The effectiveness of individualized morphosyntactic target identification and explicit intervention using the SHAPE CODING™ system for children with Developmental Language Disorder (DLD) and the impact of within-session dosage.

Susan H Ebbels[1,2], Mollie Gadd[1], Hilary Nicoll[1], Lucy Hughes [1,2], Nicola Dawson[1,5], Caroline Burke[1], Samuel D Calder[3], Pauline Frizelle[4]

[1]Moor House Research and Training Institute, Moor House School & College, Oxted, UK
[2] Division of Psychology and Language, University College London, UK
[3] Health Sciences, College of Health and Medicine, University of Tasmania, Launceston, Australia
[4] Department of Speech and Hearing Sciences, University College Cork, Republic of Ireland
[5] Department of Experimental Psychology, University of Oxford, UK

Correspondence to: Susan Ebbels, Moor House Research and Training Institute, Moor House School & College, Mill Lane, Hurst Green, Oxted, Surrey, RH8 9HN, UK, ebbelss@moorhouseschool.co.uk Tel: 00 44 1883 712271

Conflict of Interest: Susan Ebbels, Hilary Nicoll, Caroline Burke, Lucy Hughes and Nicola Dawson are (and Mollie Gadd was at the time of the project) employed by Moor House School & College which sells resources and courses on the SHAPE CODING™ system. There are no other conflicts of interest.

Funding: this research received no specific funding

Ethics: This research was approved by the Moor House Ethics Committee

Cite as: Ebbels, S., Gadd, M., Nicoll, H., Hughes, L., Dawson, N., Burke, C., Calder, S., Frizelle, P. (in press). The effectiveness of individualized morphosyntactic target identification and explicit intervention using the SHAPE CODING™ system for children with Developmental Language Disorder (DLD) and the impact of within-session dosage. *Language Speech and Hearing Services in Schools*. https://pubs-asha-org.libproxy.ucl.ac.uk/doi/full/10.1044/2024_LSHSS-23-00098

Abstract

Purpose

We investigated the effectiveness of a highly individualized morphosyntactic intervention using the SHAPE CODING™ system delivered at different dosages.

Method

Eight children with DLD aged 8;0-10;10 received ten hours of explicit individualized intervention for morphosyntax delivered in 30-minute individual sessions once per week for 20 weeks. Following at least four baseline probe tests, two grammatical targets per session received explicit instruction until they reached criterion (90%), when the next target was introduced. To control for session length and teaching episode density, either both targets received 20 teaching episodes per session, or one target received 10 teaching episodes and the other 30. Maintenance testing of completed targets was also carried out.

Results

Scores on probe tests post-intervention were significantly higher than during the baseline phase (d=1.6) with no change during the baseline or maintenance phases. However, progress during the intervention phase was highly significant. One participant showed significantly faster progress with intervention while one (with the lowest attention score) made little progress.

When considering progress relative to cumulative intervention sessions, progress was faster with 30 teaching episodes per session and slower with 10. However, when cumulative teaching episodes was used as the predictor, all three within-session dosages showed very similar rates of progress, with the odds of a correct response increasing by 3.9% for each teaching episode. The targets that were achieved required an average of 40-60 teaching episodes.

Conclusions

With the exception of one participant, the individualized intervention was highly effective and efficient. Thus, the individualized target identification process and intervention method merit further research in a larger group of children. The cumulative number of teaching episodes per target provided across sessions appeared to be key. Thus, clinicians should aim for high teaching episode rates, particularly if the number of sessions is constrained. Otherwise, intervention scheduling can be flexible.

Keywords: Developmental Language Disorder; intervention; grammar; dosage

The effectiveness of individualized morphosyntactic target identification and explicit intervention using the SHAPE CODING™ system for children with Developmental Language Disorder (DLD) and the impact of within-session dosage.

Introduction

Language is at the heart of children's learning, social interaction and wider development, yet approximately 10% of children have a language disorder, with 7% meeting the criteria for Developmental Language Disorder (DLD) (Norbury et al., 2016). In an increasingly information-based economy, difficulties understanding and using spoken and written language put children at risk of poor educational (Nippold, 2010), employment (Conti-Ramsden et al., 2018) and mental health outcomes (Yew & O'Kearney, 2013). Consequently, providing effective and efficient support to help children with language difficulties develop their language skills becomes imperative. It is therefore unsurprising that the top research priorities identified by researchers, practitioners and individuals with lived experience of DLD all focus on establishing effective interventions (Kulkarni et al., 2022), with the second highest priority being identification of the "specific characteristics of evidence-based DLD interventions which facilitate progress towards the goals of an individual with DLD" (p.958).

Given the complexity of language, language intervention requires a commitment of both time and money from children, families, schools, clinicians and funders. Therefore, maximizing intervention efficiency is of paramount importance to all involved. A focus on efficiency requires consideration of both time and outcomes, and efforts to increase efficiency should aim to either reduce the time in intervention without limiting outcomes, or improve outcomes without increasing the time in intervention.

Time-related aspects of intervention delivery such as *session duration, session frequency, number of sessions* and *total intervention period* are often centrally determined (e.g.,

30-minute sessions, once per fortnight, for six sessions, delivered over one school term). However, clinicians have much greater control over what happens within intervention sessions in terms of the *intervention target*, *intervention steps, techniques* (comprising of *clinician* and *child actions* which are hypothesized to be *active ingredients* and combined into *teaching episodes*) and *time (not) focused on intervention targets*. Figure 1 shows these components in a hypothetical session. All of these within session factors could greatly affect the effectiveness and efficiency of intervention but may vary hugely by target, child and clinician, leading to more or less effective and efficient intervention sessions. Thus, a focus on increasing efficiency should consider the details of what happens within intervention sessions, rather than just the *session duration.*

Maximally efficient intervention is likely to use an intervention method that supports learning in areas of weakness by capitalizing on areas of relative strength, to involve highly individualized targets and last only as long as necessary for the child to achieve mastery of each target, or to plateau. Where children have multiple areas needing intervention, working on more than one area or target per session, or moving onto other targets as soon as a target has been achieved could also maximize efficiency. Dosage may also affect efficiency. We will next consider each of these factors: intervention methods, targets and dosage.

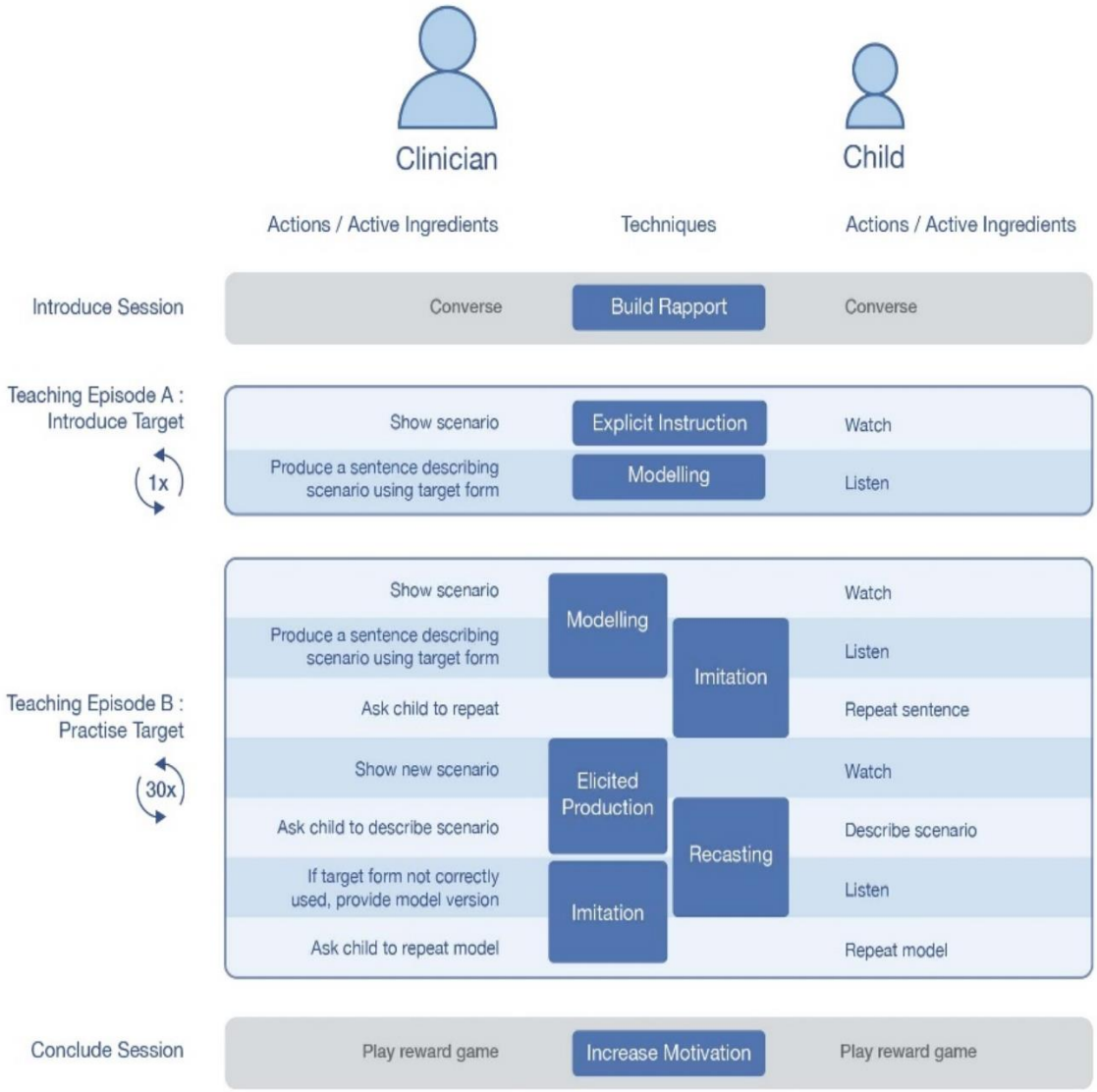


Figure 1 – Components of a hypothetical intervention session

Intervention methods

In typical language development, children extract the statistical regularities of the language(s) they are exposed to without effort or awareness. However, this form of implicit learning is challenging for children with DLD (Lammertink et al., 2017), leading to significant and persistent difficulties in forming accurate and complex sentences to express their thoughts. Indeed, while children with DLD have deficits across language domains, their grammatical abilities are particularly affected (Leonard, 2014), such that they often produce

short, simple sentences (Nippold et al., 2009) containing grammatical errors (Scott & Windsor, 2000) and have difficulty understanding and using complex sentences (Frizelle & Fletcher, 2014a, 2014b), all of which affect their ability to communicate with others and access the educational curriculum.

Interventions for grammar often employ either predominantly implicit or explicit methods of intervention (see Ebbels, 2014), or a combination. Predominantly implicit interventions make target forms more frequent and salient and have been shown to improve specific aspects of expressive syntax in young pre-school or early school-aged children with primarily expressive language deficits or delays, when between 15 and 60 hours of intervention are delivered 1:1 by an SLP or parent (Fey et al., 1993; Fey et al., 1997; Owen van Horne et al., 2017; Plante et al., 2014). Few studies of implicit interventions include older children or those with more wide-ranging difficulties including receptive language.

Interventions using explicit methods of instruction tell the children the grammatical rules of their language, often using visual cues in order to capitalize on the relative strengths of children with DLD in declarative memory and visuo-spatial skills (Lum et al., 2012). Of the explicit interventions for morphosyntax, the SHAPE CODING™ system has the strongest evidence base. Small-scale randomized control trials conducted in specialist settings demonstrate improvements in targeted aspects of grammar (and generalization to non-targeted items following the same rule) in school-aged children (aged 5-16 years) including those with receptive language deficits, when intervention is delivered for 5-10 hours 1:1 by a Speech Language Pathologist, SLP (Calder, Claessen, Ebbels, et al., 2021; Ebbels et al., 2014; Ebbels et al., 2007). Given the shorter amount of time in intervention compared with implicit instruction, it seems that explicit instruction may be more efficient (at least for children with DLD over the age of 5-6 years), even though both approaches are effective. Indeed Frizelle et al. (2021) drew a similar conclusion in their review.

Intervention targets

Many intervention studies targeting morphosyntax in children with DLD consider a single target of intervention (e.g., past tense, Calder, Claessen, Ebbels, et al., 2021; causal adverbials, Curran & Owen van Horne, 2019; past tense, Owen van Horne et al., 2017) while others choose particular targets from a small range of targets (e.g., coordinating conjunctions, Ebbels et al., 2014; verb morphology, Fey et al., 2017; Plante et al., 2019). However, different children with DLD may have difficulties with a wide range of grammatical structures, both syntactic and morphological in comprehension and/or expression. Therefore, clinicians need to be able to take an individualized approach to the identification and prioritization of targets for intervention, and to decisions around how much intervention to provide on each target depending on an individual's response to intervention, as well as potentially provide intervention for more than one target at a time.

Dosage

To increase intervention effectiveness and efficiency, we first need to describe interventions in detail, both in terms of the nature of the intervention components as well as aspects of time. The notion of a *teaching episode* (Warren et al., 2007), which may have multiple repetitions during a session, is helpful in this respect. Baker (2012) depicts a *teaching episode* as containing *active ingredients* which can be split into *therapeutic inputs* and *client acts* (see also Figure 1 clinician versus child actions / active ingredients). Once a *teaching episode* for a particular intervention has been defined, the *number of teaching episodes per session* can be counted and the *teaching episode rate* calculated. Warren et al. (2007) describe a framework for considering aspects of intervention dosage using the terms *dose*, *dose form*, *dose frequency*, *total intervention duration*, and *cumulative intervention intensity.* However, these parameters are difficult to apply to interventions when intervention sessions include more than one target and/or more than one type of *teaching episode* per target and/or when the *time focused on intervention targets* is less than the *session duration* (see also Figure 1).

Calculation of some of Warren et al.'s (2007) dosage parameters is also not possible if there are any deviations from the planned *session frequency* (which they call *dose frequency,* e.g., sessions are missed), or *number of teaching episodes per target per session* (e.g., fewer *teaching episodes* may be achieved in early intervention sessions when a target and intervention techniques are being introduced, with more achieved in later sessions; or the *teaching episode rate* may vary with child factors such as variable attention, emotional states, tiredness or health). In addition, if we wish to evaluate progress during intervention, it would be useful to distinguish *cumulative* counts (at different points during the intervention) from *total* counts (when intervention has been completed). Therefore, in this study we consider the effects on progress of *teaching episodes per target per session* (which are affected by the *teaching episode rate* and *session duration*), *cumulative teaching episodes* and *cumulative intervention sessions.* We will also consider the *total number of teaching episodes* and *total number of intervention sessions* required to reach criterion per target.

Research considering dosage relating to language intervention is in its infancy, but some studies have begun to consider these aspects. An observational study of US speech pathology practices in schools considered aspects of time and the details of what occurred within sessions (Schmitt et al., 2017). They found considerable variability in the *total number of sessions* provided for children over an academic year (16 – 154 sessions) which was not explained by child factors, such as severity. They also found children's *time in intervention sessions per week* averaged 36 minutes, with an average *session frequency* of 1.3 sessions per week (thus an average *session duration* of 28 minutes), with a *total time in intervention sessions* for the academic year of 16.9 hours (1,014 minutes). However, observation of video recordings showed that only 49% of this time was focused on language, i.e., the *total intervention time* was approximately half of the *total time in intervention sessions*. Thus, large efficiency gains could potentially be made by increasing this percentage, while still allowing time for building rapport and motivation (see Figure 1). In addition, Schmitt et al. (2017) found no significant

relationship between children's pre-intervention skills and the *total intervention time* received. Moreover, they found that children receiving intervention with *high session frequency* and *low time focused on intervention targets per session*, or *low session frequency* and *high time focused on intervention targets per session* had better general language outcomes than children receiving *high session frequency* and *high time focused on intervention targets per session* or *low session frequency* and *low time focused on intervention targets per session* intervention.

Thus, it seems that the *total time in intervention sessions* in minutes is less of a predictor than the balance between *session frequency* and *time focused on intervention targets within sessions*. Of course, what is missing from Schmitt et al.'s (2017) analysis are intervention targets, the hypothesized *active ingredients,* how they were bundled into *teaching episodes* and the *teaching episode rate* during sessions. In a recent systematic review of quantitative dosage in which some of these measures were considered (Frizelle et al., 2021) a point of 'diminishing returns' was noted, where 'more is not always better'. The authors highlighted that, in general, if there is a high *number of teaching episodes per session*, the *session frequency* can be reduced; and more frequent short sessions may yield similar results to less frequent long sessions, i.e., it may be the *total number of teaching episodes* that is important, rather than their distribution. This can be seen in Table 1, which summarizes the dosage features of previous studies that have manipulated and controlled aspects of dosage in the morphosyntax domain, using real rules and words (as opposed to studies of artificial grammar). For each study, the predictor variables are shown in red and the variables which are held constant are in black. It is worth noting that many of these studies focused only on the actions of clinicians and not child actions when considering the *active ingredients* in *teaching episodes.*

Of these six studies, only Calder, Claessen, Leitão, et al. (2021) found a difference between groups receiving different numbers of *total teaching episodes*, indicating perhaps that this is a crucial variable. The other studies which varied the *total teaching episodes* were Bellon-Harn et al. (2014) and Balthazar and Scott (2018). Bellon-Harn et al. (2014) also varied

the *total intervention period* and only had six children in each group and thus low power. Balthazar and Scott (2018) was more similar to Calder, Claessen, Leitão, et al. (2021) in both method of intervention, number of participants and session frequency, but the participants were much older (10-14 years versus 5-6) and the intervention targets were very different (complex syntax versus regular past tense). However, the lack of a significant group difference in Bellon-Harn et al. (2014) and Balthazar and Scott (2018) suggests that more intervention in terms of both *total intervention period* and *total teaching episodes* is not necessarily better (at least for these targets, participants and intervention methods).

In each of the other three studies in Table 1, *total teaching episodes* was held constant and no differences between groups emerged, even though between these three studies they manipulated different combinations of *number of teaching episodes per session, teaching episode rate, total intervention time, session frequency, session duration, time focused (not) on intervention target, total number of intervention sessions* and *total intervention period.* The findings of Plante et al. (2019) are particularly important as they manipulated the *teaching episode rate during sessions* and therefore *time focused (not) on intervention targets per session* and *total intervention time*, while keeping all other variables constant. Their finding of no difference between the two groups suggests that the same outcomes can be achieved in half the *total intervention time* by doubling the *teaching episode rate* and hence halving the *time focused on intervention targets in sessions*.

Table 1: dosage details of previous studies. Manipulated variables shown in red.

	Plante et al. (2019)	***Meyers-Denman & Plante (2016)***	***Bellon-Harn (2012)***	***Bellon-Harn (2014)***	***Calder et al. (2021b)***	***Balthazar & Scott (2018)***
Number of children	20	16	12	12	29	30
age of children with DLD	4:1-5:11	4;10-5;10	4;0-5;3	4;0-5;11	5;9-6;8	10:10-14:11
target	individually chosen morphological targets (1 per child)	individually chosen morphological targets (1 per child)	semantic and morphosyntactic skills	coordinate and subordinate clauses and number of words in clauses	regular past tense	production of complex sentences (adverbial, object complement, and relative clauses)
Content of teaching episode / active ingredients / techniques	adult enhanced conversational recast	adult enhanced conversational recast	cloze procedures, expansions and models (balance varied by child)	cloze procedures, expansions and models (balance varied by child)	child correct production (following visually supported explicit instruction, modelling, with hierarchical feedback)	verbal explanation, modelling & sentence repetition. Complex sentence manipulation (with decreasing clinician scaffolding and support)
number of teaching episodes per session	24 teaching episodes	8 vs 24 teaching episodes	140 - 260 teaching episodes	140 - 260 teaching episodes	50 teaching episodes	15 modelling & repetition + 15 complex sentence manipulation
teaching episode rate during sessions	1.6 vs 0.8 teaching episodes per min	0.8 teaching episodes per min	7-13 teaching episodes per min	7-13 teaching episodes per min	2 teaching episodes per min	0.5 - 0.75 teaching episodes per min
total teaching episodes	600 teaching episodes	600 teaching episodes	3,360 - 6,240 teaching episodes	3,360-6,240 vs 5,880-10,920 teaching episodes	500 vs 1,000 teaching episodes	270 vs 540 teaching episodes
total intervention time	6.25 vs 12.5 hours	12.5 hours	8 hours	480 vs 840 mins (8 vs 14 hours)	200-300 vs 400-600mins (3.3-5 vs 6.7-10 hours)	360-540 vs 720-1080 (6-9 vs 12-18 hours)
session frequency	1x per day, 5 days per week	3x vs 1x per day, 5 days per week	2x vs 4x per week	4x vs 3x per week	1x vs 2x per week	1x vs 2x per week
session duration	30 mins	10 vs 30 mins	20 mins	20 mins	20-30 mins	40-60 mins
time focused on target per session	15 vs 30 mins	10 vs 30 mins	20 mins	20 mins	20-30 mins	40-60 mins
time not focused on target per session	15 vs 0 mins	not reported	not reported	not reported	not reported	not reported
total number of intervention sessions	25 sessions	75 vs 25 sessions	24 sessions	24 vs 42 sessions	10 vs 20 sessions	9 vs 18 sessions
total time in intervention sessions	12.5 hours	12.5 hours	8 hours	480 vs 840 mins (8 vs 14 hours)	200-300 vs 400-600mins (3.3-5 vs 6.7-10 hours)	360-540 vs 720-1080 (6-9 vs 12-18 hours)
total intervention period	5 weeks	5 weeks	12 vs 6 weeks	6 vs 14 weeks	10 weeks	9 weeks
RESULTS	No difference between conditions	No difference between conditions	No difference between conditions	No difference between conditions	Higher dosage group steeper progress	No difference between conditions

The finding of potentially diminishing returns from increasing amounts of intervention as found by Bellon-Harn et al. (2014), Balthazar and Scott (2018) and Schmitt et al. (2017), but not Calder, Claessen, Leitão, et al. (2021) needs further investigation in studies with greater power, as continuing to provide intervention when a child is no longer making progress is a waste of limited resources. Thus, it is vital that progress is closely monitored in order to maximize the efficiency of intervention, with intervention ceasing if a child has limited progress left to make, or if they have reached a plateau, perhaps because other issues are preventing them from making further progress with the intervention (e.g., phonology affecting morphosyntax progress, or insensitive tests). Of course, if a child is no longer making progress with a particular target, this does not mean that they will not make progress with another target. Thus, it may be that a change of target is required, rather than cessation of intervention altogether (Frizelle et al., 2021b).

Aims of current study and Research Questions

With respect to dosage, in keeping with conclusions from Frizelle et al. (2021) calling for intervention studies for morphosyntax that manipulate within session dosage while controlling for *session frequency*, in our study, we aimed to investigate the effect of manipulating the *number of teaching episodes per target per session* on progress with intervention, while keeping the *total number of teaching episodes per session*, *session duration* and *session frequency* constant. The intervention targets were individually determined morphosyntactic targets, monitored regularly with a new target introduced as soon as a previous target had either been achieved or plateaued.

The intervention method was the SHAPE CODING™ system delivered 1:1 predominantly by a single clinician. In addition to evaluating the effectiveness of this intervention method, we also considered whether there were any differences in progress with intervention between children and targets. Our specific research questions were:

1. Is there an overall effect of using the SHAPE CODING™ system to teach highly individualized morphosyntactic targets to children with DLD?

2. Does the time in each phase (baseline, intervention, maintenance) predict progress and do any of the individual participants show differing patterns or trends?
3. Does the *cumulative number of teaching episodes* predict progress and do any individual participants show differing patterns or trends?
4. Does the *number of teaching episodes per target per session* affect rates of progress with intervention based on *cumulative intervention sessions* and/or *cumulative teaching episodes*?
5. Does the *number of teaching episodes per target per session* predict maintenance of any progress?
6. Do any specific targets show differing patterns or trends?
7. What is the *total number of intervention sessions* or *teaching episodes* required to reach criterion on a target and how much does this vary?

Method

Study Design

Our study was a single blind within-participant multiple baseline design across targets, consisting of three phases for each participant for each target: 1) weekly pre-intervention baseline tests, 2) weekly intervention using the SHAPE CODING™ system with either 10, 20 or 30 *teaching episodes per target per session* (described further below) with weekly probe tests to measure progress towards pre-defined stopping criteria, and 3) a maintenance phase during which retention probes were planned for administration for each completed target, 2, 6 and 14 weeks post-intervention. However, due to the school summer break, we could not gather longer term maintenance data for targets that were introduced and concluded later in the study, so the length between maintenance tests was reduced for these later targets in order to carry out as many maintenance retention tests as possible before the end of the study. Due to the variation in lengths of time between maintenance tests, the number of days in

maintenance was used as the main predictor for any analyses, so that we could see whether scores changed with time after the intervention ceased.

Targets were introduced into baseline and into intervention in a staggered manner such that per participant, we aimed for two targets to always be in baseline and two in intervention with any completed targets in maintenance.

Ethics

The research was approved by the Moor House Ethics Committee, reference number 2021/1/1.

Participants

This intervention study was carried out at a specialist day and residential school for children with (developmental) language disorders in the UK. All children in the school have language skills which significantly affect their learning. The majority meet the criteria for DLD but some children who attend the school have associated biomedical conditions, but language is judged to be the primary barrier to learning. This study focused on children in Key Stage 2, the UK educational stage where children are aged between 7 and 11 years. Eight children (6 boys and 2 girls) from two classes were judged to be suitable for the project, i.e., morphosyntactic difficulties were a priority area for intervention and their parents consented to participation (no parents refused consent). Their ages ranged from 8;0-10:10 at the start of the study. All eight participants had a diagnosis of DLD, with seven also having a diagnosis of speech sound disorder, one of whom additionally had an ADHD diagnosis. All eight participants had English as a main language. One had limited exposure to another language but had attended an English-speaking nursery and previous school setting. We did not index the socio-economic status of our participants as, due to the nature of the ongoing relationships between parents of participants and interventionists, we deemed it unnecessarily intrusive to ask parents for information on levels of education or income.

To profile the participants' language abilities prior to the intervention commencing and to inform the identification of appropriate treatment targets, the following language assessments were

completed 1) the Structured Photographic Expressive Language Test-Third Ed (SPELT-3: Dawson et al., 2005), 2) the narrative subtest from the Assessment of Comprehension and Expression (ACE, Adams, Cooke, Crutchley, Hesketh, & Reeves, 2001) and 3) the 'Beach' Story subtest from the Expression, Reception, and Recall of Narrative Instrument (ERRNI; Bishop, 2004). The 'Beach' story was chosen, as it was found to elicit more complex sentences from typically developing children than the alternative 'Fish' Story (Frizelle et al., 2018). In addition, we also carried out the Test for the Reception of Grammar-Electronic (TROG-E; Bishop, 2003) to measure children's understanding of syntax. Summaries of these scores are shown in Table 2, with individual scores in the Supplemental Information 1. These scores show that all participants had difficulties with both expressive and receptive grammar.

Table 2 - Mean raw and standard scores on standardised assessments pre-intervention

Assessment	Raw		Standard/Scaled	
	Mean	SD	Mean	SD
SPELT	28.12	10.47	60.57[a]	21.62
ACE grammar	6.25	2.38	3.38[b]	0.74
ACE information	8.88	4.55	6.00[b]	2.88
ERRNI MLU	5.51	1.20	69.25[a]	8.71
ERRNI content	8.62	2.50	70.38[a]	4.60
TROG	7.12	3.31	65.00[a]	10.92

[a] = standard score (1SD range 85-115) [b] = scaled score (1SD range: 7-13)

Measures and Materials

Intervention Target Identification

Following transcription of the language samples gathered from the assessment battery described above, the samples were analyzed against a list of 133 potential morphosyntactic targets split into 11 target groups: main clauses, questions, conjoining, adverbials, complement clauses, passives,

relative clauses, tense and aspect, subject verb agreement, noun phrases and negatives. Each target has a code which indicates its target group and its relative position within that target group (e.g., NP9 is the 9th structure/target in the Noun Phrase target group). Performance for each child on the language samples was recorded on a copy of the target identification spreadsheet developed by the first and third authors (an updated version following changes due to learning from this project is available here: https://shapecoding.com/resources/grammar-spreadsheet/. This webpage includes a video demonstration of the target identification process detailed below). The first stage identified potential targets for further testing on probe tests. Formulae on the spreadsheet highlight when a target form is not used correctly at least twice, or if it is attempted but this results in errors for more than 25% of the production attempts. Such targets are then tested further using a probe test (see below)[1].

Probe testing

Probe tests were administered for those potential targets with the highest priority in the suggested hierarchy on the target identification spreadsheet. The third, sixth and first authors (listed in order of contribution levels) created probe tests for all potential targets consisting of 10 stimulus pictures (plus one model and one practice item) presented in PowerPoint slides designed to elicit oral attempts at each target form. For examples of three probe tests, see Supplemental Information 2. Where a participant's production accuracy fell below 90% on the probe test, that target was identified for intervention. However, if performance rose above 90% during the baseline phase, the target was dropped from the project before intervention started.

Probes were administered once a week during the baseline and intervention phases and less frequently during the maintenance phase by the sixth author who was blind to whether each target for each participant was in the baseline, intervention or maintenance phase. Probe sessions were always on a different day from intervention sessions[2]. The probe tests aimed to establish the participants' baseline ability and then any change with intervention in their ability to generalize intervention targets

to untreated contexts (both linguistic contexts and with a different adult), using lexical items (where possible) that were not used in intervention.

All probe tests were scored by the sixth and first authors with regular discussions and recording of scoring criteria to ensure consistent scoring between the two scorers[3].

Pre-treatment baseline phase

Following the identification of intervention targets, four weekly baseline measurements were administered for the top four priority targets for each participant. Intervention then started for the first two targets and baseline measures continued for the next two targets (unless they had already reached criterion and were dropped from the study). Each time a participant reached a stopping criterion for a target (see "stopping criteria" section), intervention ceased for that target, a new target was introduced from baseline into intervention, and baseline testing began for the next target in the hierarchy. The aim was to have at least four baseline test points for each target before intervention began on that target. In most cases, the baseline was longer, but its length varied according to the speed with which each participant reached the stopping criterion for the previous target. In a couple of cases, a new target was needed before the next target had a baseline with at least four timepoints, in this case, intervention was only provided on one target until the next target had at least four baseline timepoints.

Intervention

All participants received intervention during weekly 30-min 1:1 intervention sessions over 20 weeks using the SHAPE CODING™ system with a pre-determined protocol for each target with the following broad steps (see Figure 2, more details are provided below): 1) Introduce the target structure with a *SHAPE CODING template* and an *explicit rule* (see Appendix A for the *SHAPE CODING templates*, rules and examples for all targeted structures); 2) Child and adult produce the target structure with *SHAPE CODING template* and then, after five consecutive correct productions, without the *SHAPE CODING template* (the template was brought back if two consecutive errors were then made); and 3) Error detection.

Study Procedure

To reflect the school timetable, intervention was given in approximate 10-week blocks over the course of two terms: January – April 2022 and April – July 2022. Intervention was delivered by the second author, an SLP working in the school, who had 3 years' experience using the SHAPE CODING system and had successfully completed both the Part 1 and Part 2 SHAPE CODING courses[4].

Teachers were blind to which specific targets had been assigned to each participant and when they were being targeted; they were asked to support language in a more general way in the classroom, rather than focusing on specific morphosyntactic targets. Teachers were informed of the specific targets on completion of the maintenance assessment phase. The participants did not receive any additional SLP intervention targeting expressive grammar.

Intervention sessions took place individually in a room within the school. Where possible, students received intervention on the same day and at a similar time each week, however this varied marginally, as sessions were conducted on a fortnightly timetable in order to allow students to attend the greatest possible range of school lessons.

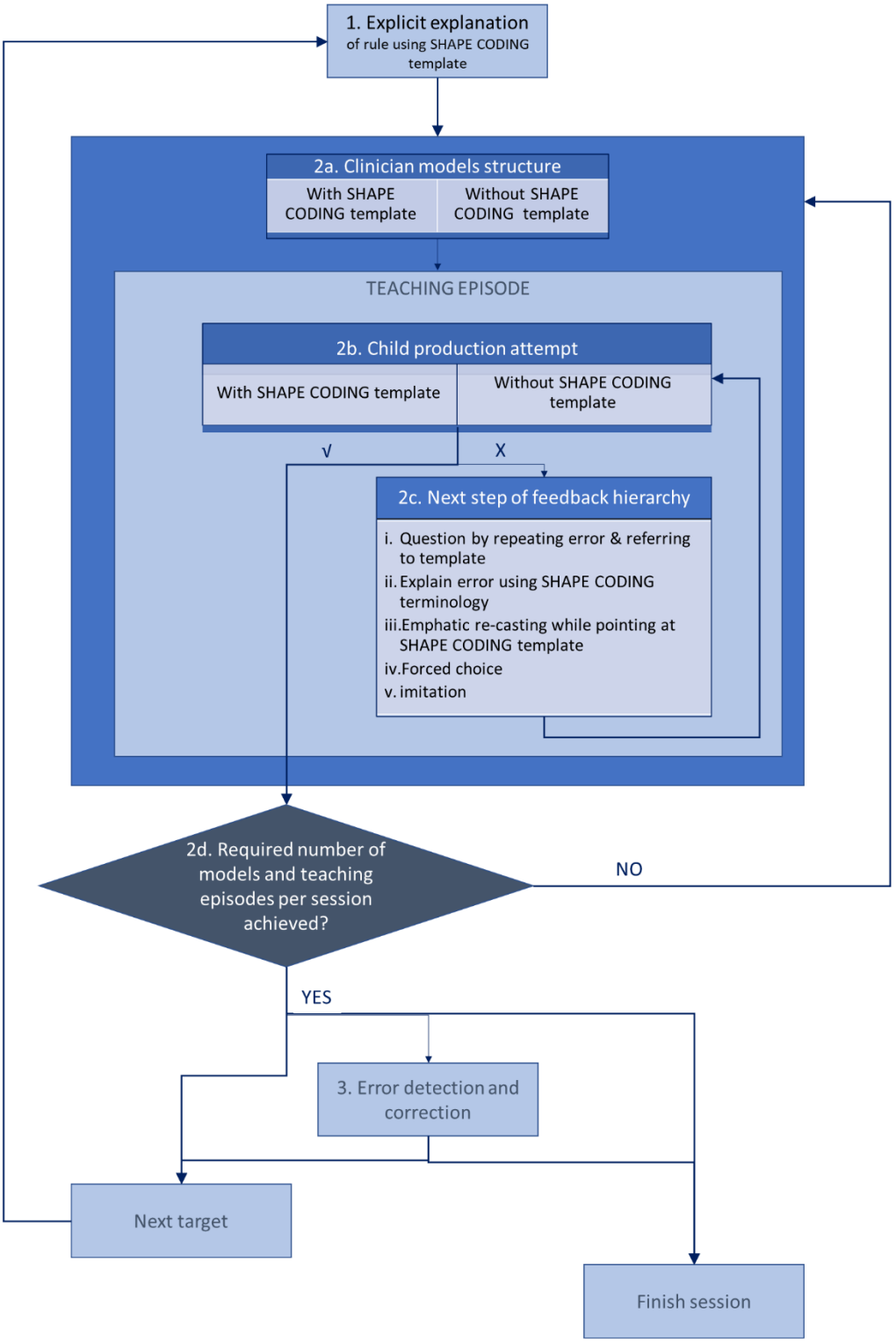


Figure 2 – Summary of intervention steps and techniques showing how the ***techniques*** *were combined into* ***teaching episodes****.*

At least half of the verbs / nouns in the probe tests were not used in the intervention. We also avoided using any of the exact probe sentences (with matching nouns and verbs) in the intervention.

Below, we use the framework developed by (Frizelle et al., 2021) to unpack the intervention components involved in the *dose form* to include the *intervention context*, *method of instruction*, *intervention techniques* and *procedures*.

Intervention context

The SHAPE CODING™ system set the overall context for the intervention and was implemented using activities such as looking at pictures and videos or acting out scenarios using toy animals. The system uses specific visual cues to target aspects of language known to be challenging for children and adolescents with language disorders, including colors for word classes, shapes for phrases and their roles and positions within sentences, arrows to depict tense and single versus double lines for singular versus plural.

In keeping with Plante et al., (2014) we aimed for a degree of variability in the linguistic input (linguistic context). When the target was morphological, the SLP always used 10 linguistically unique utterances per session with a different root word, regardless of the allocated *number of teaching episodes per session*. This ensured that the within session variability was constant for the three dosages. There was no restriction on using similar root words between sessions.

In relation to syntactic goals, verb variability was determined by the specific target. Given the nature of syntactic structures, it was not always possible to find 10 unique verbs to represent a particular structure. For example, there are a limited number of verbs that take three obligatory arguments (e.g. *the boy put the cat in the basket*). For such targets, using 10 unique verbs was not possible. However, we endeavored to ensure a high level of noun variation for all targets.

Method of Instruction

The treating SLP used *explicit explanation* of the target rule, supported by *visual templates* from the SHAPE CODING system (see Appendix A). Broad intervention steps for each target were

developed by the third and first authors and made available to the treating SLP (second author), see here: https://shapecoding.com/intervention-steps/. *Explicit methods* of instruction were used in conjunction with implicit techniques such as *modelling*.

Intervention Techniques, Procedures and Feedback Hierarchy

The precise techniques, procedures and feedback hierarchy used across all targets are summarized in Figure 2 and are described here: https://shapecoding.com/general-techniques/ (although this study focused on production teaching episodes and thus only included the adult model portion of intervention step 2a, as in Figure 2).

Each participant worked on two intervention targets per session. The order of focus on the two targets alternated from session to session. For some participants both targets were morphological, for others both were syntactic and for others the targets were mixed.

Teaching Episodes and Dosage

A *teaching episode* in this study included intervention steps 2b and 2c and could include some or all of the *active ingredients/techniques* in those steps, depending on the child's skill level and the stage of intervention. For the purposes of our analyses, we deemed an individual *teaching episode* to be complete when a child had correctly said the target structure (regardless of the level of visual support provided by the *SHAPE CODING templates* or the level of the feedback hierarchy needed to achieve this). Thus, our dosage manipulation used the number of correct child productions as our measure of the *number of teaching episodes per session*. In addition, we also controlled the total number of adult models of the target structure per session to match the number of *teaching episodes* to remove this as a potential confound[5]. This does not mean that we do not consider the adult models to be part of the teaching; indeed, we would argue that our intervention includes two different types of *teaching episodes*: adult models and child supported production attempts until successful. We aimed for these to be of equal number so our analyses would be simplified by only using one type of *teaching episode* as the predictor (in effect as if 2a were part of our *teaching episode*). However, where the number of

correct child and adult productions differed, we used the number of child correct productions as our predictor. We leave it to future studies to examine whether ratios between adult models and child productions other than 1:1 are more efficient.

In each session, we aimed for a total of 40 *teaching episodes* shared between two targets, either 20:20, or 10:30. The participants were randomly assigned to one of two groups. In the first intervention term, four participants received 20 *teaching episodes* for each of the two targets per session and the other four participants received 10 *teaching episodes* for one target and 30 for the other. The participants were randomly assigned as to whether their odd or even targets received 30 or 10 *teaching episodes*. Pilot work indicated that at least 40 *teaching episodes* per 30-minute session was achievable for SLPs to administer and this is also in keeping with work carried out by Calder, Claessen, Ebbels, and Leitão (2020, 2021) which involved 50 *teaching episodes* in a 30-minute session on a single target. In the second intervention term, the allocation was reversed, such that those who had been assigned to the 20:20 condition were now assigned to the 10:30 condition and vice versa. This protocol allowed for all children to work on different targets at either 10, 20 or 30 *teaching episodes per target per session*, while keeping *session duration* constant at 30 minutes.

The planned *session frequency* was once weekly with the *total time in intervention sessions* being 10 hours. The *total intervention period* was two school terms, which covered a period of 27 weeks[6].

Therefore, we aimed to keep the following constant: *teaching episode rate during sessions*, *session frequency*, *session duration*, *total number of intervention sessions*, *total time in intervention* and *total intervention period*, while manipulating the *number of teaching episodes per target per session.* In order to maximize intervention efficiency, each child worked on each target to a pre-specified criterion (see below). This means that the following dosage parameters varied according to individual response to intervention: *total number of sessions per target, total intervention time per target, total teaching episodes per target*.

During sessions the SLP kept track of the *number of teaching episodes per target per session* using a pen and paper. In addition, she also tracked the *time focused on each target per session*. In the 10:30 condition the treating SLP aimed to complete 10 *teaching episodes* in 8 minutes and 30 teaching episodes in 22 minutes. In the case that it proved difficult to reach the required number of *teaching episodes* with a given child (perhaps due to attention and/or behavior), the treating SLP aimed for the *number of teaching episodes* of the two targets to be in the desired ratio, even if the actual number required was not achieved. This ensured that the *teaching episode rate* was roughly equivalent across all conditions. In the 20:20 condition, the treating SLP aimed to complete 20 teaching episodes in 15 minutes.

In cases where a child was absent and missed an intervention session, they were not tested on those targets until they had an intervention session so that the *cumulative teaching episodes* and *cumulative intervention sessions* increased between each test point during the intervention phase. This sometimes meant a gap longer than one week between testing points if sessions were missed, or a holiday intervened. However, if the child was available for the testing, but not the intervention, testing of targets in baseline continued, in order to extend the baseline.

Stopping criteria

We implemented two stopping criteria throughout the intervention. The first criterion (*target achieved*) was to identify when participants had made sufficient progress on a given target that they no longer required intervention on that target. 90% production accuracy was deemed an appropriate indication of this.

The second criterion (*target discontinued*) was for those participants who were not, or no longer, making progress on a target. There are obvious ethical issues in continuing to focus on a target with little or no (continued) progress when a child could benefit from input on a different target. The minimum *cumulative teaching episodes per target* before the discontinue target rule was applied was 180 teaching episodes. This value was based on the average number of teaching episodes required to

observe progress (and needed to be divisible by 10, 20 and 30 to accommodate the three possible dosages). For each participant and target, we were then able to run a linear regression to examine whether the proportion correct on that target changed significantly as a function of cumulative teaching episodes (once the minimum of 180 teaching episodes had been reached). If the outcome was non-significant, intervention was discontinued for that target.

All targets were included in the analyses below, including those which were still in progress at the end of the study and those which were discontinued due to limited progress.

Participants' attention in sessions and intervention fidelity

All sessions were video recorded and 20% (26 sessions) were randomly selected for analysis by the fourth author, who examined the participants' attention within sessions and adherence to intervention procedures and techniques (i.e., the planned versus received dosage; uniqueness of lexical items; and the ordering of the intervention steps and feedback hierarchy)

Participants' attention in sessions

Participants' attention in sessions was a potential moderating factor, which could affect the delivery and outcomes of intervention (Carroll et al., 2007). To explore this, we used the Attention subscale of the Pivotal Behaviour Rating Scale (Revised; Mahoney & Wheedon 1999), which rates children's attention in four different areas: (1) their general ability to focus on an activity; (2) their level of persistence within tasks; (3) their active involvement in the session and (4) their compliance with adult suggestions. Each area is rated out of five, where 1=very low, 2=low, 3=moderate, 4=high and 5=very high and an average score for attention is calculated.

Mean attention scores for individual participants across analyzed sessions are shown in Table 3. On average, their scores were high, but there was variability within the sample.

ID6 showed the lowest average attention score (but was not the child with the ADHD diagnosis). He scored in the "moderate" range, indicating that there were sustained periods during which he appeared focused on what he was doing and other extended periods in which he seemed

disinterested in the activities (Mahoney & Wheedon, 1999). In contrast, ID7's attention was rated as "very high" across all activities and sessions, reflecting that this participant was highly involved throughout the observed sessions. ID5's scores fell between the "moderate" and "high" range. All remaining participants were rated as showing a "high" to "very high" level of engagement, staying focused on the activities and appearing to derive satisfaction from their participation.

Table 3 – Mean attention scores in rated video sessions and intervention sessions received

	Participant								
	ID1	ID2	ID3	ID4	ID5	ID6	ID7	ID8	**Mean**
Mean attention score across videoed sessions	4.92	4.40	4.44	4.92	3.96	3.13	5	4	**4.34**
Total number of intervention sessions received	16	17	19	20	17	20	20	17	**18.25**

Planned versus received dosage

The *total number of intervention sessions* received per participant ranged from 16 to 20 sessions out of the 20 planned sessions (see Table 3). Missed sessions were mainly due to high levels of absence as a result of sickness during 2022 among both adults and children in UK schools.

A high level of implementation fidelity was achieved for aspects of the intervention related to number of *teaching episodes*. On average, 96.4% of the planned child correct productions were achieved. In addition, the SLP produced 97.2% of planned adult *models*. Inter-rater reliability for treating SLP vs researcher-recorded child correct productions and adult *models* was assessed using a two-way mixed, absolute agreement, single-measures intraclass correlation coefficient (ICC; Hallgren, 2012) using SPSS statistical package version 27. The resulting ICC was in the excellent range; ICC = 0.967 (0.951 – 0.978) indicating a high level of agreement across raters, with minimal measurement error. This shows that tracking child and adult productions with a high degree of accuracy live during intervention sessions is feasible.

The average time spent per session on focused intervention activities was 23min 1s, compared to the planned 30 minutes. However, within each session, the proportion of time allocated to each of the two intervention targets matched closely with the planned proportion. In the 20:20 *teaching episodes* condition, this was exactly as planned with a time ratio of 50:50. In the 10:30 *teaching episodes* condition, time was allocated at a ratio of 30:70, compared to a planned allocation of 25:75. This reflected the time needed to lay out the *SHAPE CODING templates*, introduce and model each target before eliciting child productions (intervention step 1). These values were used to calculate the achieved dosage details shown in Table 4.

Of the average 6 min 59s per allocated 30-minute session which were not focused directly on intervention targets, this included time taken to collect the child from lessons and time spent on rewards, such as short games, at the end of the session. These activities were not video-recorded and therefore it is not possible to provide a detailed breakdown of their relative time weightings. The section of the intervention sessions which was videoed lasted an average of 24 min 38s. Within this, time not focused on targets included time a) presenting a general introduction to the session, including talking through a visual timetable (mean = 41.6s), b) lost to child distractions, e.g., talking about topics unrelated to the intervention (mean = 33.7s), c) transitioning between intervention targets, including collecting and organizing intervention materials (mean = 16.8s), d) lost to technical problems (mean = 3.2s) and e) other one-off interruptions including the student getting stuck in his chair and a staff member entering to look for another student. A summary of timings for activities within video recorded sessions is provided in Supplemental Information 3.

Table 4: Planned and achieved dosage details. Dosage manipulations are shown in red

	Planned			Achieved		
number of children	8					
age of children with DLD	8;0-10;10					
target	individually determined morphosyntactic targets					
Content of teaching episode / active ingredients / techniques	concluded with child accurate production (potential active ingredients: explicit explanation, modelling, visual templates, elicitation, hierarchical feedback, error detection and correction)					
number of teaching episodes per target per session	10 teaching episodes	20 teaching episodes	30 teaching episodes	9.6 teaching episodes	19.3 teaching episodes	28.9 teaching episodes
teaching episode rate during sessions	1.33 teaching episodes per min			1.4 teaching episodes per min	1.7 teaching episodes per min	1.8 teaching episodes per min
total teaching episodes per target	varies according to response to intervention					
total intervention time per target	varies according to response to intervention					
session frequency	1x per week (minus absences and holidays)					
session duration	30 mins					
time focused on target 1 per session	8 mins	15 mins	22 mins	6 mins 54s	11 mins 30s	16 mins 6s
time focused on target 2 per session	22 mins	15 mins	8 mins	16 mins 6s	11 mins 30s	6 mins 54s
time not focused on targets per session	0 mins			6 mins 59s		
total number of intervention sessions per target	varies according to response to intervention					
total number of intervention sessions	20 sessions			16-20 sessions		
total time in intervention sessions per target	varies according to response to intervention					
total time in intervention sessions	10 hours			8-10 hours		
total intervention period	27 weeks					

Ordering of intervention steps and feedback hierarchy

Video analysis showed 100% adherence to the ordering of steps for intervention. However, although the intervention protocol specified the withdrawal of the *SHAPE CODING templates* after five consecutive correct child productions, about one third of the time the SLP left templates visible for longer than specified, to provide additional support for participants practicing new structures. The final stage of the intervention protocol (intervention step 3, *error detection*) was only evidenced in nine of the 26 analyzed videos. This technique was employed with IDs 1,2,3,4 and 7, depending on the child's progress with an individual target but not with the remaining participants (IDs 5 and 6), indicating that the SLP did not judge them to be ready for this metacognitive stage of intervention or there was insufficient time, during the videoed sessions. This may also have been related to attention and motivation as these two participants had relatively lower attention scores (see Table 3).

In relation to the order of the pre-specified feedback hierarchy following incorrect child productions, on a total of 67 occasions across the analyzed dataset, the SLP started with feedback (ii) *explicit explanation* (e.g., "I didn't hear the diamond word), rather than feedback (i), *question by repeating the child's error with a questioning tone*. Thus, these two steps were sometimes used interchangeably.

Table 5 shows the total percentage of feedback episodes for each child which resulted in a correct child production. For the vast majority of cases (82%), the feedback hierarchy was not required, as the child production (often supported by a *SHAPE CODING template*) was correct. However, the number of incorrect productions requiring some level of feedback varied markedly by individual. While ID7 required only 7 instances of feedback out of a total 149 correct productions (feedback rate of 4.7%), IDs 1 and 2 each required feedback on over 31% of their productions (38 out of 120 for ID1 and 36 out of 113 for ID2).

Table 5: Percentage instances of each feedback type per child, compared to total correct child productions and percentage of all attempted teaching episodes not completed, analysed across 26 intervention sessions.

Step in feedback hierarchy	ID 1	ID2	ID3	ID4	ID5	ID6	ID7	ID8	**Mean**
No feedback required	68.3	68.1	87.5	86.8	86.7	78.9	95.3	82.8	**81.8**
(i) Question by repeating child's error	19.2	14.2	5.6	2.4	0.7	2.9	2.7	9.1	**7.1**
(ii) Explain error	9.2	17.7	4.4	9.6	11.6	16.4	2.0	7.1	**9.7**
(iii) Emphatic recasting	3.3	-	2.5	1.2	-	1.9	-	1.0	**2.0**
(iv) Forced choice	-	-	-	-	-	-	-	-	**-**
(v) Imitation	-	-	-	-	-	-	-	-	**-**
% teaching episodes not completed	0.8	0.9	0.0	0.0	0.7	**6.3**	0.0	1.0	**1.2**

The most common step of the feedback hierarchy which resulted in an accurate child production was feedback (ii): *explicitly explain the error using SHAPE CODING terminology*. The next most frequent step was feedback (i): *question by repeating the child's error with a questioning intonation*. Finally, there were just 12 examples (2%) of feedback (iii): *emphatic recasting with emphasis on the error* within the analyzed dataset. Feedbacks (iv) and (v) (*forced choice* and *imitation*) were not needed, as in almost all cases, the child had already produced a correct attempt at the target before these steps.

There were a small number of instances where the *teaching episode* was not completed (i.e., no correct child production was achieved) and where the SLP produced the target herself, without asking the child to repeat her model. This occurred on only one occasion for IDs 1,2,5 and 8 and was not observed in any of the videos analyzed for IDs 3,4 or 7. However, it is notable that for ID6, 6% of *teaching episodes* in his observed sessions were incomplete (i.e.,

the feedback hierarchy was not continued until a correct child production was obtained). This was evidently related to ID6's relatively poorer attention control (see Table 3). Thus, maintaining his attention by moving on to the next *teaching episode* was higher priority than persisting with repeated attempts at correct production to complete the *teaching episode* and risking losing his attention and motivation.

Variability and uniqueness of lexical items

The fidelity of this appears in Supplemental Information 4.

Data analysis

Data and code for replicating all analyses are available at https://osf.io/32ubp/?view_only=f1ee0fb2a07a4b3198d1b051fa0056b5. All statistical analyses were performed using R Statistical Software version 4.3.0 using the glmmTMB package (version 1.1.7) (R Core Team, 2018). This package fits generalized linear mixed models to a range of statistical distributions, including the binomial distribution of data with binary responses (e.g., correct vs incorrect). This accounts for random effects, such as differences between individuals, dosage or targets, as well as the nested structure of repeated measures data (Brown, 2021).

We split our analyses into seven research questions. Combining across research questions was not possible or desirable because the models would have been too complex, leading to difficulties with both convergence and interpretation. However, because we carried out 14 analyses in total on the same data set, we adjusted the *p*-value we accepted as significant for all analyses to $p = 0.05/14 = 0.004$.

Results

The raw data from all participants are shown in Figures 3 and 4. Each column shows all data points from one participant and each plot shows an individual target for an individual participant. The first number in each title is the order of introduction of the targets, the

combined letter and number code indicates the specific target (e.g., NG2 = second target in the negatives group) and the final number indicates the dosage (10, 20 or 30 *teaching episodes per target per session*). The x-axis shows days since the start of the project and thus the staggering of targets across time can be clearly seen. Visual inspection of these graphs shows much variation both between and within participants and targets. However, in general, it seems that scores on targets are lowest during the baseline phase (in red), increase during their intervention phase (in green) and are higher in the maintenance phase (in blue) than the baseline phase. Given the complexity of the data, for the statistical analyses in Research Questions 1 and 2, we collapsed the data into the three phases: baseline, intervention and maintenance and ignored the staggering of targets over time. For Research Questions 3-7, we collapsed all three phases by considering *cumulative teaching episodes* or *cumulative intervention sessions* (which are at zero during the baseline phase and do not increase further during the maintenance phase).

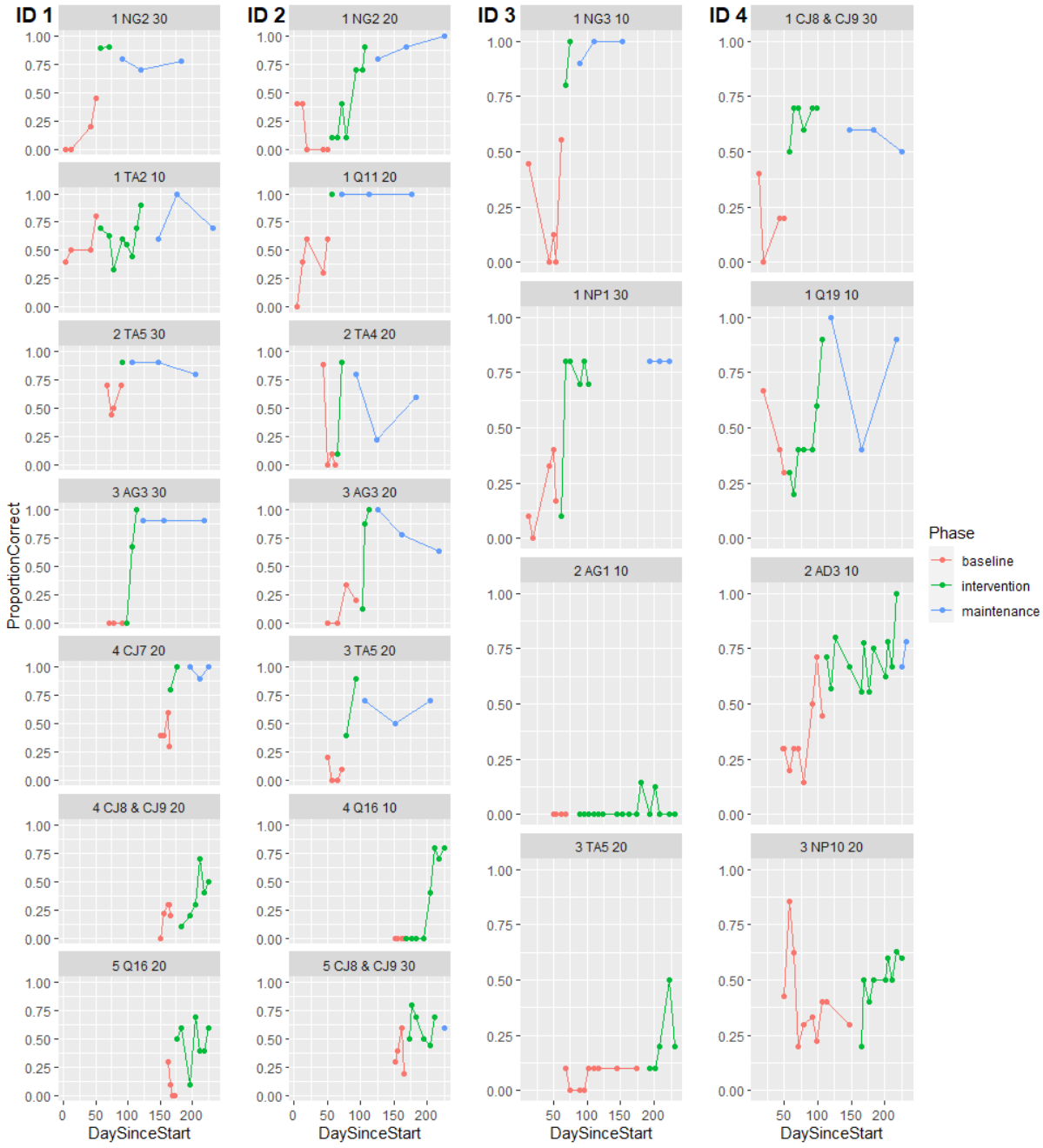


Figure 3 – Repeated data for participants 1-4 showing multiple baseline across targets

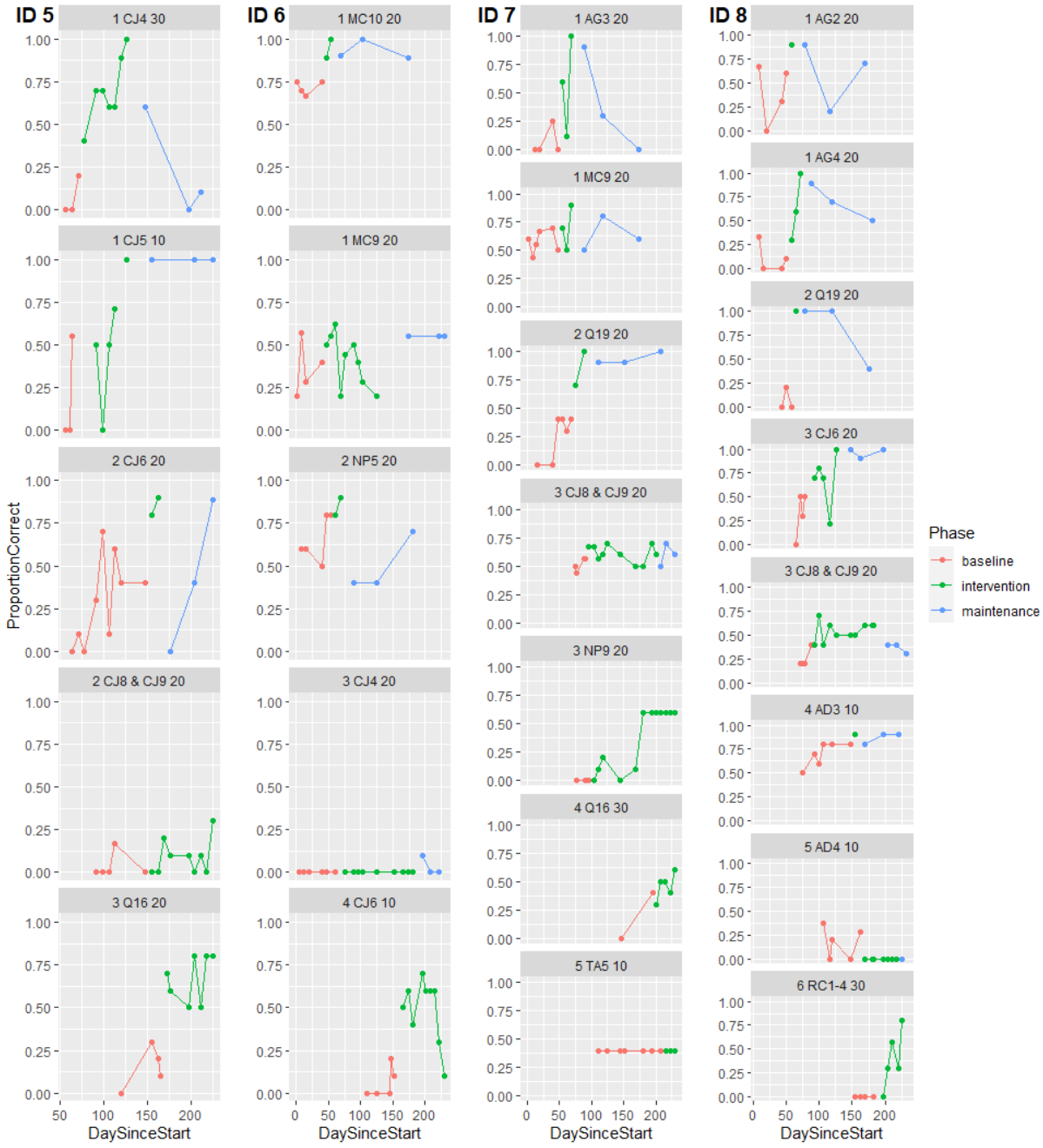


Figure 4 – Repeated data for participants 5-8 showing multiple baseline across targets

Research Question 1 – Is there an overall effect of using the SHAPE CODING™ system to teach highly individualized morphosyntactic targets to children with DLD?

In order to establish whether there was an overall effect of intervention, we first compared scores pre-intervention, during the baseline phase (*n*=222, mean = 0.27, SD = 0.25) with those post-intervention, during the maintenance phase (*n*=97, mean = 0.69, SD = 0.29), an effect size of *d*=1.6 (CI=1.36-1.90). These data are shown in the violin and box plot in Figure 5. To analyze the significance of this effect, we ran a generalized linear mixed model with the fixed effect of phase (baseline vs. maintenance), combining across all datapoints within these two phases (so ignoring time within the phases), but including random intercepts of participant, dosage group and target (this was the random effects structure that explained the most variance and converged). This model showed a significant overall effect of phase where relative to the baseline phase, the odds of a correct response was significantly higher in the maintenance phase, *OR*: 6.78 (*CI*: 5.56-8.27), *p*<.001, indicating the general effectiveness of the intervention.

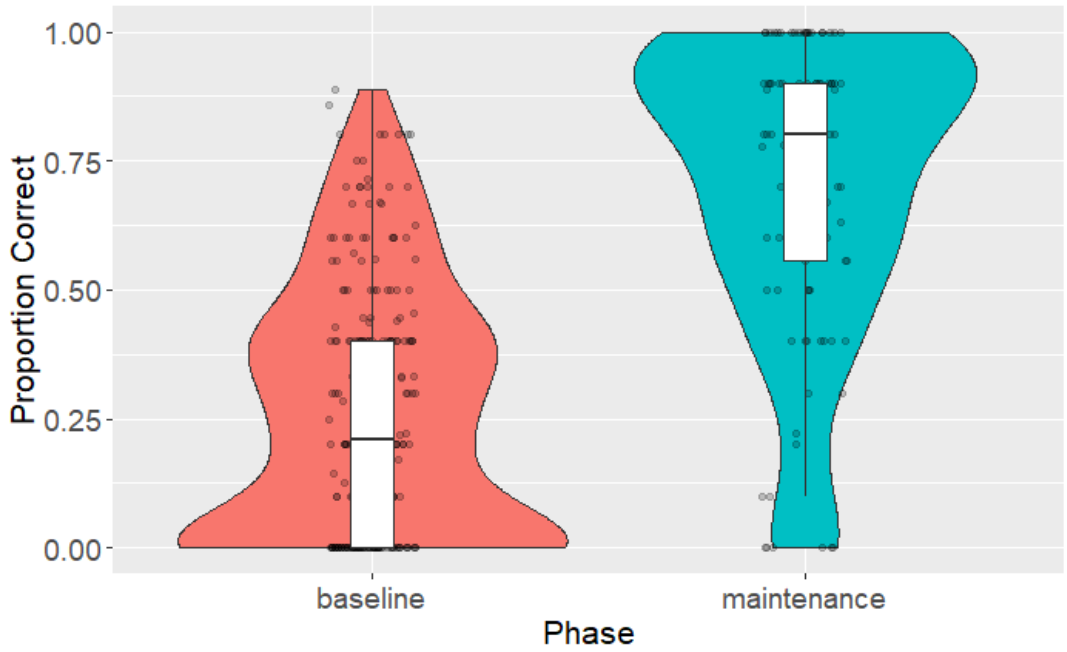


Figure 5: Violin and box plot showing the proportion correct during the baseline and maintenance phases.

Research Question 2 – Does the time in each phase (baseline, intervention, maintenance) predict progress and do any individual participants show differing patterns or trends?

The analysis in Research Question 1 ignored the effect of time within phases and included participants as a random factor. This means that the performance of individual participants could not be evaluated and any change in performance over time within a phase was masked. In the following analysis, we specifically consider the fixed effects of participant and days within each phase and the interaction between them, to establish whether all eight participants responded to the intervention in a similar manner. Thus, we included three fixed effects: participant (sum-coded i.e., comparing each to the grand mean rather than to a reference participant), days in phase, and their interaction. In this analysis, we were not interested in the effect of the individual targets (see Research Question 6) or dosage (see Research Questions 4 and 5). Thus, target, dosage group and the interaction of target with days in phase were included as potential random effects.

In order to analyze each phase in detail, we considered each of the three phases separately. This had the advantage that some data points could appear in more than one phase: i.e., the last data point of the baseline phase was included as the intercept for the intervention phase and the last data point of the intervention phase was included as the intercept for the maintenance phase. For the intervention and maintenance phases, the best model in terms of random effects included intercept for dosage group and by-target slopes with days in phase. For the baseline phase, the by-target slope with days in phase was no better than the intercept for target. However, for consistency, the baseline phase was analyzed with the same random factors as the intervention and maintenance phases. The full model which included the fixed effects of days in phase, participant and their interaction, improved the model fit for all phases. The results are shown in Table 6.

Table 6: Model for each phase separately with fixed factors of days in phase and participant (sum-coded), significant results (p<.004) shown in bold.

	Baseline Phase			**Intervention Phase**			**Maintenance Phase**		
Predictors	*Odds Ratios*	*95% Confidence Interval*	*p*-value	*Odds Ratios*	*95% Confidence Interval*	*p*-value	*Odds Ratios*	*95% Confidence Interval*	*p*-value
Days in Phase	1.014	1.003 – 1.024	0.010	**1.083**	**1.033 – 1.134**	**0.001**	0.993	0.984 – 1.002	0.136
Days in Phase * ID1	1.017	0.992 – 1.043	0.188	1.010	0.992 – 1.029	0.285	1.007	0.992 – 1.023	0.365
Days in Phase * ID2	0.983	0.958 – 1.010	0.209	**1.035**	**1.015 – 1.056**	**0.001**	1.009	0.994 – 1.024	0.253
Days in Phase * ID3	1.002	0.982 – 1.023	0.822	0.978	0.924 – 1.034	0.432	1.026	0.986 – 1.068	0.210
Days in Phase * ID4	0.986	0.968 – 1.004	0.123	1.028	1.002 – 1.054	0.035	0.993	0.976 – 1.011	0.443
Days in Phase * ID5	1.019	0.997 – 1.041	0.085	0.997	0.983 – 1.011	0.672	0.983	0.962 – 1.004	0.103
Days in Phase * ID6	0.994	0.970 – 1.020	0.670	**0.927**	**0.903 – 0.952**	**<0.001**	1.013	0.996 – 1.031	0.122
Days in Phase * ID7	0.995	0.980 – 1.010	0.517	1.014	1.001 – 1.028	0.031	0.986	0.973 – 0.998	0.027
Days in Phase * ID8	1.004	0.985 – 1.024	0.681	1.015	0.101 – 1.030	0.032	0.984	0.969 – 0.999	0.037
Random Effects									
σ^2	3.29			3.29			3.29		
τ_{00}	0.05 $_{\text{Dosage Group}}$			0.53 $_{\text{Dosage Group}}$			0.40 $_{\text{Dosage Group}}$		
	2.03 $_{\text{Target}}$			2.47 $_{\text{Target}}$			3.67 $_{\text{Target}}$		
τ_{11}	0.00 $_{\text{Target. Days in Phase}}$			0.01 $_{\text{Target. Days in Phase}}$			0.00 $_{\text{Target. Days in Phase}}$		
ρ_{01}	-0.25 $_{\text{Target}}$			0.11 $_{\text{Target}}$			0.07 $_{\text{Target}}$		
ICC	0.41			0.86			0.59		
N	3 $_{\text{Dosage Group}}$			3 $_{\text{Dosage Group}}$			3 $_{\text{Dosage Group}}$		
	26 $_{\text{Target}}$			26 $_{\text{Target}}$			21 $_{\text{Target}}$		
Observations	222			300			131		
Marginal R^2 / Conditional R^2	0.088 / 0.464			0.235 / 0.896			0.146 / 0.646		

Table 6 shows no significant change during the baseline phase (when using the adjusted alpha level of p=.004). During the intervention phase, there was a highly significant increase in scores on target probes with days in intervention in the grand mean across all participants, p=.001. However, two participants differed significantly from the grand mean – ID2 made significantly greater progress and ID6 made significantly less progress (investigated further below). During the maintenance phase there was no significant change or interaction. Thus, progress was maintained.

Given that ID6 had a significantly shallower slope during the intervention phase than the other participants, we analyzed progress of just this participant in the intervention phase. A mixed model analysis included a fixed factor of days since the start of intervention and random intercepts for dosage and target (the model did not converge with by-target slope as a random variable). This analysis showed no significant increase with time during the intervention phase for participant ID6, OR = 0.99 (CI=0.98-1.01), p= .34. Thus, the intervention does not seem to have been effective for this participant.

Research Question 3 – Does the *cumulative number of teaching episodes* predict progress and do any individual participants show differing patterns or trends?

Dosage was manipulated within the study to try to establish whether differing numbers of *teaching episodes per sess*ion affect the effectiveness or efficiency of the intervention (see Research Questions 4 and 5). However, the differing dosage (10, 20 or 30 *teaching episodes per session*) means that after a given number of intervention sessions, some targets had received three times as many *teaching episodes* as others. Therefore, our next analysis uses *cumulative teaching episodes* as the predictor variable; this is the total number of *teaching episodes* delivered by each testing point. This predictor also has the advantage of including the baseline, intervention and maintenance phases in the same analysis.

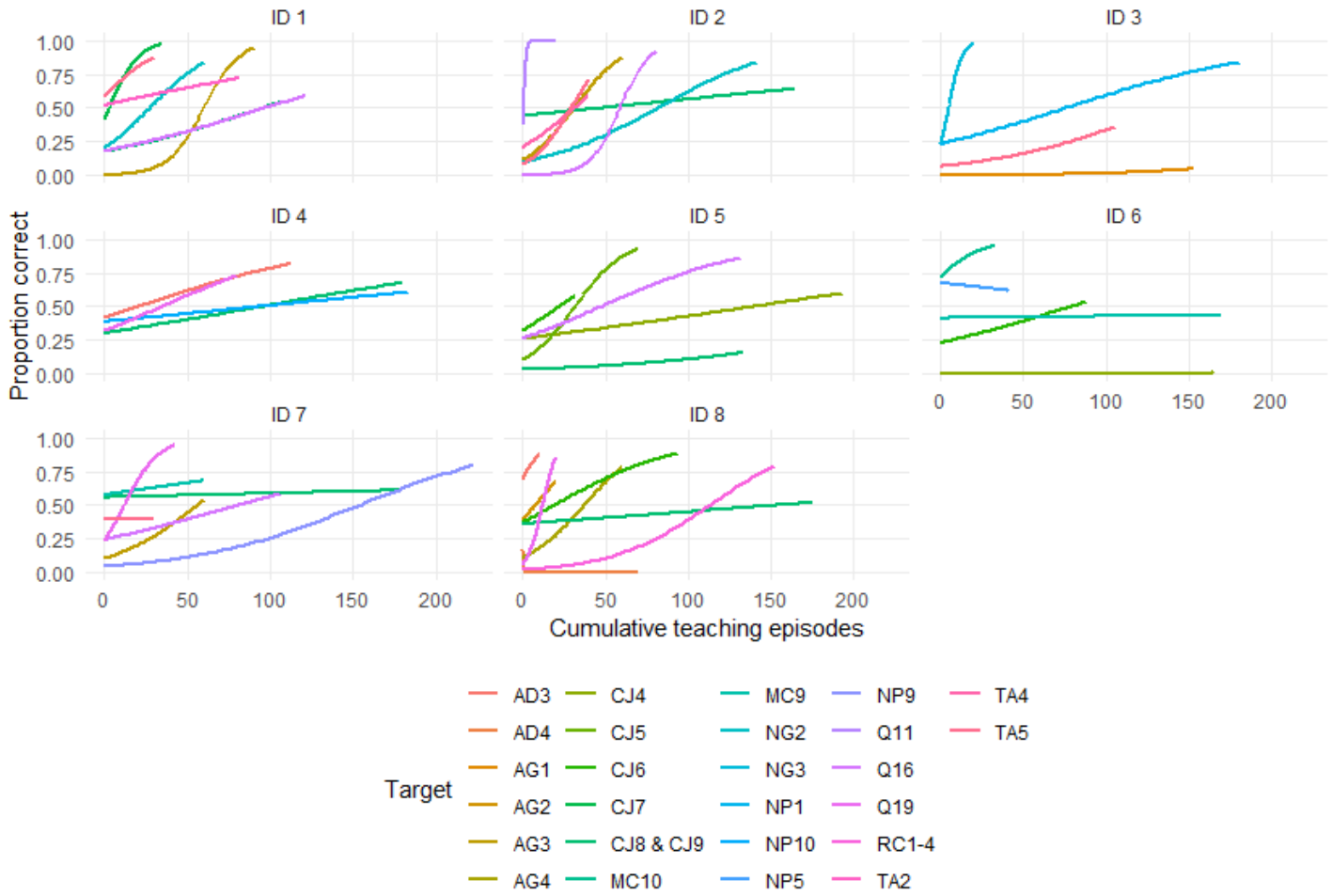


Figure 6 – Proportion correct by ***cumulative teaching episodes*** *on individual targets (separate lines for each target, see Appendix A for explanations of target codes, more details and examples).*

Figure 6 shows the proportion correct for each participant and target with *cumulative teaching episodes* on the x-axis. As each *teaching episode* ended with a correct child production, we counted how many times in total the child had correctly produced the target during intervention sessions to date as a proxy for the number of *cumulative teaching episodes*. Because the targets were all individually assigned and intervention ceased according to pre-defined stopping criteria, the different participants received intervention on different numbers of targets, for differing amounts of time. Figure 6 shows all participants progressed on at least one target. However, the rate of progress varied greatly both within and between participants, with some participants showing much faster progress on some targets than others.

In order to establish whether using *cumulative teaching episodes* as the predictor gave similar results to the analysis in Research Question 2 using *cumulative days in intervention*, we ran a mixed effects model with two fixed effects: participant (sum-coded) and *cumulative teaching episodes* and their interaction. In this analysis, we were not interested in the effect of the individual targets (see Research Question 6), or the dosage group (see Research Questions 4 and 5), so these were included as potential random effects. The best fitting model that converged included the fixed effects of participant and *cumulative teaching episodes* and their interaction and the random intercept of dosage group and by-target random slope with *cumulative teaching episodes*. The results are shown in Table 7.

Table 7 shows very similar results to the intervention phase analysis in Research Question 2 (Table 6). The odds of a correct answer increased significantly ($p<.001$) with *cumulative teaching episodes* (3.9% per *teaching episode*). ID2 still showed significantly greater progress than the grand mean and ID6 significantly less.

Given that ID6 had a shallower slope than the grand mean across all participants, we then repeated the analysis with just this participant. This showed no significant increase with cumulative teaching episodes, $OR = 1.01$ (CI=1.00-1.02), $p= .17$.

Table 7 - Model showing odds ratios with fixed predictors of cumulative teaching episodes and participant (sum-coded), significant results (p<.004) shown in bold.

Predictors	*Odds Ratios*	*95% Confidence Interval*	*p*-value
Cumulative teaching episodes	**1.039**	**1.018 – 1.060**	**<0.001**
Cumulative teaching episodes * ID 1	1.006	1.000 – 1.012	0.059
Cumulative teaching episodes * ID 2	**1.009**	**1.004 – 1.015**	**0.001**
Cumulative teaching episodes * ID 3	0.984	0.968 – 1.001	0.062
Cumulative teaching episodes * ID 4	1.006	1.001 – 1.011	0.018
Cumulative teaching episodes * ID 5	1.002	0.996 – 1.007	0.531
Cumulative teaching episodes * ID 6	**0.987**	**0.980 – 0.995**	**0.001**
Cumulative teaching episodes * ID 7	1.004	1.000 – 1.008	0.076
Cumulative teaching episodes * ID 8	1.002	0.998 – 1.007	0.323
Random Effects			
σ^2	3.29		
$\tau_{00\ Dosage\ Group}$	0.22		
$\tau_{00\ Target}$	1.96		
$\tau_{11\ Target.Cumulative\ teaching\ episodes}$	0.00		
$\rho_{01\ Target}$	0.32		
ICC	0.83		
$N_{\ Dosage\ Group}$	3		
$N_{\ Target}$	26		
Observations	572		
Marginal R^2 / Conditional R^2	0.198 / 0.867		

Thus, the results from Research Question 3 match those of Research Question 2 leading us to conclude that on average the participants benefited from the intervention, with the exception of ID6. ID2 seems to have shown particularly strong progress.

Research Question 4 – Does *number of teaching episodes per target per session* affect rates of progress with intervention based on *cumulative intervention sessions* and/or *cumulative teaching episodes*?

We carried out two separate analyses of dosage (planned *number of teaching episodes per target per session:* 10, 20 or 30) and its interaction with a) *cumulative intervention sessions* and b) *cumulative teaching episodes*. Both these predictors include data from the baseline and maintenance phases. Here we are not interested in the performance of individual participants, or targets (see Research Question 3), so these were included as random effects. The best fit for the random effects included by-participant and by-target random intercepts and slopes for the analyses of *cumulative intervention sessions* or *teaching episodes*. Both analyses used sum coding for the dosage group and the results are shown in Table 8.

Table 8 shows a highly significant main effect of both *cumulative intervention sessions* and *teaching episodes* ($p<0.001$). The analysis with *cumulative intervention sessions* shows a highly significant interaction where the slope is significantly shallower for 10 *teaching episodes per session* and significantly steeper for 30. However, the analysis with *cumulative teaching episodes* shows no significant interactions. Thus, the interaction for *cumulative intervention sessions* appears to be due entirely to the difference in *cumulative teaching episodes*, where the odds of a correct response increases by 3.9% per *teaching episode*, regardless of the number of *teaching episodes per target per session*.

Table 8 – Odds of a correct response as predicted by teaching episodes per session, cumulative intervention sessions or teaching episodes and their interactions, significant results (p<.004) shown in bold.

	Cumulative intervention sessions			**Cumulative teaching episodes**		
Predictors	*Odds Ratios*	*Confidence Interval*	*p-value*	*Odds Ratios*	*Confidence Interval*	*p-value*
Cumulative intervention sessions	**1.914**	**1.413 – 2.591**	**<0.001**			
Cumulative intervention sessions * Teaching episodes per session =10	**0.835**	**0.762 – 0.916**	**<0.001**			
Cumulative intervention sessions * Teaching episodes per session =20	1.017	0.947 – 1.093	0.637			
Cumulative intervention sessions * Teaching episodes per session =30	**1.176**	**1.079 – 1.283**	**<0.001**			
Cumulative teaching episodes				**1.039**	**1.017 – 1.061**	**<0.001**
Cumulative teaching episodes * Teaching episodes per session =10				1.000	0.993 – 1.006	0.976
Cumulative teaching episodes * Teaching episodes per session =20				1.000	0.996 – 1.004	0.958
Cumulative teaching episodes * Teaching episodes per session =30				1.000	0.995 – 1.006	0. 940
Random Effects						
σ^2	3.29			3.29		
τ_{00}	1.97 Target			2.02 Target		
	0.40 Participant			0.43 Participant		
τ_{11}	0.52 Target. Cumulative intervention Sessions			0.00 Target. Cumulative teaching episodes		
	0.01 Participant. Cumulative intervention Sessions			0.00 Participant. Cumulative teaching episodes		
ρ_{01}	0.20 Target			0.28 Target		
	0.66 Participant			0.68 Participant		
ICC	0.79			0.84		
N	26 Target			26 Target		

	8 Participant	8 Participant
Observations	572	572
Marginal R^2 / Conditional R^2	0.190 / 0.826	0.180 / 0.868

We don't report the fixed effects of *teaching episodes per target per session* as we are only interested in the interaction of this with intervention.

Research Question 5 – Does *number of teaching episodes per target per session* predict maintenance of any progress?

In order to establish whether the number of *teaching episodes per target per session* affected the maintenance of progress, we analyzed just the data points in the maintenance phase using days in maintenance and its interaction with *teaching episodes per target per session* (sum-coded) as predictors. This includes all maintenance datapoints regardless of when they were collected. The best model included random intercepts for participant and target and the by-target and by-participant slope with days in maintenance and is shown in Table 9. This showed no significant change with days in maintenance and no interactions with *teaching episodes per target per session*, thus showing progress was maintained with little difference between the dosages.

Research Question 6 – Do any specific targets show differing patterns or trends?

Figure 7 shows the data for each of the 27 unique targets with *cumulative teaching episodes* on the x-axis. This shows that most participants made progress with most targets, but the rates of progress differed greatly between targets and participants, with some targets showing very rapid progress and others much slower progress. It was not possible to analyze these data statistically as there were many targets and several were only a target of intervention for a single participant.

Table 9 – Odds of a correct response as predicted by teaching episodes per session, days in maintenance and their interactions

Predictors	*Odds Ratios*	*Confidence Interval*	*p-value*
Days in maintenance	0.993	0.983– 1.003	0.145
Days in maintenance * teaching episodes per session [10]	1.004	0.991 – 1.017	0.525
Days in maintenance * teaching episodes per session [20]	0.997	0.987 – 1.007	0.494
Days in maintenance * teaching episodes per session [30]	0.999	0.987 – 1.012	0.918

Random Effects	
σ^2	3.29
$\tau_{00\ Participant}$	0.64
$\tau_{00\ Target}$	3.49
$\tau_{11\ Target.Days\ in\ Maintenance}$	0.00
$\tau_{11\ Participant.Days\ in\ Maintenance}$	0.00
$\rho_{01\ Target}$	0.15
$\rho_{01\ Participant}$	-0.55
ICC	0.58
N_{Target}	21
$N_{Participant}$	8
Observations	131
Marginal R^2 / Conditional R^2	0.065 / 0.610

Figure 7 - Proportion correct by participant and target (see Appendix A for descriptions of each target and examples of each target).

We considered which targets were discontinued due to limited progress. Table 10 shows that across the eight participants, 26 targets were achieved, 11 were still in progress at the end of the study, and 10 were discontinued due to limited progress. Consideration of the individual discontinued targets showed that one target accounted for five uses of the discontinue rule, in IDs 2, 4, 5, 7 and 8 (CJ8/9: coordinating noun, verb and adjective phrases using *but not* and *or*). Indeed, no participant reached the 90% correct criterion for this target. The other targets affected only one participant each (ID3: NP1: plural *-s*, ID6: MC9: Subject + Verb + Object + PP and CJ4: coordinated verb and adjective phrases with *and*, ID7: NP9: possessive -*'s* + Noun, ID8: AD4: *if* and *unless*). For ID8 on *if/unless*, we stopped intervention because he was able to produce the target forms reliably in intervention sessions, but in the probe tests, he persisted in asking questions rather than producing a scorable response, despite repeated prompting[7]. Thus, the test did not seem to reflect his ability.

Table 10 - Number of targets achieved, discontinued or in progress at the end of the study by participant

	Participant									
	ID1	ID2	ID3	ID4	ID5	ID6	ID7	ID8	**Mean**	**Total**
Targets achieved (90% correct)	5	5	1	2	3	2	3	5	**3.25**	**26**
Targets still in progress at end of study	2	1	2	0	2	1	2	1	**1.5**	**11**
Targets discontinued	0	1	1	2	0	2	2	2	**1.25**	**10**
TOTAL number of targets	**7**	**7**	**4**	**4**	**5**	**5**	**7**	**8**	**6**	**47**

We compared the baseline scores of discontinued targets with those of achieved targets. We compared both their last baseline score (immediately before intervention) and their mean baseline score, with target as a random effect. These analyses showed a trend for achieved targets to have higher scores at baseline, but these differences were not significant for either the last, t=1.57(29.9), p=.13, d=0.48 or mean baseline scores, t=1.36(31.6), p=.18, d=0.34. We also analyzed, for the achieved targets, whether the *total teaching episodes*

required to reach criterion correlated with baseline scores. We found no significant correlations with either the last, r = -.34, p=.09 or mean baseline score, r = -.33, p=.10.

Research Question 7 – How many *intervention sessions* or *teaching episodes* are required to reach criterion on a target and how much does this vary?

Here we considered just the 26 individual targets across the eight participants which were achieved and the number of *cumulative teaching episodes* and *intervention sessions* after which criterion was reached. Given that Research Question 4 showed *cumulative teaching episodes* was a much better predictor than *cumulative intervention sessions* (which depends on the number of *teaching episodes per target per session*), we report only *cumulative teaching episodes* in the first column of Table 11. The other columns show for each dosage the *total intervention time per target* (calculated by dividing the *cumulative teaching episodes* to criterion by the achieved *teaching episode rate* for each dosage from Table 4) and the *total number of intervention sessions per target* required to reach criterion (calculated by dividing *cumulative teaching episodes* to criterion by the achieved number of *teaching episodes per session* for each dosage from Table 4),

Table 11 shows wide variation in the required *cumulative teaching episodes* to reach the 90% criterion, ranging from 10 to 193 *teaching episodes*. This variation does not seem to be due to particular participants; as the five targets requiring the highest *cumulative teaching episodes* (90, 93, 113, 141, 193) to reach criterion were spread across five different participants. This also does not appear to be strongly related to baseline scores as the correlations between *cumulative teaching episodes* to reach criterion and baseline scores, were small and non-significant (see Research Question 6).

Table 11: Cumulative teaching episodes, total intervention time and number of sessions required to reach 90% correct criterion (only targets which reached criterion are included).

	***Cumulative teaching episodes* to 90% criterion**	**Total intervention time required (mins)**			**Total number of sessions required**		
		at 10 doses	**at 20 doses**	**at 30 doses**	**at 10 doses**	**at 20 doses**	**at 30 doses**
Mean	58	41	34	32	6.0	3.0	2.0
Median	41	29	24	23	4.3	2.1	1.4
Min	10	7	6	6	1.0	0.5	0.3
Max	193	138	114	107	20.1	10.0	6.7

The mean and median *cumulative teaching episodes* to criterion are particularly informative and indicate that the average target reached the 90% criterion after 40-60 *teaching episodes.* At the three slightly differing *teaching episode rates* for the three dosages reported in Table 4, a mean of 58 *teaching episodes* would take 32, 34 or 41 minutes. At 20 or 30 *teaching episodes per target per session*, this would require two or three intervention sessions (each of 16.1 or 11.5 minutes' length respectively, see Table 4). For the maximum *cumulative teaching episodes* (193), at 20 or 30 *teaching episodes per target per session*, 107 or 114 minutes would be required for one target, which could be completed in 7 or 10 (11-16 minute) sessions. Of course another target could also be completed during longer sessions and other targets may need to be added subsequently.

Discussion

We delivered highly individualized 1:1 grammatical intervention to eight children with DLD where specific targets were identified for each participant. Short probe tests identified when a target was either achieved or should be discontinued, with the aim of maximizing the efficiency of the intervention for each participant and target. In total, between them, the eight participants received intervention on 47 morphosyntactic targets, 27 of which were unique. Each participant received intervention on two targets per session at one of three different dosages (10, 20 or 30 *teaching episodes per target per session*).

General effectiveness

Across all participants and targets, scores on probe tests were significantly higher after intervention than during the baseline phase. There was no significant change during the baseline or maintenance phases, but during the intervention phase there was highly significant progress. This indicates that overall the intervention was effective and progress was maintained.

Differences between participants

Two of the eight participants showed a different response to intervention with one (ID2) showing significantly faster progress with intervention and one (ID6) showing significantly less progress; indeed, for this participant, the intervention does not appear to have been effective. It is of interest that the participant who made faster progress had been in the school previously and had received similar intervention in the two previous years (focused on different targets), thus was already familiar with the SHAPE CODING™ system. ID6 had the lowest attention score (see Table 3).

Effect of Dosage

All three dosages led to significant progress which was maintained after intervention ceased. When considering progress relative to *cumulative intervention sessions*, progress was faster with 30 *teaching episodes per session* and slower with 10 *teaching episodes per session*. However, when *cumulative teaching episodes* was used as the predictor, all three dosages showed very similar rates of progress, with the odds of a correct response increasing by 3.9% for each *teaching episode*. Thus, it seems that the *cumulative teaching episodes* were key to intervention progress and the distribution of those *teaching episodes* mattered less.

This finding is similar to the findings from Plante, Mettler, Tucci, and Vance (2019) where children receiving 24 *teaching episodes per session* achieved similar results, regardless of whether the intervention took 15 or 30 minutes to deliver (i.e., whether the recasts were

densely or sparsely distributed). There are indications of a similar effect in Calder, Claessen, Leitão, et al. (2021), where the *number of teaching episodes per session* was held constant, but the *session frequency*, *total number of intervention sessions* (and hence *total number of teaching episodes*) was doubled for some of the children. Those who received double the *cumulative* and *total number of teaching episodes* made greater progress with intervention, indicating again that perhaps it is the *total/cumulative teaching episodes* which is important. Similar findings were not however found in studies by Bellon-Harn et al. (2014) or Balthazar & Scott (2018), which also varied the same aspects, but focused on syntactic rather than morphological targets. However, Bellon-Harn et al. (2014) also varied the *total intervention period* and had a very small number of participants in their groups study design and hence low power. In addition, both the Bellon-Harn et al. (2014) and Balthazar & Scott (2018) studies involved greater variability in the deliverer (where multiple clinicians delivered the intervention, unlike Calder, Claessen, Leitão, et al. (2021) where a single clinician delivered all intervention sessions).

Our finding that the *cumulative number of teaching episodes* seems to be key rather than how those are distributed, gives clinicians much flexibility with delivery, particularly in the school context. For example, we could expect to achieve the same outcomes after two (16 minute) sessions with 30 *teaching episodes*, or six (7 minute) sessions with 10 *teaching episodes*, or three (11.5 minute) sessions with 20 *teaching episodes*. This means that there could be flexibility according to the child's level of attention and engagement, availability and school timetabling considerations. To maximize efficiency with children with higher levels of attention and engagement, clinicians could also target more than one structure per session in longer (e.g., 25-minute) sessions, as we did in this study.

In our study, all children received weekly intervention, so the *session frequency* was constant, meaning that the *total intervention period per target* in weeks varied according to the

number of *teaching episodes per session* which affected the *number of intervention sessions* required to reach criterion. An alternative protocol could keep the *total intervention period, total intervention time* and *total teaching episodes* constant, while varying the *session frequency* and *session duration*. Thus, decreasing the *session duration* would increase the *number of intervention sessions* required, but these could be provided over the same *total intervention period* if the *session frequency* is increased. This "little and often" model might be better for a child with a shorter attention span in order to minimize *time not focused on intervention targets* and thus maximize outcomes.

Another alternative might be to keep *session duration* constant, but increase the *session frequency*. In this way, the *total intervention period* could be shortened, without reducing the *total teaching episodes* or *total intervention time per target.* Thus, perhaps the same outcomes could be achieved in six sessions twice a week over a three-week period as in six sessions once a week over a six-week period. This "short and sharp" model might be useful for a clinician who can only spend a limited period of time in a particular school, or where the child is only available for a limited period (e.g., during a school holiday). We did not manipulate *session frequency* in this study, so this study provides no information on the effectiveness, efficiency or maintenance of progress of this possible model, but the study by Calder, Claessen, Leitão, et al. (2021) found faster progress during intervention with more frequent sessions. However, the maintenance of progress in a "short and sharp" model versus a model with wider spaced sessions requires further investigation.

Effect of intervention targets

In general, the participants appeared to show progress across a range of targets. However, the discontinue criterion due to limited progress was reached for seven of the eight participants for at least one target. One particular target (CJ8 & 9) was discontinued in all five participants for whom it was an area of focus, indicating that there was an issue with this target.

This target focused on coordinating noun, verb and adjective phrases using *but not* and *or* and showed initial progress with intervention, but then performance plateaued. Consideration of the actual items on this probe test showed that most participants made progress with the items involving coordinated verb and adjective phrases and noun phrases embedded in a verb or prepositional phrase (e.g., *the lady could sing or dance; the car is blue but not fast, the boy wants a cat but not a dog*), but made very limited progress on items involving coordinated subject noun phrases, e.g., *Mum or Dad will wash the dog, the man but not the lady is reading* and *the lady, but not the boy, is hugging the girl*, where they either produced two coordinated clauses ("Mum will wash the dog, or Dad will wash the dog"), or added the phrase preceded by *but not* to the end of the clause ("the man is reading, but not the lady", "the lady is hugging the girl, but not the boy" – despite this latter example being ambiguous) . Therefore, in a revised version of the target identification spreadsheet available from www.shapecoding.com, we have now split this structure off into a separate target which has its own separate probe test for use in future studies and clinical practice[8]. In addition, as coordination of the subject with *or* and *but not* is less common in everyday conversations, we now recommend this structure should not be targeted if participants can understand it and are able to express the meaning in another way (e.g., *Mum will wash the dog, or Dad will* or *The man is reading, but the lady isn't*).

The other targets which were discontinued each only affected one participant, and therefore the issue seems to have been more specific to the individual participant and subsets of the target. For example, ID3 and ID7 made good progress with plural *-s* and possessive '*-s* respectively where the added *-s* is realized as [s] or [z], but they made no progress on those items ending in a sibilant, requiring the [ɪz] allomorph, where they continued to make errors, despite specific focus on this during intervention sessions. We view the use of the discontinue criterion as important for not spending limited intervention time on a target where little

(further) progress is being made, especially when there are other structures or areas which could be targeted instead.

It was not possible to distinguish pre-intervention which targets were likely to be discontinued versus achieved, as these two groups of targets did not differ in baseline scores. Thus, it seems likely it was the nature of the targets and/or probe tests (as described above) that affected progress with intervention rather than baseline scores.

Intervention required to achieve targets

When we considered just the targets which were achieved, we found wide variation in the *total teaching episodes* and therefore the *total intervention sessions* required, again indicating the need for an individualized approach. However, the average achieved target reached criterion following 40-60 *teaching episodes* which, at 20 or 30 *teaching episodes per session,* could be completed in two or three (11-16 min) intervention sessions (and two or more targets could be combined in longer sessions, as in this study). Even for those targets requiring the maximum *cumulative teaching episodes*, 7 or 10 (11-16 min) intervention sessions would be sufficient. This is far less than the *total number of teaching episodes* in previous research projects. For example, Cleave et al. (2015) found that effective recasting intervention requires 10-20 sessions of 30-60 minutes with 300-1000 exposures per target. The study by Calder, Claessen, Leitão, et al. (2021) using very similar intervention to this study (focused only on the past tense, but with younger children), involved either 500 or 1000 *teaching episodes*, depending on whether the children received intervention once or twice a week.

There are multiple potential reasons for the lower *total teaching episodes* (and hence *total intervention time* and *number of intervention sessions*) required to achieve targets in our study: the *teaching episode rate during sessions*, and the nature of the targets, the participants and the *intervention techniques*. Next, we consider each of these in turn.

Teaching episode rate during sessions

In typical clinical practice, when focusing on morphosyntactic targets, the *number of teaching episodes* and hence *cumulative teaching episodes* after a given number of sessions is likely to be lower than 40 *teaching episodes* in a 30-minute session. Indeed, clinicians working with children in elementary schools in the US responding to Finestack and Satterlund's (2018) survey said that the children received between 11 and 20 *teaching episodes per session*, but in the opinion of over 90% of respondents, the ideal number of *teaching episodes per session* would be around 40 (which matches the total number we aimed for over the two targets in each session). Given the highly predictive nature of *cumulative teaching episodes* on progress on probe tests, this is clearly an area which would benefit from greater focus in clinical practice. If the number of teaching episodes could be increased within the same *total intervention time* (i.e., the *teaching episode rate* within sessions increased), this could greatly improve the efficiency of intervention. Indeed, Plante et al. (2019) found that if they doubled the *teaching episode rate* (using recasting intervention, where the *teaching episode*, which they call a *dose*, was defined as an adult attentional cue plus a unique recast, rather than a child production), they could achieve the same results in half the time. However, in their study, in total the children received 6.25 or 12.5 hours of intervention with 600 *teaching episodes*. This is far higher than the 10-193 *teaching episodes* (or 6-138 minutes) required for the achieved targets in our study. In contrast to our study, Plante et al. (2019) did not report the number of child correct productions (our measure of a completed *teaching episode*), which we viewed as an important *active ingredient* of our *teaching episodes.*

Another way to increase *teaching episodes per session* could be to decrease the *time not focused on targets per session.* In our study, on average, 23 minutes of each planned 30-minute session (77%) was spent focusing on targets. This was a greater percentage than the 49% found in Schmitt et al. (2017). The remaining time was spent collecting the children from

class (which was not included in Schmitt et al.'s calculation), building rapport and rewarding engagement, introducing the session, organizing intervention materials, transitioning between targets and managing interruptions and distractions. Many of these are unavoidable practicalities of providing intervention and some may themselves be active ingredients (see Figure 1). However, by being aware of these and the balance between the amount of time spent directly on these versus the amount spent focusing on targets, clinicians may be able to make small changes which would increase the amount of intervention session time focused directly on targets, thereby increasing efficiency.

In our study, we carried out weekly probe sessions; these could have increased the *teaching episodes* beyond those which were included in the intervention and hence counted in our analysis, as each probe test included ten test items with no feedback, one model item and one practice item. Depending on whether or not just producing a structure without any of the other *intervention techniques* functions as an *active ingredient* of a *teaching episode*, this could potentially have increased the *cumulative teaching episodes* by ten per week (as any child scoring 90% or 100% correct would cease intervention, so these correct productions would not be included). However, even with probe test productions included, the *total number of teaching episodes* for achieved targets was still much lower than in previous studies.

Targets

The targets in this study were individually selected for each participant and were chosen on the basis that the child scored below 90% correct on probe tests in all baseline sessions. Thus, for some children, the baseline scores could be relatively high. In intervention projects where 10-20 sessions will be provided on a single target, children with relatively high scores on that target are unlikely to be included. Indeed, in Plante et al.'s (2019) study, targets were chosen to be below 30% correct pre-intervention. Because the children in our study had multiple targets and could move on to the next target as soon as the target was achieved, we

were able to include targets where they may have had relatively high scores pre-intervention, but still below 90%.

The higher pre-intervention scores for some targets means that these targets were likely to be achieved with a lower number of *teaching episodes* and hence in fewer *intervention sessions*. However, we found no significant correlation among the achieved targets between baseline scores and *cumulative teaching episodes* to criterion.

Another potential reason for the low number of *teaching episodes* and *intervention sessions* required in this project is that our analysis of dosage to criterion only considers those targets which were achieved, which may have been inherently easier. However, we found no significant differences in baseline scores between achieved and discontinued targets.

Most previous studies considering dosage in the area of sentence production and grammar have focused primarily on morphology. While such morphological targets were included in our study, we also included a wider range of targets, including syntactic targets that aimed to increase sentence complexity in addition to accuracy. Thus, it is possible (but unlikely) that syntactic targets require fewer *teaching episodes* and sessions than morphological targets.

Participants

The children in this study were older (8-11 years) than in most previous studies considering dosage (see Table 1). However, in Frizelle et al.'s (2021) review of quantitative dose, they posited that perhaps younger children require fewer sessions than older children, suggesting that "younger children might react faster and more easily to intervention than older children" (p.748). Thus, it seems unlikely that the older age of the children in the current study accounted for the faster progress than in other similar studies. However, it may be that even if younger children respond faster than older children to predominantly implicit methods of instruction, the reverse could be the case for a predominantly explicit intervention. Thus, future

studies should consider the interaction between the ages of the children, the method of instruction, *intervention techniques*, how they are combined, what constitutes a *teaching episode* and the *total number of teaching episodes*.

Intervention techniques / active ingredients

The intervention in this project uses a range of *intervention techniques,* all or some of which could be *active ingredients.* We provided *SHAPE CODING templates* to simplify explanations of grammatical rules, capitalizing on the relative strengths in declarative memory and visuo-spatial skills of children with DLD (Lum et al., 2012) to support correct production. We required participants to attempt production of targets and provided support and a feedback hierarchy until they were successful. This contrasts with many other language intervention studies which do not require correct production from participants. However, Smith-Lock et al. (2015) found children who were provided with a feedback hierarchy to support correct production following recasting made greater progress than children who only heard recasts, suggesting that child correct production of a target form and the techniques in the feedback hierarchy could be important *active ingredients* that increase the effectiveness of a *teaching episode*. Indeed Frizelle et al. (2021) concluded that whether the child is given the opportunity to produce the target plays a role in improving outcomes.

In our intervention, we aimed to give the minimum possible support to facilitate a correct child production. Therefore, as a participant showed increased proficiency with a target, fewer supports (such as the *SHAPE CODING template*) were provided. We thus aimed for errorless learning with repeated effortful, but successful, retrievals, as this has been found to be successful in other areas of learning (e.g., Rowland, 2014), including word learning in children and young adults with DLD (e.g., Gordon et al., 2020; Leonard & Deevy, 2020). This appeared to be successful as 82% of productions were correct on the first attempt. When the participants did make an error, a feedback hierarchy was provided which again aimed to

give the participants the minimum support necessary to correct their own error. In the majority of cases, the first two steps of the feedback hierarchy (*question by repeating the error,* or *explain the error*) led to a successful child production of the target and the remaining steps in the feedback hierarchy were not required.

Limitations and future directions

While our study provides a detailed investigation of the effects of intervention and within session dosage on progress, we acknowledge that our small sample size of eight participants limits the extent to which our conclusions can be generalized to the broader population of children with DLD, particularly as our participants all had severe levels of DLD, attended the same educational institution, and received intervention provided predominantly by a single clinician. Thus, further investigation in larger studies is required in order to establish whether the results hold across all targets and a wider range of participants and whether progress generalizes to standardized tests and general measures of sentence length, complexity and/or accuracy.

We only evaluated three different dosages (10, 20 or 30 *teaching episodes per target per session*) and only at one *session frequency* (once per week). Therefore, our finding that it is the *total number of teaching episodes* that is key, and not the *total number of intervention sessions* may not hold at other dosages and session frequencies. Thus, future studies could extend and replicate this work with larger, more diverse samples in different settings and with different session frequencies, with intervention delivered by a range of clinicians.

Clinical implications

We found faster progress with intervention than in previous studies. This was perhaps due to the fidelity to a higher *teaching episode rate*, the highly individualized nature of the

intervention and the errorless learning approach with repeated effortful but successful retrievals, supported when necessary by *SHAPE CODING templates* and a feedback hierarchy.

Identifying multiple individualized targets which could all receive a similar intervention method has several advantages. It gives clinicians the flexibility to focus on more than one target at a time and also to discontinue a target and start a new one if a participant is not making progress. In addition, participants can use the same intervention method over an extended period of time, and this increased familiarity may lead to faster progress, as seen with ID2.

Our finding that *cumulative teaching episodes* was the crucial dosage factor indicates that clinicians could focus on how to maximize efficiency by aiming for a high *teaching episode rate*; this can be operationalized by aiming for a particular number of *teaching episodes per target per session*. It will be important to try to minimize *time spent not focused on intervention targets*, including moving between locations, setting up intervention materials and managing the child's attention levels and motivation. For some children and settings this may mean that more shorter sessions are more efficient, and for others, fewer longer sessions. We recognize that many clinicians have little control over intervention delivery schedules. However, most clinicians have control over what happens within sessions.

Our study indicates that the primary factor clinicians should focus on is increasing the *teaching episode rate* within sessions and decreasing the proportion of *time not focused on intervention targets*. This could potentially lead to children achieving the same outcomes in less intervention time, or better outcomes in the same amount of time. This could reduce the amount of time that children spend in intervention (and thus away from educational, social or relaxation time) and increase the time clinicians have available, either for planning or seeing other children. Thus, clinicians, funders, children and their families would all benefit from a greater focus on maximizing efficiency via high *teaching episode rates*.

Conclusions

Our study design enabled us to look in detail at the progress of eight participants on a range of morphological and syntactic targets. In general, the intervention appears to have been highly effective and efficient (with 1-5 targets achieved per participant and another 1-2 still in progress after 8-10 hours of intervention), but progress varied between targets and participants. The results for individual targets have led us to modify the target identification process for the future with the aim of increasing effectiveness.

One participant (with low attention) does not appear to have benefited, but the other seven showed highly significant progress with intervention. One participant who had received similar intervention previously showed faster progress, indicating that greater gains could perhaps be achieved in a shorter time if the same intervention approach were continued (but with new targets). We will continue to refine our target identification process and stopping criteria in response to future trials to further improve effectiveness and efficiency and we hope to test the methods further in larger studies in order that generalization to more general language measures can be evaluated.

Despite the need for further larger-scale trials, we believe the results of this study suggest that this method is effective for children with DLD who have sufficient attention levels to attend to the SHAPE CODING procedures and feedback hierarchy but who require support with constructing a range of sentence structures accurately. The ability to use the morphosyntactic forms targeted here will allow children with DLD to express their thoughts and ideas more clearly, will impact their self–efficacy, and is central to both their academic and social functioning.

Acknowledgments

We would like to thank the participants, their families and teachers for their support and flexibility throughout this study.

Data Availability Statement

All data for this study are available from

https://osf.io/32ubp/?view_only=f1ee0fb2a07a4b3198d1b051fa0056b5

References

Baker, E. (2012). Optimal intervention intensity. *International Journal of Speech-Language Pathology*, *14*(5), 401-409. https://doi.org/10.3109/17549507.2012.700323

Balthazar, C. H., & Scott, C. M. (2018). Targeting Complex Sentences in Older School Children With Specific Language Impairment: Results From an Early-Phase Treatment Study. *Journal of Speech, Language, and Hearing Research*, *61*(3), 713-728. https://doi.org/doi:10.1044/2017_JSLHR-L-17-0105

Bellon-Harn, M. L., Byers, B. A., & Lappi, J. (2014). Treatment Intensity: Effects of Interactive Book Reading on Narrative Abilities in Preschool Children With SLI. *Communication Disorders Quarterly*, *35*(4), 226-236. https://doi.org/10.1177/1525740114524051

Brown, V. A. (2021). An Introduction to Linear Mixed-Effects Modeling in R. *Advances in Methods and Practices in Psychological Science*, *4*(1), 2515245920960351. https://doi.org/10.1177/2515245920960351

Calder, S. D., Claessen, M., Ebbels, S., & Leitão, S. (2021). The Efficacy of an Explicit Intervention Approach to Improve Past Tense Marking for Early School-Age Children With Developmental Language Disorder. *Journal of Speech, Language, and Hearing Research*, *64*(1), 91-104. https://doi.org/doi:10.1044/2020_JSLHR-20-00132

Calder, S. D., Claessen, M., Leitão, S., & Ebbels, S. (2021). Evaluating two different dose frequencies and cumulative intervention intensities to improve past tense production for early school-aged children with developmental language disorder. *International Journal of Language & Communication Disorders*, *56*(6), 1278-1295. https://doi.org/https://doi.org/10.1111/1460-6984.12667

Cleave, P. L., Becker, S. D., Curran, M. K., Van Horne, A. J. O., & Fey, M. E. (2015). The efficacy of recasts in language intervention: a systematic review and meta-analysis. *American journal of speech-language pathology / American Speech-Language-Hearing Association*, *24*(2), 237-255. MEDLINE:25654306 (In File)

Conti-Ramsden, G., Durkin, K., Toseeb, U., Botting, N., & Pickles, A. (2018). Education and employment outcomes of young adults with a history of developmental language disorder. *International Journal of Language & Communication Disorders*, *53*(2), 237-255. https://doi.org/10.1111/1460-6984.12338

Curran, M., & Owen van Horne, A. J. (2019). Use of Recast Intervention to Teach Causal Adverbials to Young Children With Developmental Language Disorder Within a Science Curriculum: A Single Case Design Study. *American Journal of Speech-Language Pathology*, *28*(2), 430-447. https://doi.org/doi:10.1044/2018_AJSLP-17-0164

Ebbels, S. H. (2014). Effectiveness of intervention for grammar in school-aged children with primary language impairments: A review of the evidence. *Child Language Teaching and Therapy*, *30*(1), 7-40. (In File)

Ebbels, S. H., Maric, N., Murphy, A., & Turner, G. (2014). Improving comprehension in adolescents with severe receptive language impairments: a randomised control trial of intervention for coordinating conjunctions. *International Journal of Language & Communication Disorders*, *49*(1), 30-48. (In File)

Ebbels, S. H., van der Lely, H. K. J., & Dockrell, J. E. (2007). Intervention for verb argument structure in children with persistent SLI: a randomized control trial. *Journal of Speech Language and Hearing Research*, *50*, 1330-1349. (In File)

Fey, M. E., Cleave, P., Long, S. H., & Hughes, D. L. (1993). Two Approaches to the Facilitation of Grammar in Children with Language Impairment: An Experimental Evaluation. *Journal of Speech and Hearing Research*, *36*, 141-157. (In File)

Fey, M. E., Cleave, P. L., & Long, S. H. (1997). Two models of grammar facilitation in children with language impairments: phase 2. *Journal of Speech Language and Hearing Research*, *40*, 5-19. (In File)

Fey, M. E., Leonard, L. B., Bredin-Oja, S. L., & Deevy, P. (2017). A Clinical Evaluation of the Competing Sources of Input Hypothesis. *Journal of Speech, Language, and Hearing Research*, *60*(1), 104-120. https://doi.org/10.1044/2016_JSLHR-L-15-0448

Finestack, L. H., & Satterlund, K. E. (2018). Current Practice of Child Grammar Intervention: A Survey of Speech-Language Pathologists. *American Journal of Speech-Language Pathology*, 1-23. https://doi.org/10.1044/2018_AJSLP-17-0168

Frizelle, P., & Fletcher, P. (2014a). Profiling relative clause constructions in children with specific language impairment. *Clinical Linguistics & Phonetics*, *28*(6), 437-449. https://doi.org/10.3109/02699206.2014.882991

Frizelle, P., & Fletcher, P. (2014b). Relative clause constructions in children with specific language impairment. *International Journal of Language & Communication Disorders*, *49*(2), 255-264. https://doi.org/https://doi.org/10.1111/1460-6984.12070

Frizelle, P., Thompson, P. A., McDonald, D., & Bishop, D. V. M. (2018). Growth in syntactic complexity between four years and adulthood: evidence from a narrative task. *Journal of child language*, 1-24. https://doi.org/10.1017/S0305000918000144

Frizelle, P., Tolonen, A.-K., Tulip, J., Murphy, C.-A., Saldana, D., & McKean, C. (2021). The Impact of Intervention Dose Form on Oral Language Outcomes for Children With Developmental Language Disorder. *Journal of Speech, Language, and Hearing Research*, *64*(8), 3253-3288.

Gordon, K. R., McGregor, K. K., & Arbisi-Kelm, T. (2020). Optimising word learning in post-secondary students with Developmental Language Disorder: The roles of retrieval difficulty and retrieval success during training. *International Journal of Speech-Language Pathology*, 1-14. https://doi.org/10.1080/17549507.2020.1812719

Kulkarni, A. A., Chadd, K. E., Lambert, S. B., Earl, G., Longhurst, L. M., McKean, C., . . . Norbury, C. F. (2022). Editorial Perspective: Speaking up for developmental language disorder – the top 10 priorities for research. *Journal of Child Psychology and Psychiatry*, *63*(8), 957-960. https://doi.org/https://doi.org/10.1111/jcpp.13592

Lammertink, I., Boersma, P., Wijnen, F., & Rispens, J. (2017). Statistical Learning in Specific Language Impairment: A Meta-Analysis. *Journal of Speech, Language, and Hearing Research*, *60*(12), 3474-3486. https://doi.org/doi:10.1044/2017_JSLHR-L-16-0439

Leonard, L. B. (2014). *Children with specific language impairment*. MIT press.

Leonard, L. B., & Deevy, P. (2020). Retrieval Practice and Word Learning in Children With Specific Language Impairment and Their Typically Developing Peers. *Journal of Speech, Language, and Hearing Research*, *63*(10), 3252-3262. https://doi.org/doi:10.1044/2020_JSLHR-20-00006

Lum, J., Conti-Ramsden, G., Page, D., & Ullman, M. (2012). Working, declarative and procedural memory in specific language impairment. *Cortex*, *48*(9), 1091-1250. (In File)

Nippold, M. A. (2010). It's Not Too Late to Help Adolescents Succeed in School. *Language, Speech and Hearing Services in Schools*, *41*, 137-138. (In File)

Nippold, M. A., Mansfield, T. C., Billow, J. L., & Tomblin, J. (2009). Syntactic Development in Adolescents With a History of Language Impairments: A Follow-Up Investigation. *American Journal of Speech-Language Pathology*, *18*(3), 241-251. WOS:000268697300005 (Not in File)

Norbury, C. F., Gooch, D., Wray, C., Baird, G., Charman, T., Simonoff, E., . . . Pickles, A. (2016). The impact of nonverbal ability on prevalence and clinical presentation of language disorder: evidence from a population study. *Journal of Child Psychology and Psychiatry*, *57*(11), 1247-1257. https://doi.org/doi:10.1111/jcpp.12573

Owen van Horne, A. J. O., Fey, M., & Curran, M. (2017). Do the Hard Things First: A Randomized Controlled Trial Testing the Effects of Exemplar Selection on Generalization Following Therapy for Grammatical Morphology. *Journal of Speech, Language, and Hearing Research*, *60*(9), 2569-2588. https://doi.org/10.1044/2017_JSLHR-L-17-0001

Plante, E., Mettler, H. M., Tucci, A., & Vance, R. (2019). Maximizing Treatment Efficiency in Developmental Language Disorder: Positive Effects in Half the Time. *American Journal of Speech-Language Pathology*, *28*(3), 1233-1247. https://doi.org/doi:10.1044/2019_AJSLP-18-0285

Plante, E., Ogilvie, T., Vance, R., Aguilar, J. M., Dailey, N. S., Meyers, C., . . . Burton, R. (2014). Variability in the Language Input to Children Enhances Learning in a Treatment Context. *American Journal of Speech-Language Pathology*, *23*(4), 530-545. https://doi.org/10.1044/2014_AJSLP-13-0038

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological bulletin*, *140*, 1432-1463. https://doi.org/10.1037/a0037559

Schmitt, M. B., Justice, L. M., & Logan, J. A. R. (2017). Intensity of language treatment: contribution to children's language outcomes. *International Journal of Language & Communication Disorders*, *52*(2), 155-167. https://doi.org/https://doi.org/10.1111/1460-6984.12254

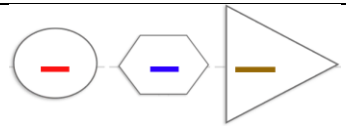
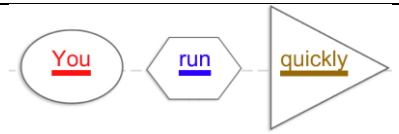
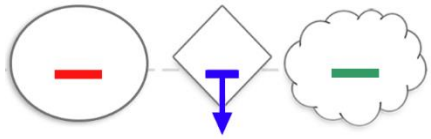
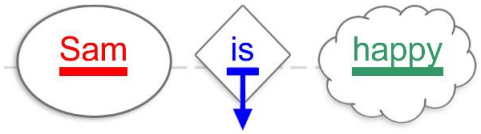
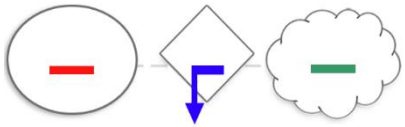
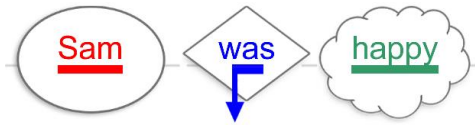
Scott, C. M., & Windsor, J. (2000). General Language Performance Measures in Spoken and Written Narrative and Expository Discourse of School-Age Children With Language Learning Disabilities. *Journal of Speech, Language, and Hearing Research*, *43*(2), 324-339. https://doi.org/10.1044/jslhr.4302.324

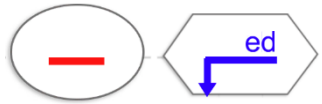
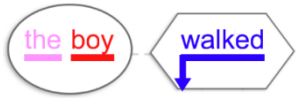
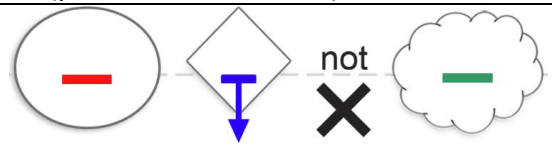
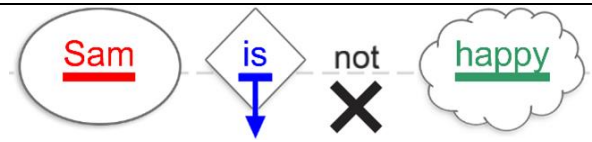
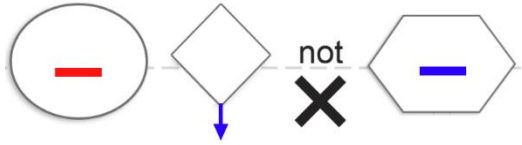
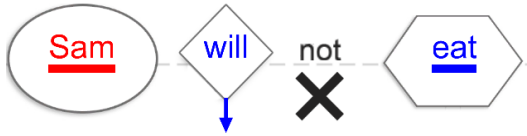
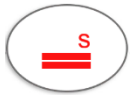
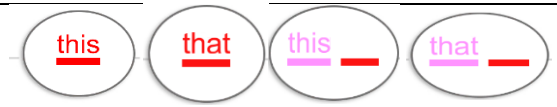
Smith-Lock, K. M., Leitão, S., Prior, P., & Nickels, L. (2015). The Effectiveness of Two Grammar Treatment Procedures for Children With SLI: A Randomized Clinical Trial. *Language, Speech, and Hearing Services in Schools*, *46*(4), 312-324. https://doi.org/10.1044/2015_LSHSS-14-0041

Warren, S. F., Fey, M. E., & Yoder, P. J. (2007). Differential treatment intensity research: a missing link to creating optimally effective communication interventions. *Mental Retardation and Developmental Disabilities Research Reviews*, *13*, 70-77. (Not in File)

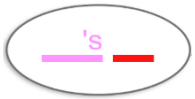
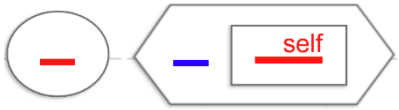
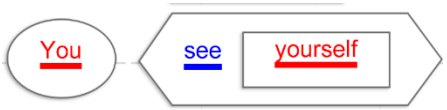
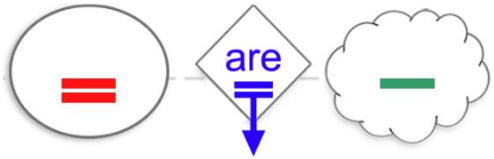
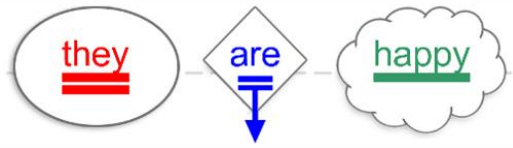
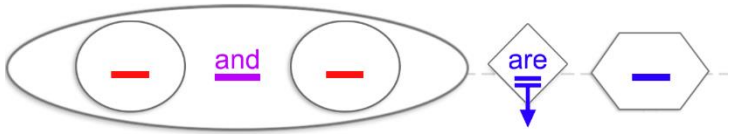
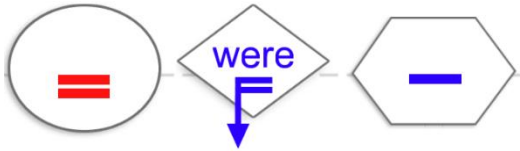
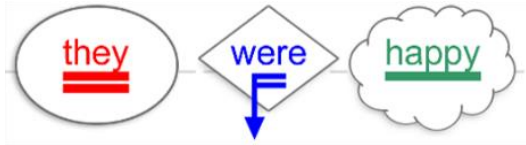
Yew, S. G. K., & O'Kearney, R. (2013). Emotional and behavioural outcomes later in childhood and adolescence for children with specific language impairments: meta-analyses of controlled prospective studies. *Journal of Child Psychology and Psychiatry*, *54*(5), 516-524. https://doi.org/https://doi.org/10.1111/jcpp.12009

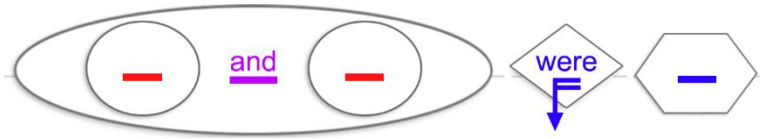
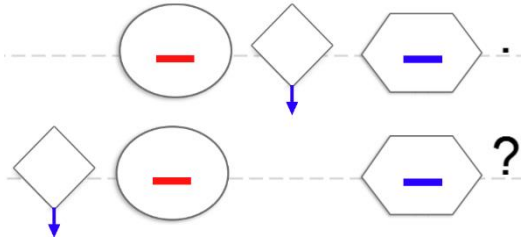
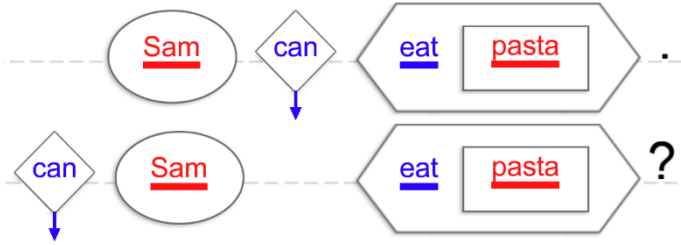
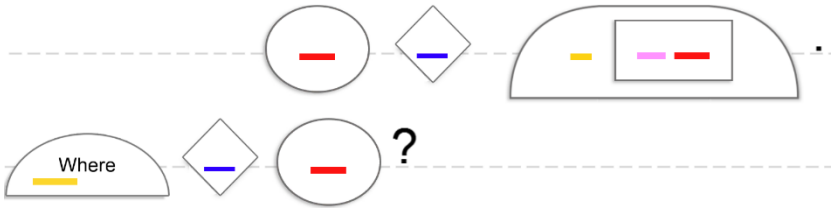
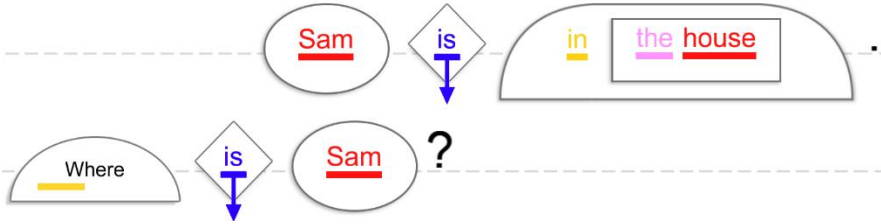
Appendix A – Target codes and descriptions for structures targeted in the study, with the SHAPE CODING™ template and child-friendly rule, and a coded example sentence.

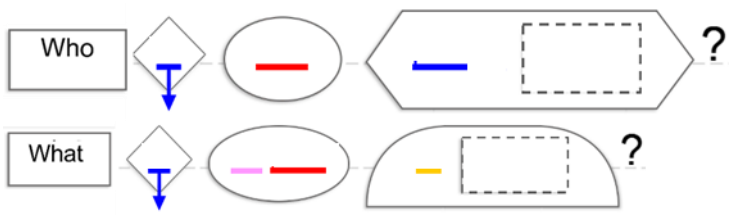
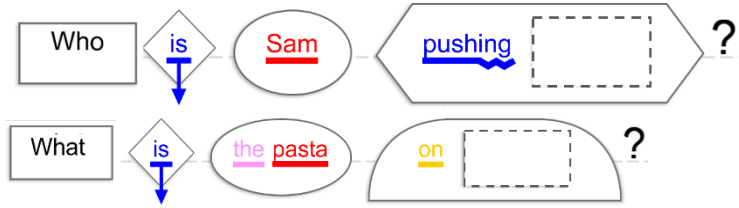
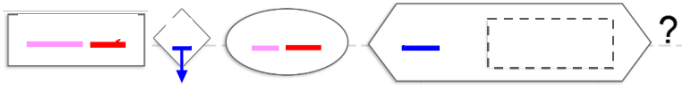
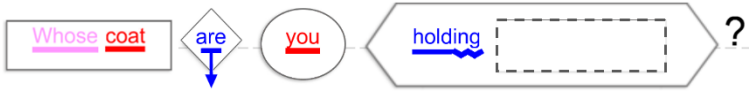
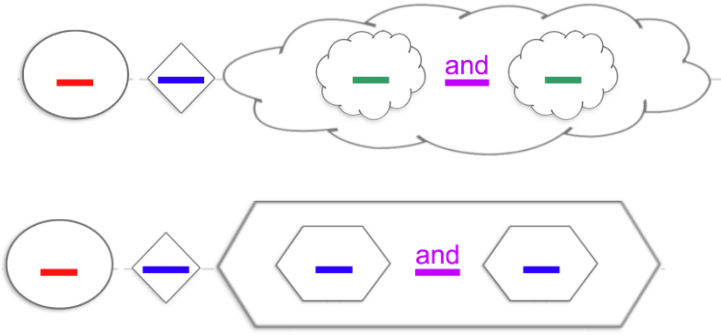
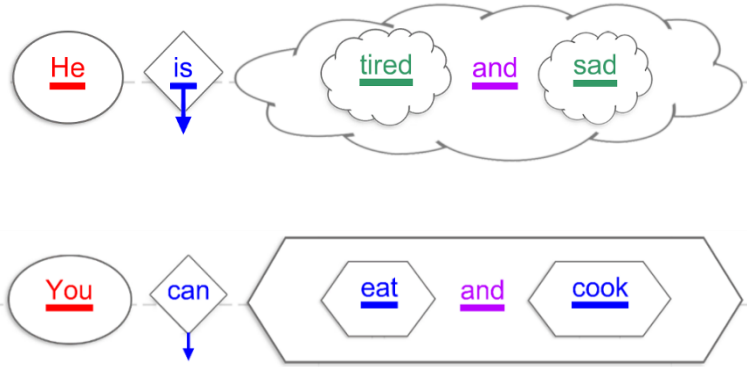
Code: Structure	SHAPE CODING template plus rule	Example
MC9: Subject moves an object to a new place (Subject + Verb + Object + Prepositional Phrase)	 Oval moves rectangle to a new place (semi-circle)	
MC10: Adverbs of manner	 Make brown word from green word by adding *-ly*. Brown word tells you how the oval is doing the blue word (pointy triangle goes with pointy hexagon)	
TA2: present tense copula/aux (*is/are/am*)	 Need a blue word (*is, are,* am) in the diamond between oval and cloud	
TA4: past tense copula/aux (*was/were*)		

	When talking about past time, we need a past (back) arrow on the blue word in the diamond (this changes *is* and *am* to *was*, and *are* to *were*).	
TA5: sentences requiring the past tense	 Adding back arrow for past time onto hexagon blue word adds -*ed* (pronounced /t, d, Id/)	
NG2: auxiliary/copula + *not*	 The *not* cross goes after a diamond	
NG3: modal + *not*	 The *not* cross goes after a diamond	
NP1: Plural -*s*	 More than one needs two red lines. Add -*s* (pronounced /s,z,iz/)	
NP5: Demonstratives *this* vs *that*	 *This* is for nearby, *that* is for further away. Can be red or pink word	

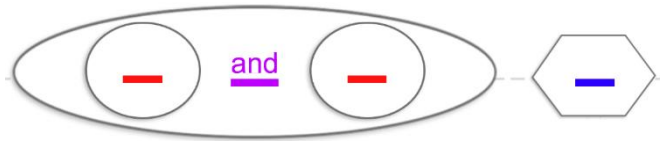
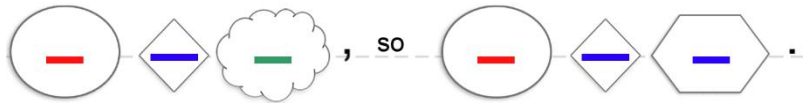
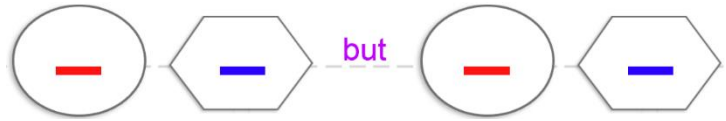
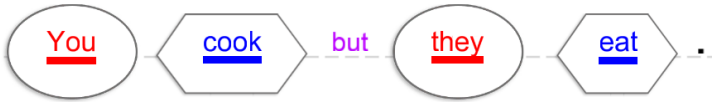
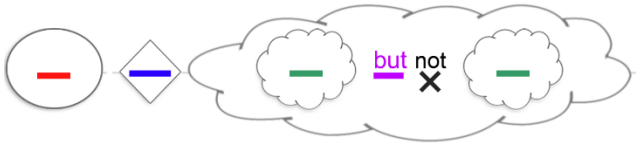
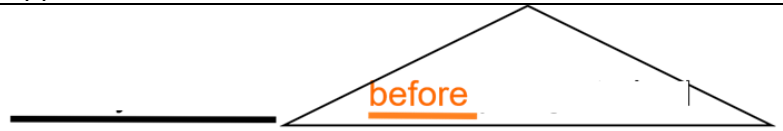
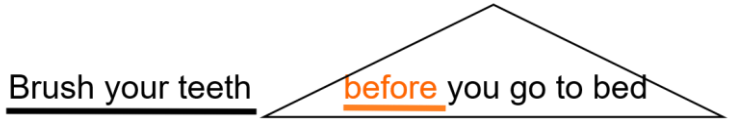
EXPLICIT SHAPE CODING™ INTERVENTION FOR DLD

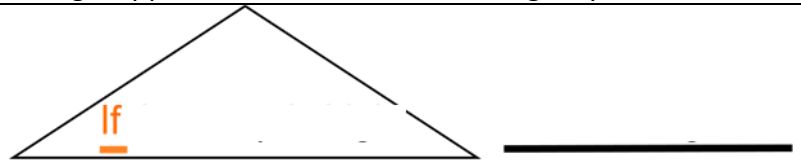
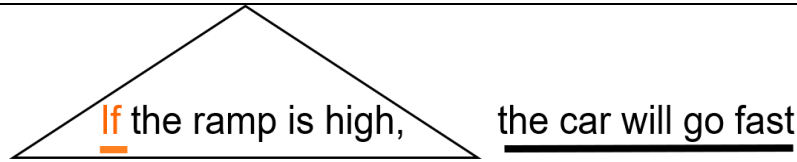
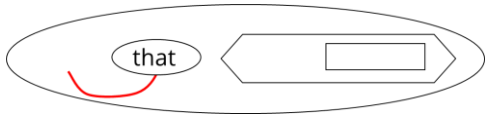
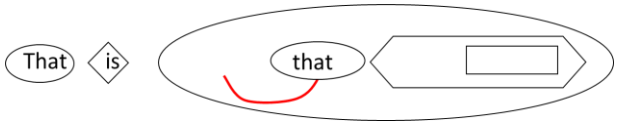
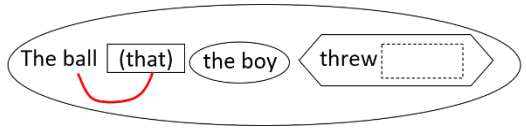
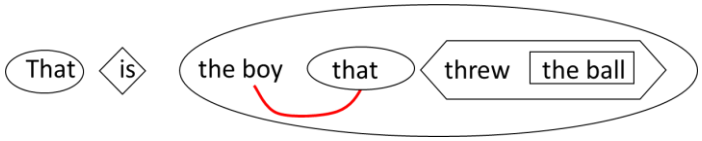
NP7: Possessive *-s* + Noun	 To show something belongs, add –'*s* to turn red word into pink word	
NP10: Reflexive pronoun	 When oval and rectange are the same person use *myself, yourself, himself, herself, ourselves, themselves* in the rectangle	
AG1: *are* with plural Noun Phrase	 Two red lines in oval needs two blue lines in diamond, *are* in present tense	
AG2: *are* with coordinated Noun Phrases	 Two red lines in big oval (one in each small oval) needs two blue lines in diamond, *are* in present tense	
AG3: *were* with plural Noun Phrase		

	Two red lines in oval needs two blue lines in diamond, w*ere* in past tense	
AG4: *were* with coordinated Noun Phrases	 Two red lines in big oval (one in each small oval) needs two blue lines in diamond, *were* in past tense	
Q11: Question formation with movement of modal	 To ask a *yes/no* question, move the diamond to the front	
Q15: *Where, why, how* questions	 Move the *Wh* shape to the front and then move the diamond to second position	

Q16: *Who, what* object questions requiring movement	 Move the *Wh* rectangle to the front and then move the diamond to second position. To understand these questions, put the rectangle back in place.	
Q19: *Whose, which* Noun object questions requiring movement	 Move the *Wh* rectangle to the front and then move the diamond to second position. To understand these questions, put the rectangle back in place.	
CJ4: Coordinated Verb and Adjective Phrases with *and*	 Join two clouds in a big cloud, or two hexagons in a big hexagon.	

EXPLICIT SHAPE CODING™ INTERVENTION FOR DLD

CJ5: Coordinated Noun Phrases with *and*	 Join two ovals together in a big oval	
CJ6: Causal conjunct *so*	 *So* joins two sentences. The first sentence causes the second to happen.	
CJ7: Coordinated clauses with *but, or*	 Join two sentences together with *but*. The second sentence is a surprise.	
CJ8/9: Coordinated phrases with *but not, or*	 Join two shapes the same together with *but not*. The first one happens, the second one doesn't.	
AD3: Adverbial subordinate clauses with temporal conjunctions	 The sentence in the triangle tells you when the main sentence (black line) happens. The main sentence happens 1st with *before*, 2nd with *after*, 2nd with *when* (but straight away, triangle	

before, after, when, until	starts it), 1st with *until* (triangle stops it). Doesn't matter if triangle appears second or first, meaning stays the same.	
AD4: Adverbial subordinate clauses with conditional conjunctions *if, unless*	 *If* works same as *when* and *unless* same as *until*, difference is they might never happen.	
RC1-4: Unembedded or presentational relative clauses	  Put a whole sentence inside an oval to give more information.	 

Supplemental information 1

Individual-level raw and standard scores on standardised measures

Assessment	ID code	Age at start	Raw score	Scaled/Standard score[1]
SPELT	1	9.79	21	<40
	2	9.75	14	<40
	3	8.70	15	<40
	4	9.61	42	89
	5	8.01	32	72
	6	9.32	32	64
	7	10.89	30	n/a
	8	8.94	39	82
ACE grammar	1	9.79	7	3
	2	9.75	2	3
	3	8.70	8	4
	4	9.61	8	3
	5	8.01	4	3
	6	9.32	5	3
	7	10.89	7	3
	8	8.94	9	5
ACE information	1	9.79	10	6
	2	9.75	8	5
	3	8.70	11	8
	4	9.61	10	6
	5	8.01	4	4
	6	9.32	5	3
	7	10.89	18	12
	8	8.94	5	4
ERRNI MLU	1	9.79	5.50	65
	2	9.75	4.10	64
	3	8.70	5.50	69
	4	9.61	6.17	69
	5	8.01	5.47	69
	6	9.32	4.16	64
	7	10.89	5.26	64
	8	8.94	7.92	90
ERRNI content	1	9.79	10	72
	2	9.75	10	72
	3	8.70	10	75
	4	9.61	6	65
	5	8.01	5	65

Assessment	ID code	Age at start	Raw score	Scaled/Standard score[1]
	6	9.32	6	65
	7	10.89	11	74
	8	8.94	11	75
TROG	1	9.79	7	62
	2	9.75	8	67
	3	8.70	2	55
	4	9.61	12	85
	5	8.01	4	55
	6	9.32	5	55
	7	10.89	9	65
	8	8.94	10	76

[1]Missing standard scores indicate that norms were unavailable for the child's age at the point of testing. ACE scores are scaled scores (1SD range: 7-13), all others are standard scores (1SD range: 85-115)

Supplemental information 2

Three example probe tests (score sheets only shown here, each probe test includes a PowerPoint presentation with pictures)

NG2 "not" after diamond

Model: "The banana is yellow but..... the apple is not yellow. Can you say that?"
Practice item: The man is walking but..... Target: *The lady is not walking*

Item	Target	Response	Score
1	this boy is not yawning		
2	this ball is not on the table		
3	this man is not clapping		
4	the lady is not angry		
5	this bird is not flying		
6	this house is not in the box		
7	the ice cream is not hot		
8	this girl is not under the bridge		
9	the ladder is not square		
10	the girl is not drinking		
		Score 0 if auxiliary/copula and/or "*not*" omitted, or if appear in wrong order **TOTAL**	

Q16 Who and what movement in questions

Model: Mrs Jones asks lots of questions. Today she is asking about the thing or person under the blue box (point): For example, “the girl is kicking something”. Mrs Jones asks ‘what’ – “What is the girl kicking?” Can you ask that?

Practice item: The man is pushing someone. Mrs Jones asks "who". She says…. Prompt: "who….." Target: *who is the man pushing?*

Type	Item	Target	Response	Score
SVO	1	Who is the man kissing?		
SVO	2	Who is the man chasing?		
PP:SVA	3	What is the mug on?		
SVO(A)	4	What is the lady pouring (into the cup)?		
SVOA	5	What is the lady putting in the machine?		
SVOA	6	What is the lady sticking on the wall?		
PP:SVOA	7	Who is she putting the nappy on?		
PP:SVOA	8	What are they putting the rubbish in?		
SVOO	9	What is the man feeding the baby?		
How:SVOA	10	What is she spreading the butter with?		
			Score n/a if no attempt to ask a question **TOTAL**	

AG3 Plural aux/cop past tense TESTING FOR USE OF PLURAL COP/ AUX (with plural subject)

NB. Do not test if child does not mark plurals in Subject

Model: We are going to talk about things that happened before now. We are going to talk about yesterday or last week.
"Yesterday, the puppies were sleeping". Can you say that?
Practice item: The boys were playing football.

Type	Item	Target	Response	Score
reg pl	1	The bottles were empty.		
reg pl	2	The flowers were gorgeous/alive/red.		
reg pl	3	The toys were on the bed.		
reg pl	4	The balls were on the table.		
reg pl	5	The girls were jumping.		
reg pl	6	The boys were reading a book.		
irreg pl	7	The people were listening (to music).		
irreg pl	8	The men were laughing.		
irreg pl	9	The people were clapping.		
irreg pl	10	The sheep were eating.		
Score 0 for singular past tense aux/cop with plural subject. Score N/A for pl. present tense aux/cop or omitted aux/cop.			**TOTAL**	

Supplemental information 3

Summary of timings within video-recorded intervention sessions (raw data in seconds)

Child and session no.	Recording length	Time on Target 1	Time on Target 2	Total time on targets	General intro	T1-T2 transition	Child distraction	Tech problems	Interruptions	Other	Total non-target time
ID6, 1	1233	610	528	1138	11	34	50	0	0	0	95
ID7, 1	1581	689	853	1542	28	11	0	0	0	0	39
ID5, 1	1917	1362	511	1873	40	4	0	0	0	0	44
ID2, 1	1663	980	528	1508	80	5	70	0	0	0	155
ID3, 2	1441	408	935	1343	82	16	0	0	0	0	98
ID4, 2	2319	1506	759	2265	39	0	15	0	0	0	54
ID6, 4	1487	517	816	1333	15	53	86	0	0	0	154
ID7, 2	1572	727	822	1549	15	8	0	0	0	0	23
ID1,4	1320	328	923	1251	11	24	0	34	0	0	69
ID8, 4	1297	756	432	1188	24	10	75	0	0	0	109
ID8, 6	1471	718	713	1431	30	10	0	0	0	0	40
ID5, 6	1375	188	645	833	88	42	412	0	0	0	542
ID2, 3	1718	1078	633	1711	7	0	0	0	0	0	7
ID3, 5	1321	852	421	1273	15	33	0	0	0	0	48

Child and session no.	Recording length	Time on Target 1	Time on Target 2	Total time on targets	General intro	T1-T2 transition	Child distraction	Tech problems	Interruptions	Other	Total non-target time
ID4, 5	1381	527	821	1348	4	29	0	0	0	0	33
ID1, 6	1118	311	719	1030	23	17	0	48	0	0	88
ID8, 10	1051	454	471	925	110	16	0	0	0	0	126
ID1, 13	1563	823	623	1446	108	9	0	0	0	0	117
ID7, 15	1637	811	740	1551	35	0	0	0	51	0	86
ID2, 14	1410	407	905	1312	48	25	0	0	0	25	98
ID6, 6	1626	698	777	1475	0	34	117	0	0	0	151
ID3, 7	1215	824	350	1174	41	0	0	0	0	0	41
ID4, 8	1671	726	917	1643	10	18	0	0	0	0	28
ID5, 9	1675	1219	247	1466	161	21	27	0	0	0	209
ID7, 19	1366	851	496	1347	19	0	0	0	0	0	19
ID3, 18	1035	682	273	955	38	17	25	0	0	0	80
Mean s	1479.35	732.77	648.38	1381.15	41.62	16.77	33.73	3.15	1.96	0.96	98.19
Min, s	24;38	12;12	10;48	23;01	00;42	00;17	00;34	00;03	00;02	00;01	01;38

Supplemental information 4

Fidelity regarding variability and uniqueness of lexical items

The variability and uniqueness of lexical items used in intervention was judged according to whether more than ten different root words were used for each morphological target or ten different verbs for each syntactic structure. This was achieved for 100% of targets and sessions, with the exception of one target (verbs requiring three obligatory arguments), where three out of five of the analyzed sessions working on this target did include ten or more different verbs and the others included nine or six different verbs respectively. This was achieved because, although all three arguments are obligatory for very few verbs (e.g., *I put the book on the table*), other verbs can use this structure, but the prepositional phrase is optional. Several of these verbs were included in the sessions e.g., *The elephant carried the bean bag (to the treasure chest)*. Optimal fidelity (100%) was achieved across all targets for ensuring that no more than half of items which appeared in probe tests were also employed during intervention sessions.

Footnotes

[1] Both the first and second authors filled out the spreadsheet for three of the participants and their agreement on whether a potential target required a probe test or not was 92%. The five other participants were each scored by just one of these two authors. Most disagreements were either because one author had missed a child production matching a target form or because when the child had produced an error, the authors had not always agreed on which target to mark the error. For example, for "then a parrot came and chattering to the monkey", one author had marked the error on the past tense target and the other on both past tense and verb phrase coordination. As the use of verb phrase coordination was in fact accurate, the subsequent agreed decision was to credit the participant with this. Identified differences in marking were resolved through discussion and scores for other participants amended as appropriate before the probe testing started.

[2] In cases of absence of a participant or the testing SLP on their test day, the first author carried out the testing (never on the same day as she delivered any intervention). The first author carried out tests on 24 days (often to catch up a single participant) and the sixth author on 66 days (usually with multiple participants).

[3] The main issues were around whether to score a particular utterance as incorrect versus not applicable, particularly when the participant had not attempted a particular construction. Decisions around this were added to scoresheets to ensure consistency. Examples of answers coded as “not applicable”: 1) the target was plural auxiliary are/were, but the participant used a singular subject, so a plural verb not required, or the auxiliary was missing entirely - these would both be coded as an error on a different target, 2) the target was use of adverbial clauses, but the participant failed to use the target subordinating

conjunction, 3) the target was formation of "wh" questions, but the participant did not attempt to ask a question (perhaps misunderstanding the task).

[4] On the occasions that the treating SLP was absent, sessions were covered by the first author (who devised the SHAPE CODING system). If the participant was absent, sessions were made up on another day where possible, either by the treating SLP or the first author. Over the whole study, the first author provided eight intervention sessions to six different participants (once or twice each).

[5] If recasting was needed (which was rare – see Table 5), this would mean that the adult had provided two models for one child correct production. The treating SLP would then not provide the initial model for a later attempt so that the total number of models during a session equalled the total number of child correct productions.

[6] The *total intervention period* was longer than the number of weekly sessions due to the inclusion of holidays (with a one-week mid-term break within each term and a two week break for Easter between the two terms) and some missed sessions.

[7] This participant seemed to get into a pattern with the probe test which he persisted with throughout. He asked questions (e.g., "if nobody answers, what shall she do?", "unless the car starts, what shall he do?") instead of producing a statement (e.g., "if nobody answers, she will leave the pizza", "he will call a mechanic, unless his car starts"). For unless, the questions were ungrammatical and were therefore scored as zero, but with if, the questions were grammatical, but not following the desired structure and therefore un-scorable.

[8] We have used the new target codes in this paper. This target therefore has a double code (CJ8 & CJ9) because it has now been split into CJ8 and CJ9.