S., ORY, M., OGEDEGBE, G., ORWIG, D., ERNST, D., and CZAJKOWSKI, S., 2004, Enhancing treatment fidelity in health behavior change studies: best practices and recommendations from the NIH Behavior Change Consortium. Health Psychology, 23(5), 443-451.

BOYLE, M., 2004, Semantic feature analysis treatment for anomia in two fluent aphasia syndromes. American Journal of Speech-Language Pathology, 13(3), 236-249.

BOYLE, M., and COELHO, C. A., 1995, Application of Semantic Feature Analysis as a treatment for aphasic dysnomia. American Journal of Speech-Language Pathology, 4(4), 94-98.

CARROLL, C., PATTERSON, M., WOOD, S., BOOTH, A., RICK, J., and BALAIN, S., 2007, A conceptual framework for implementation fidelity. Implementation Science, 2(1), 1-9.

CHERNEY, L. R., PATTERSON, J. P., RAYMER, A., FRYMARK, T., and SCHOOLING, T., 2008, Evidence-based systematic review: effects of intensity of treatment and constraint-induced language therapy for individuals with stroke-induced aphasia. Journal of Speech, Language, and Hearing Research, 51(5), 1282-1299.

CHERNEY, L. R., SIMMONS-MACKIE, N., RAYMER, A., ARMSTRONG, E., and HOLLAND, A., 2013, Systematic review of communication partner training in aphasia: methodological quality. International Journal of Speech-Language Pathology, 15(5), 535-545.

CICCHETTI, D. V., 1994, Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. Psychological Assessment, 6(4), 284-290.

CLARKE, G., 1998, Intervention fidelity in the psychosocial prevention and treatment of adolescent depression. Journal of Prevention & Intervention in the Community, 17(2), 19-33.

COCHRANE, W. S., and LAUX, J. M., 2008, A survey investigating school psychologists' measurement of treatment integrity in school-based interventions and their beliefs about its importance. Psychology in the Schools, 45(6), 499-507.

COELHO, C. A., MCHUGH, R. E., and BOYLE, M., 2000, Semantic feature analysis as a treatment for aphasic dysnomia: A replication. Aphasiology, 14(2), 133-142.

COHEN, J., 1960, A coefficient of agreement for nominal scales. Educational and Psychosocial Measurement, 20, 37-46.

CONLEY, A., and COELHO, C., 2003, Treatment of word retrieval impairment in chronic Broca's aphasia. Aphasiology, 17(3), 203-211.

CRAIG, P., DIEPPE, P., MACINTYRE, S., MICHIE, S., NAZARETH, I., and PETTICREW, M., 2008, Developing and evaluating complex interventions: the new Medical Research Council guidance. British Medical Journal, 337, a1655.

DIETZ, A., KNOLLMAN-PORTER, K., HUX, K., TOTH, K., and BROWN, B., 2014a, Supported reading comprehension for people with aphasia: Visual and linguistic supports. Journal of Medical Speech-Language Pathology, 21(4), 319-331.

DIETZ, A., WEISSLING, K., GRIFFITH, J., MCKELVEY, M., and MACKE, D., 2014b, The impact of interface design during an initial high-technology AAC experience: a collective case study of people with aphasia. Augmentative and Alternative Communication, 30(4), 314-328.

DUSENBURY, L., BRANNIGAN, R., FALCO, M., and HANSEN, W. B., 2003, A review of research on fidelity of implementation: implications for drug abuse prevention in school settings. Health Education Research, 18(2), 237-256.

EDMONDS, L. A., and BABB, M., 2011, Effect of verb network strengthening treatment in moderate-to-severe aphasia. American Journal of Speech-Language Pathology, 20(2), 131-145.

EDMONDS, L. A., and KIRAN, S., 2006, Effect of semantic naming treatment on crosslinguistic generalization in bilingual aphasia. Journal of Speech, Language, and Hearing Research, 49(4), 729-748.

EDMONDS, L. A., NADEAU, S. E., and KIRAN, S., 2009, Effect of Verb Network Strengthening Treatment (VNeST) on lexical retrieval of content words in sentences in persons with aphasia. Aphasiology, 23(3), 402-424.

EGAN, J., WORRALL, L., and OXENHAM, D., 2004, Accessible Internet training package helps people with aphasia cross the digital divide. Aphasiology, 18(3), 265-280.

FAROQI-SHAH, Y., FRYMARK, T., MULLEN, R., and WANG, B., 2010, Effect of treatment for bilingual individuals with aphasia: A systematic review of the evidence. Journal of Neurolinguistics, 23(4), 319-341.

FIXSEN, D.L., NAOOM, S.F., BLASE, K.A., FRIEDMAN, R.M., and WALLACE, F., 2005, Implementation research: A synthesis of the literature (FMHI Publication No. 231). Tampa, FL: University of South Florida, Louis de la Parte Florida

Mental Health Institute, The National Implementation Research Network. Available at: http://ctndisseminationlibrary.org/PDF/nirnmonograph.pdf (Accessed: 2 August 2016).

GOFF, R. A., 2013, Examining the effectiveness of intensive language action therapy in individuals with nonfluent aphasia. PhD Thesis, University of South Florida [Online]. Available at: http://scholarcommons.usf.edu/cgi/viewcontent.cgi?article=6013&context=etd&sei-redir=1&referer=https%3A%2F%2Fscholar.google.co.uk%2Fscholar%3Fq%3DExamining%2Bthe%2Beffectiveness%2Bof%2Bintensive%2Blanguage%2Baction%2Btherapy%2Bin%2Bindividuals%2Bwith%2Bnonfluent%2Baphasia.%26hl%3Den%26as_sdt%3D0%26as_vis%3D1%26oi%3Dscholart%26sa%3DX%26ved%3D0ahUKEwiWyOaYg6POAhWByRQKHQF_AuYQgQMIHDAA#search=%22Examining%20effectiveness%20intensive%20language%20action%20therapy%20individuals%20nonfluent%20aphasia.%22 (Accessed: 2 August 2016).

GRESHAM, F. M., MACMILLAN, D. L., BEEBE-FRANKENBERGER, M. E., and BOCIAN, K. M., 2000, Treatment integrity in learning disabilities intervention research: Do we really know how treatments are implemented? Learning Disabilities Research & Practice, 15(4), 198-205.

GRIFFITH, J., DIETZ, A., and WEISSLING, K., 2014, Supporting narrative retells for people with aphasia using augmentative and alternative communication: Photographs or line drawings? Text or no text? American Journal of Speech-Language Pathology, 23(2), S213-S224.

HALLGREN, K. A., 2012, Computing inter-rater reliability for observational data: An overview and tutorial. Tutorials in Quantitative Methods for Psychology, 8(1), 23-34.

HEILEMANN, C., 2013, Investigating aspects of treatment fidelity in a new conversation-based therapy for people with agrammatic aphasia and their conversation partners. PhD Thesis, University College London [Online]. Available at: https://epub.ub.uni-muenchen.de/17910/1/Master_Thesis_Heilemann_2013.pdf (Accessed: 2 August 2016)

HEILEMANN, C., BEST, W., JOHNSON, F., BECKLEY, F., EDWARDS, S., MAXIM, J., and BEEKE, S., 2014, Investigating treatment fidelity in a conversation-based aphasia therapy. Aphasie und Verwandte Gebiete, 2, 14-26.

HICKEY, E., BOURGEOIS, M., and OLSWANG, L., 2004, Effects of training volunteers to converse with nursing home residents with aphasia. Aphasiology, 18(5-7), 625-637.

HINCKLEY, J. J., and CARR, T., 2005, Comparing the outcomes of intensive and non-intensive context-based aphasia treatment. Aphasiology, 19(10-11), 965-974.

HINCKLEY, J. J., and DOUGLAS, N. F., 2013, Treatment fidelity: its importance and reported frequency in aphasia treatment studies. American Journal of Speech-Language Pathology, 22(2), S279-284.

HOFFMANN, T.C., GLASZIOU, P.P., BOUTRON, I., MILNE, R., PERERA, R., MOHER, D., ALTMAN, D.G., BARBOUR, V., MACDONALD, H., JOHNSTON, M. AND LAMB, S.E., 2014, Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ, 348,1687.

HOGUE, A., LIDDLE, H. A., SINGER, A., and LECKRONE, J., 2005, Intervention fidelity in family-based prevention counseling for adolescent problem behaviors. Journal of Community Psychology, 33(2), 191-211.

KADERAVEK, J. N., and JUSTICE, L. M., 2010, Fidelity: An essential component of evidence-based practice in speech-language pathology. American Journal of Speech-Language Pathology, 19(4), 369-379.

KAZDIN, A. E., 1986, Comparative outcome studies of psychotherapy: Methodological issues and strategies. Journal of Consulting and Clinical Psychology, 54(1), 95-105.

KEMPLER, D., and GORAL, M., 2011, A comparison of drill- and communication-based treatment for aphasia. Aphasiology, 25(11), 1327-1346.

KIRAN, S., 2008, Typicality of inanimate category exemplars in aphasia treatment: Further evidence for semantic complexity. Journal of Speech, Language, and Hearing Research, 51(6), 1550-1568.

KIRAN, S., and JOHNSON, L., 2008, Semantic complexity in treatment of naming deficits in aphasia: Evidence from well-defined categories. American Journal of Speech-Language Pathology, 17(4), 389-400.

KIRAN, S., SANDBERG, C., and SEBASTIAN, R., 2011, Treatment of category generation and retrieval in aphasia: Effect of typicality of category items. Journal of Speech, Language, and Hearing Research, 54(4), 1101-1117

KIRAN, S., and THOMPSON, C. K., 2003, The role of semantic complexity in treatment of naming deficitstraining semantic categories in fluent aphasia by controlling exemplar typicality. Journal of Speech, Language, and Hearing Research, 46(3), 608-622.

LANE, K. L., BOCIAN, K. M., MACMILLAN, D. L., and GRESHAM, F. M., 2004, Treatment integrity: An essential—but often forgotten—component of school-based interventions. Preventing School Failure: Alternative Education for Children and Youth, 48(3), 36-43.

LAWTON, J., JENKINS, N., DARBYSHIRE, J., HOLMAN, R., FARMER, A., and HALLOWELL, N., 2011, Challenges of maintaining research protocol fidelity in a clinical care setting: A qualitative study of the experiences and views of patients and staff participating in a randomized controlled trial. Trials, 12(1), 108.

LEONARD, C., ROCHON, E., and LAIRD, L., 2008, Treating naming impairments in aphasia: Findings from a phonological components analysis treatment. Aphasiology, 22(9), 923-947.

LICHSTEIN, K. L., RIEDEL, B. W., and GRIEVE, R., 1994, Fair tests of clinical trials: A treatment implementation model. Advances in Behaviour Research and Therapy, 16(1), 1-29.

LINNAN, L. and STECKLER, A., 2002, Process evaluation for public health interventions and research: An overview. In L. Linnan and A. Steckler (eds), Process evaluation for public health interventions and research (San Francisco: Jossey-Bass), pp. 1–23.

LONG, M. E., GRUBAUGH, A. L., ELHAI, J. D., CUSACK, K. J., KNAPP, R., and FRUEH, B. C., 2010, Therapist fidelity with an exposure-based treatment of ptsd in adults with schizophrenia or schizoaffective disorder. Journal of Clinical Psychology, 66(4), 383-393.

LOWELL, S., BEESON, P. M., and HOLLAND, A. L., 1995, The efficacy of a semantic cueing procedure on naming performance of adults with aphasia. American Journal of Speech-Language Pathology, 4(4), 109-114.

MATTIS, S., 1988, Dementia Rating Scale: Professional manual. (Odessa, FL: Psychological Assessment Resources).

MCINTYRE, L. L., GRESHAM, F. M., DIGENNARO, F. D., and REED, D. D., 2007, Treatment integrity of school-based interventions with children in the journal of

applied behavior analysis 1991-2005. Journal of Applied Behavior Analysis, 40(4), 659-672.

MONCHER, F. J., and PRINZ, R. J., 1991, Treatment fidelity in outcome studies. Clinical Psychology Review, 11(3), 247-266.

MOWBRAY, C. T., HOLTER, M. C., TEAGUE, G. B., and BYBEE, D., 2003, Fidelity criteria: Development, measurement, and validation. American Journal of Evaluation, 24(3), 315-340.

NETEMEYER, R. G., BEARDEN, W. O., and SHARMA S., 2003, Validity. In R. G. Netemeyer (eds), Scaling procedures: Issues and applications (Thousand Oaks: Sage), pp. 71-87.

PAPATHANASIOU I. and MIHOU A., 2006, Elaborative Semantic Feature Analysis: A case study. [Poster]. 2006 ASHA's Convention, Miami, USA, November 16 - 18 2006.

PAPATHANASIOU I., PAPADIMITRIOU D., GAVRILOU V., and MIHOU A., 2008, Psychometric propertis of BDAE in normal adult population: The effect of age and gender (Greek edition). Psychology, 15(4), 398-410.

PEACH, R. K., and REUTER, K. A., 2010, A discourse-based approach to semantic feature analysis for the treatment of aphasic word retrieval failures. Aphasiology, 24(9), 971-990.

PEREPLETCHIKOVA, F., and KAZDIN, A. E., 2005, Treatment integrity and therapeutic change: issues and research recommendations. Clinical Psychology: Science and Practice, 12(4), 365-383.

RAVEN, J., 2004, Coloured progressive matrices and Crichton vocabulary scale (London, England: Pearson).

RIDER, J. D., WRIGHT, H. H., MARSHALL, R. C., and PAGE, J. L., 2008, Using semantic feature analysis to improve contextual discourse in adults with aphasia. American Journal of Speech-Language Pathology, 17(2), 161-172.

ROSE, M., and DOUGLAS, J., 2006, A comparison of verbal and gesture treatments for a word production deficit resulting from acquired apraxia of speech. Aphasiology, 20(12), 1186-1209.

ROSE, M., DOUGLAS, J., and MATYAS, T., 2002, The comparative effectiveness of gesture and verbal treatments for a specific phonologic naming impairment. Aphasiology, 16(10-11), 1001-1030.

ROSE, M. L., RAYMER, A. M., LANYON, L. E., and ATTARD, M. C., 2013, A systematic review of gesture treatments for post-stroke aphasia. Aphasiology, 27(9), 1090-1127.

ROSE, M., and SUSSMILCH, G., 2008, The effects of semantic and gesture treatments on verb retrieval and verb use in aphasia. Aphasiology, 22(7-8), 691-706.

ROSSION, B., and POURTOIS, G., 2004, Revisiting Snodgrass and Vanderwart's object pictorial set: The role of surface detail in basic-level object recognition. Perception, 33(2), 217-236.

SCHNEIDER, S., and FRENS, R., 2005, Training four-syllable CV patterns in individuals with acquired apraxia of speech: Theoretical implications. Aphasiology, 19(3-5), 451-471.

SCHOENWALD, S. K., GARLAND, A. F., CHAPMAN, J. E., FRAZIER, S. L., SHEIDOW, A. J., and SOUTHAM-GEROW, M. A., 2011, Toward the effective and efficient measurement of implementation fidelity. Administration and Policy in Mental Health and Mental Health Services Research, 38(1), 32-43.

SNODGRASS, J. G., and VANDERWART, M., 1980, A standardized set of 260 pictures: Norms for name agreement, image agreement, familiarity, and visual complexity. Journal of Experimental Psychology: Human Learning and Memory, 6(2), 174-215.

STEIN, K. F., SARGENT, J. T., and RAFAELS, N., 2007, Intervention research: Establishing fidelity of the independent variable in nursing clinical trials. Nursing Research, 56(1), 54-62.

STUFFLEBEAM, D. L., 2000, Guidelines for developing evaluation checklists: The Checklists Development checklist (CDC). Available at: https://www.wmich.edu/sites/default/files/attachments/u350/2014/guidelines_cdc.pdf (Accessed: 2 August 2016).

SWEIFACH, J. S., and LINZER, N., 2015, Beneficence vs. Fidelity: Serving social work clients in the aftermath of catastrophic events. Journal of Social Work Values and Ethics, 12(1), 3-12.

WAMBAUGH, J. L., and WRIGHT, S., 2007, Improved effects of word-retrieval treatments subsequent to addition of the orthographic form. Aphasiology, 21(6-8), 632-642.

WRIGHT, H. H., MARSHALL, R. C., WILSON, K. B., and PAGE, J. L., 2008, Using a written cueing hierarchy to improve verbal naming in aphasia. Aphasiology, 22(5), 522-536.

YEATON, W. H., and SECHREST, L., 1981, Critical dimensions in the choice and maintenance of successful treatments: Strength, integrity, and effectiveness. Journal of Consulting and Clinical Psychology, 49(2), 156-167.

YODER, P., and SYMONS, F., 2010, Interobserver Agreements and Reliability of Observational Variables. In P. Yoder and F. Symons (eds), Observational Measurement of Behavior (New York: Springer), pp. 159-182.

Appendix. Detailed Description of the ESFA therapy (Individual therapy approach)

Elaborated Semantic Feature Analysis (ESFA) Therapy Background. ESFA therapy is a modified version of Semantic Feature Analysis (SFA) therapy, as it is based on the SFA approach, but also prompts the individual, after word retrieval, to elaborate the features of the word elicited on the SFA chart into a sentence. It also includes provision of elaborate cueing hierarchies to elicit features when participants cannot produce them. Moreover, during ESFA therapy, participants are encouraged to write the features on the chart, as a self-cuing strategy. ESFA treatment approach was applied to improve word retrieval of object nouns in aphasia. The purpose of this approach is to enable the individual to transfer their naming abilities to connected speech.

Stimuli. Therapy targets were chosen based on the results of three oral confrontation-naming sessions. In each of these sessions the participant was asked to name the Rossion & Pourtois coloured version of the 260 Snodgrass and Vanderwart line drawings of nouns (Rossion and Pourtois 2004). The pictures were presented on a computer screen in a timed PowerPoint presentation in a random order to each participant and were scored as either correct or incorrect. Participants had 13 seconds to respond before the next picture was shown. Correct responses were intelligible productions of the target word within this timeframe.

The pictures that a client failed to name on at least two trials were selected as treatment and probe stimuli. This process of stimulus selection resulted in a set of treatment and probe items that were specific to each client in terms of content and number of items. Seventy percent of the incorrect responses were selected as treatment material, while the other 30% was used as untreated generalisation stimuli. The exact same procedure was followed for the stimulus

selection for the group therapy. The only difference was that the final stimulus selection took place by comparing the results of all administrations of all the clients participating in the group and then selecting for treatment those words that were not named by all group participants.

Intervention Providers. ESFA Therapy providers in this study were the three research speech and language therapists (SLTs) who were trained in ESFA and delivered the treatment in the Thales aphasia project. All three participants had a Master's degree, four to nine years of clinical experience and had worked with PwA from two to seven years.

Modes of Delivery, Location, Dose. ESFA therapy was delivered through two different modes: individual therapy vs. a combination of individual and group therapy, although at this section only individual therapy approach is described. The sessions took place mainly in the participants' home and some in hospital settings. Participants were randomised to receive either 36 hours of individual therapy (three one-hour sessions per week for 12 weeks) or 36 hours of a combination of individual and group therapy (two 45-minute individual sessions and one 1½ - hour group session per week for 12 weeks).

Main Therapy procedures. ESFA therapy approach is based on SFA therapy (Boyle and Coelho 1995); however, it differs in aspects, including provision of elaborate cueing hierarchies to elicit features when participants cannot produce them. Moreover, during ESFA therapy, participants are encouraged to write the features on the chart, as a self-cueing strategy, but as writing is not target of ESFA therapy, writing errors are not corrected by the therapist.

In terms of ESFA therapy procedure, during the therapy session, for each item trained, the clinician initially asked the client to draw a picture from the treatment material set and then to name it. Then, presenting a semantic features chart [same as that shown in Boyle (2004), but

translated in Greek language], the therapist prompted the client to think of and say words semantically related to the target word (semantic features). The chart included six categories: *superordinate category, use, action, physical properties, location,* and *association.* To elicit feature production, the therapist asked questions or provided the client with sentence-completion cues. For instance, for the *superordinate category*, a question such as "*What category does it belong to?"* was provided. Similarly, for the category *use*, a statement as "*You use it to/for* _____" was given. After the oral word production, which is the focus of ESFA therapy, the clinician prompted the client to write down the elicited features in the chart. For clients with writing difficulties, the therapist helped them to write the features with the use of an alphabet table (e.g. pointing to the letters they needed). For clients who could not write, the therapist filled in the chart.

After the chart completion and the retrieval of the word by the client, when the SFA procedure was completed for the target word, the therapist encouraged the client to produce phrases with the target word and each of its features. If needed, the clinician and client would say the words together or the clinician would point to the target and a feature for the client to put together in a phrase. Then, the client was encouraged to produce a sentence, including the target word and at least one of the elicited semantic features. For example for the item 'table', the individual was asked during SFA to produce features such as: furniture, for dining, wooden, kitchen, chair, tea, eat, and then to elaborate these features in sentences such as: we eat at the table, we have tea at the table, the table is for dining, the table is a piece of furniture in the kitchen, etc. Elaboration of features was achieved by asking the individual to choose as many features as they wanted (one as a minimum) and to put them together into a sentence. The same strategy was followed for all treatment items. Participants had first to produce the sentence orally and then if they could to write the sentence down. It did not matter if people made errors

in their sentences, e.g. syntactic or morphological errors as long as the sentence was meaningful. After its completion (SFA stage), the chart was used as help/ cueing as and when needed.

At the end of each session the client was asked to name all the words that had been worked on during the previous therapeutic sessions: if a target word was retrieved correctly for three consecutive sessions, without prompt or help by the therapist, and the client was able to produce correct sentences without cues or reference to the chart, this word was removed from the therapy process and another new word replaced it. Subsequently, at the beginning of each therapy session, the client was asked to name the pictures that they had not named correctly in the previous session and to produce one sentence for each of these target words. If the client did not name the picture correctly, the chart analysis was repeated with these targets before moving on to new targets.

Additional Therapy Principles. In terms of the order of chart completion, there was flexibility. At the first therapy sessions the therapists would start for animate nouns, e.g. 'dog' with the first category (*superordinate* category), e.g. 'what is it?' or 'what group does it belong to?' and for inanimate nouns, e.g. 'scissors' with the *action* category, e.g. 'what do we do with it?' or the *use* category, e.g. 'we use it for…?', and then work their way through the other features in the following order: *physical properties, location*, and *association.* However, as the participants became familiar with the technique, they were let to spontaneously generate features out of sequence. When this happened, the features were written in the appropriate boxes on the chart, and if and when needed the clinician resumed eliciting features in the prescribed order, skipping over the categories that the participant had spontaneously completed. If a category was not applicable for a target word, such as when *use* and *action* categories are similar (e.g. for paintbrush: to paint), then this category was skipped by the therapist and only

those deemed appropriate for the target item were elicited. If a participant named the target picture on confrontation or during the features generation, the therapist still asked for all features to be produced, in order for the participant to build up semantic links, promoting spreading activation to related semantic concepts. This also aimed to develop feature generation as a compensatory strategy by encouraging the establishment of the technique and its use and, through repeated practice, to increase the chances of a more automatic use of the technique when lexical retrieval difficulties were encountered. The client was prompted to produce as many features as possible for each category, which were then written in the category box, as more related words facilitate the connections of the semantic network. Some categories elicited more features compared to others: the *physical properties category*, for example, typically had several entries, whereas the box for *superordinate category* had fewer. The production of more than one feature for each category was not an integral component of ESFA though; one semantic feature for each category was the basic requirement. The number of the pictures worked on in each session depended on the client's performance.

During the therapy, the therapist provided cues to clients, following a specific cueing hierarchy based on the type of paraphasias produced. The hierarchy followed is demonstrated on the integrity checklists (Supplemental Materials). If the client was not able to produce the word after cueing, they were led through the entire SFA chart, with cues provided as needed, to produce the target word. When the client could not produce the target work even when all features had been listed, the clinician produced the word orally and then the participant repeated it and named all of its features.

Tailoring. ESFA therapy as described by the treatment protocol is a therapy that is administered to all clients in the same way, with differences depending only to the therapy mode (individual vs group therapy). However, therapists' cues and help during the process took

place according to client's aphasia type and abilities or performance. In more detail, in terms of chart completion, the clinician prompted the client to write down the elicited features in the chart. For clients with writing difficulties though, the therapist helped them to write the features with the use of an alphabet table (e.g. pointing to the letters they needed). For clients who could not write, the therapist filled in the chart. Regarding the phrase production, therapist encouraged the client to produce phrases with the target word and each of its features. If needed however, the clinician and client would say the words together or the clinician would point to the target and a feature for the client to put together in a phrase. During the sentence production, when the individual was asked to choose as many features as they wanted (one as a minimum) and to put them together into a sentence, help was given to participants according to their abilities; people with global aphasia for instance, needed more cues from the therapist compared to people with fluent aphasia, while over time, therapist's help was reduced. Finally, during the therapy, the therapist provided cues to clients, following a specific cueing hierarchy based on the type of paraphasias produced by the clients. The hierarchy followed is demonstrated on the integrity checklists (Supplemental Materials).

Treatment Integrity. The present study investigated ESFA adherence to the treatment protocol (Treatment Integrity – TI), through observation of therapy videos. Moreover, different strategies, including training, use of the treatment manual, supervision and support by developers, and peer support were used to maintain and enhance integrity. ESFA therapy was delivered as planned in a high degree, i.e. 91.4% (94.6% for individual therapy approach and 86.7% for group therapy approach).