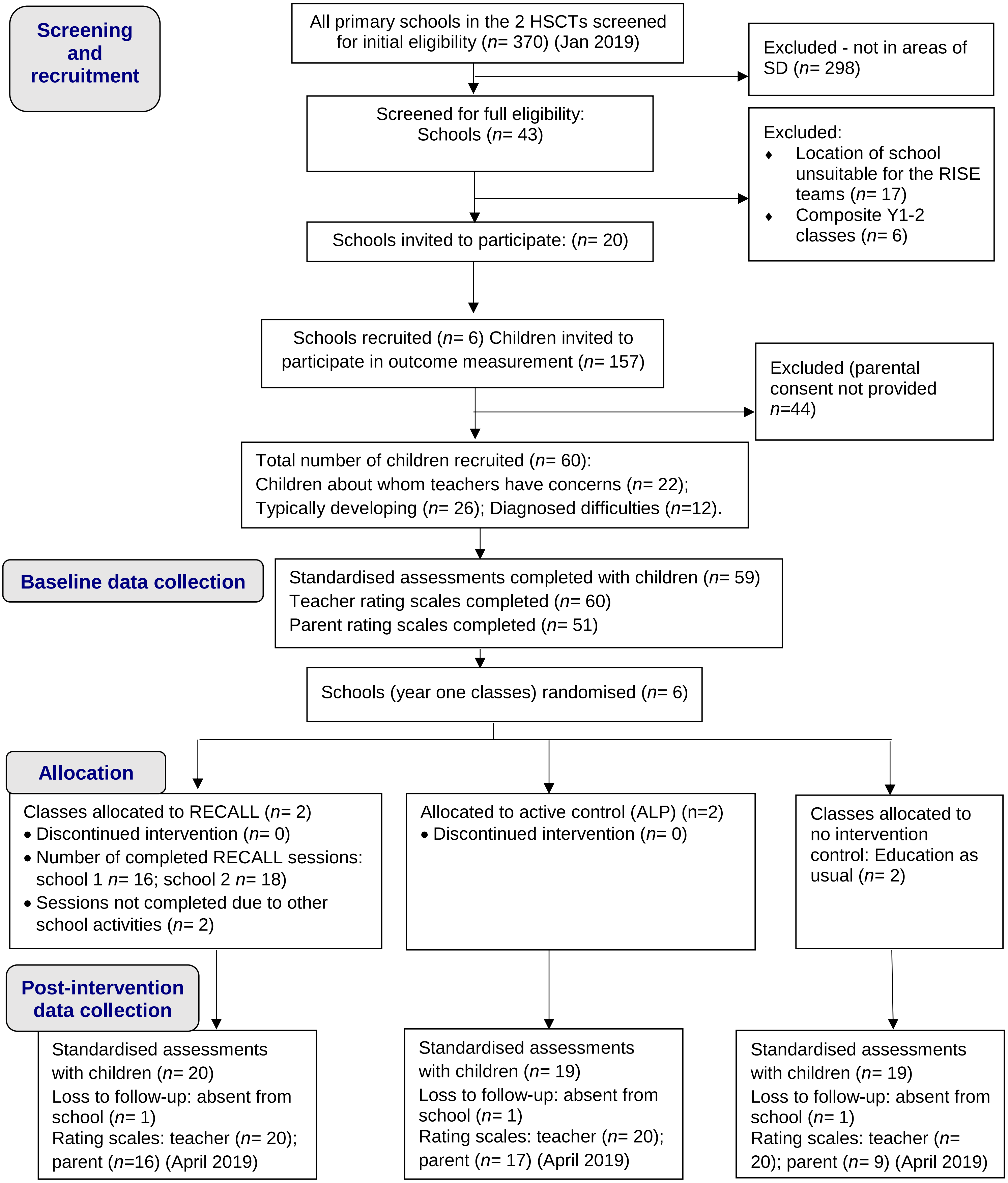
Participant Flow



Screening and recruitment

All primary schools in the 2 HSCTs screened for initial eligibility (*n*= 370) (Jan 2019)

Excluded - not in areas of SD (*n*= 298)

Screened for full eligibility: Schools (*n*= 43)

Excluded:
- Location of school unsuitable for the RISE teams (*n*= 17)
- Composite Y1-2 classes (*n*= 6)

Schools invited to participate: (*n*= 20)

Schools recruited (*n*= 6) Children invited to participate in outcome measurement (*n*= 157)

Excluded (parental consent not provided *n*=44)

Total number of children recruited (*n*= 60):
Children about whom teachers have concerns (*n*= 22);
Typically developing (*n*= 26); Diagnosed difficulties (*n*=12).

Baseline data collection

Standardised assessments completed with children (*n*= 59)
Teacher rating scales completed (*n*= 60)
Parent rating scales completed (*n*= 51)

Schools (year one classes) randomised (*n*= 6)

Allocation

Classes allocated to RECALL (*n*= 2)
- Discontinued intervention (*n*= 0)
- Number of completed RECALL sessions: school 1 *n*= 16; school 2 *n*= 18)
- Sessions not completed due to other school activities (*n*= 2)

Allocated to active control (ALP) (n=2)
- Discontinued intervention (*n*= 0)

Classes allocated to no intervention control: Education as usual (*n*= 2)

Post-intervention data collection

Standardised assessments with children (*n*= 20)
Loss to follow-up: absent from school (*n*= 1)
Rating scales: teacher (*n*= 20); parent (*n*=16) (April 2019)

Standardised assessments with children (*n*= 19)
Loss to follow-up: absent from school (*n*= 1)
Rating scales: teacher (*n*= 20); parent (*n*= 17) (April 2019)

Standardised assessments with children (*n*= 19)
Loss to follow-up: absent from school (*n*= 1)
Rating scales: teacher (*n*= 20); parent (*n*= 9) (April 2019)

Baseline Characteristics

Participants	**Recruitment targets**	**Number recruited**	**Characteristics**
Health professionals	*n*= 8	*n*= 8	Professional backgound: SLT (*n*= 4) OT (*n*= 2) PT (n= 1) SEB (n= 1)
Schools (clusters)	*n*= 6	*n*= 6	Social disadvantage ranking (based on data from the NIMDM 2017 [29]): Within lowest decile for their HSCT area (*n*= 3) Within lowest quintile for their HSCT area (*n*= 3)
Children recruited for outcome measurement	*n*= 60	*n*= 60	Gender: girls (*n*= 26, 43%); boys (*n*= 34, 57%) Age at baseline: 56 months to 67 months (mean = 61 months)
	n= 30 (50% of sample)	*n*= 22 (37%)	1) children about whom teachers had concerns around listening and communication skills
	n= 12 (20%)	*n*= 12 (20%)	2) children with diagnosed developmental or learning difficulties
	n= 18 (30%)	*n*= 26 (43%)	3) typically developing children who did not have any identified listening and communication problems as recognised by the teachers

Primary outcome measures

Primary outcome	Results	
1. Compliance	95% of RECALL	
2. Fidelity	RECALL sessions delivered by the health professionals (HPs): 76% fidelity to intervention protocol. For the RECALL sessions delivered by teachers, fidelity varied between the 2 schools: school 1 (76%); school 2 (45%)	
3. Acceptability	Qualitative data were collected via semi-structured interviews with teachers and health professionals (HPs). There were mixed findings regarding the acceptability of RECALL. All of the HPs and teachers liked the listening recall, fantastical play and phoneme awareness components of the intervention. None of the HPs or teachers liked the odd one out task in its current format and it was difficult to deliver to large groups (9-10) children.	
4. Recruitment, consent and sampling	4.1 Number and proportion of schools that meet the eligibility criteria	12% (43 of 370) of schools in study area met initial eligibility criteria re socio-economic status (see Figure 1 Study flow chart)
	4.2. Number of schools approached	*n*=20
	4.3. Number of schools where consent is obtained from principals and teachers	Recruitment target achieved (*n*=6)
	4.4. Number and proportion of children identified by teachers in each of the 3 sub-groups	Teachers did not always know whether children did/did not have a diagnosis but were able to identify appropriate numbers in each sub-group (Table 2)
	4.5. Number and proportion of parents who consent	Overall rate of parental consent: 72% Some parents of children about whom teachers had concerns did not consent and the desired proportion of children in this sub-group was not achieved (*n*= 22, 37% compared to the target of *n*= 30, 50%) See Table 2 for further detail of participant characteristics and recruitment rates.
5. Attendance and loss to follow-up	5.1. Number of completed interventions	100% of interventions completed No schools dropped out of the study
	5.2. Number of completed standardised assessments, teacher rating scales and parent rating scales at post-intervention and three-month follow-up.	97% of assessments with children completed post-intervention (3% loss to follow-up) Teacher rating scales: 100% completed at both time points Parent rating scales: 70% completed Three month follow- up – this was not completed following an amendment to the study protocol.
6. Acceptability of randomisation	6.1. Consent rates	Achievement of school recruitment targets indicated that randomisation was acceptable to schools

	6.2. Reasons given for participation and non-participation by school prinicipals	Other initiatives taking place in school at time of the study
	6.3. Qualitative data gathered in the semi-structured interviews	No concerns about randomization raised by teachers during post-intervention interviews
7.Acceptability of active control intervention as a comparator to RECALL	7.1. Health professionals' perspectives on similarities/differences between the programmes, explored in the semi-structured interviews	The active control condition differed sufficiently in content from the experimental RECALL intervention but took the same amount of time to deliver, meaning it appears to be an appropriate comparator for a full-scale trial. Descriptive statistics of the children's results suggested that there may be differences between intervention groups which also supports the use of this intervention in a full-scale trial (see Tables 4 and 5)
	7.2. Observations of delivery by research team	
8. Exploration of education as usual	8.1 Qualitative data from semi-structured interviews	Teachers reported that the components of RECALL differ from the tasks delivered typically in their usual practice (education as usual). Therefore it would be appropriate to investigate RECALL in a full-scale trial.
9. Acceptability of outcome measures for the children, teachers and HPs	9.1. Number of completed assessments for each child at each time point	Baseline: 100% (all measures completed with full sample of children (*n*= 60) Post-intervention: all measures completed with 97% of children (*n*= 58) (See Table 3 for full list of these secondary outcome measures)
	9.2. Number lost to follow-up and reasons why if possible	2 children (out of 60, 3%) were absent from school due to sickness so were not assessed at the post-intervention time point
	9.3. Qualitative data obtained in semi-structured interviews	Research Assistants (RAs) reported that administering the full battery of assessments with each child was time-consuming (on average more than one hour per child). In particular, the New Reynell Developmental Language Scales (NRDLS) took a considerable amount of time to complete, whereas the Clinical Evaluation of Language Fundamentals- Preschool (CELF-P) (trialled in one school for comparison) was much quicker to administer. The RAs found it difficult to observe and simultaneously record the children's performance for the auditory attention and statue subtests of the Developmental Neuropsychological Assessment (NEPSY-II), Therefore, they doubted the accuracy of their scoring. If this test were used in a full trial, thorough training and practice should be provided to those administering it and inter-rater reliability must be measured.
10. Unexpected adverse effects,	There were no adverse events associated with this trial	

recorded by the health professionals and teachers	
11. Whether blinding is maintained at end of study, investigated in the semi-structured interviews	The teachers in the RECALL group reported that due to the nature of the tasks they were aware that it was the experimental intervention. Teachers in the active control group remained blinded to their allocation. The outcomes assessors also remained blinded.

Secondary outcome measures

Outcome measured	Skill	Standardised assessment
Trained task	Trained WM tasks	**Automated Working Memory Assessment (AWMA) [40]** • A computerised assessment administered using a laptop • 2 subtests administered in all 6 schools (*n*= 60 children): - Listening recall - Odd one out
Trained task	Phoneme awareness	**The Preschool and Primary Inventory of Phonological Awareness (PIPA) [37]** • A standardised assessment consisting of 6 subtests for children aged 3 years to 6 years 11 months • 2 subtests trialled: - Phoneme isolation subtest (administered in 5 schools, *n*= 50 children) - Phoneme segmentation subtest (administered in 1 school, *n*= 10 children)
Near-transfer	Untrained WM tasks	**Automated Working Memory Assessment** (detailed above) [40] • 4 further subtests administered in all 6 schools (*n*= 60 children): - digit recall - block recall - counting recall - non-word recall
Far-transfer	Attention	**NEPSY-II – A Developmental Neuropsychological Assessment (NEPSY) [41]** • Includes standardised performance-based measures of attention for children under 6 years • 2 subtests administered in all 6 schools (*n*= 60 children) - Auditory attention - Statue
	Language	**The New Reynell Developmental Language Scales (NRDLS) [38]** • A standardised assessment for children aged between 3 years and 7 years 6 months. • Comprehension scale administered in 5 schools (*n*= 50 children) **Clinical Evaluation of Language Fundamentals- Preschool (CELF-P) [39]** • A standardised assessment for 3 – 6 year olds that examines children's: understanding and use of syntax (grammar/sentence structure), semantics (word meanings) and grammatical morphology (markers of grammatical relationships • Core language subtests (*n*= 10) conducted in 1 school (*n*= 10)
	Behaviour in the classroom	**Behaviour Rating Scale of Executive-Function- Preschool Version (BRIEF-P) [42]** (*n*= 60) • A standardised, validated scale completed by teachers • Includes consisting of 63 items that can be used with children from 2 years to 5 years 11 months to measure behavioural characteristics associated with executive function skills including WM • Completed by teachers in all 6 schools (*n*= 60 children)
	Communication skills at home	**The Focus on Communication Outcomes Under Six – 34 (FOCUS-34) [43] (*n*= 60)** • A checklist of children's communication skills at home completed by parents to measure change over time

		- Completed by parents in all 6 schools (*n*= 60 children)

Descriptive statistics for raw scores at baseline for the full and stratified samples

Outcome Measure			Full sample (*n*= 60)		Split sample				Interpretation of results
					Typically developing group (*n*= 26)		Concerns group (*n*= 34)		
Outcome	**Task**	**Test**	**Mean (SD)**	**Skewness**[1]	**Mean (SD)**	**Skewness**	**Mean (SD)**	**Skewness**	
Trained task	Listening recall	AWMA[2]	1.16 (1.68)	1.57	1.58 (1.98)	1.08	.81 (1.33)	2.21	Both groups: scores highly skewed towards the low end- potential floor effects
	Odd one out	AWMA	7.16 (3.54)	.28	7.88 (3.98)	-.13	6.56 (3.07)	.61	Both groups - distribution approximates normality
	Phoneme awareness	PIPA Phoneme isolation*	8.90 (4.04)	-1.08	10.08 (3.75)	-1.99	7.81 (4.05)	-.60	Full sample and TD group: highly skewed towards high scores - potential ceiling effects. Children with concerns- moderately skewed
		PIPA Phoneme segmentation†	.30 (.675)	2.28	.00	-	.38 (.74)	1.95	Both groups: highly skewed towards the low end - potential floor effects.
Near-transfer (untrained WM)	Digit recall	AWMA	18.24 (4.96)	-.46	18.69 (6.41)	-.80	17.88 (3.42)	.66	Children with concerns: moderate skewness towards high scores for digit recall and counting recall.
	Block recall	AWMA	10.74 (3.24)	-.28	11.65 (3.90)	-.73	10.00 (2.41)	-.39	
	Counting recall	AWMA	6.21 (3.19)	-.42	7.00 (3.60)	-.53	5.56 (2.71)	-.90	
	Nonword recall	AWMA	4.52 (3.56)	.15	3.96 (3.14)	.03	4.97 (3.86)	.08	
Far-transfer	Auditory Attention	NEPSY-II	19.54 (6.22)	-.69	20.62 (6.47)	-.94	18.70 (5.98)	-.62	Both groups: moderate skewness towards high scores
	Statue	NEPSY-II	22.64 (5.58)	-.81	25.69 (2.95)	-.61	20.24 (6.02)	-.23	Full sample: moderate skewness towards high end. Concerns group- approximates normality.
	Language	NRDLS*	61.06 (4.58)	-.74	62.75 (3.25)	.29	59.50 (5.11)	-.57	Both groups: NRDLS scores moderately skewed towards high performance; CELF-P scores moderately skewed towards lower end
		CELF-P† (Cumulative Raw Scores)	55.40 (8.53)	.44	61.5 (10.61)	-	53.80 (8.01)	.56	
	Behaviour in the classroom	BRIEF-P[3] Global Executive Composite	99.57 (30.21)	.90	88.73 (32.13)	1.67	107.85 (26.20)	.76	For both scales of this measure: scores are highly skewed to lower end (indicating better performance) for the TD group but not for the concerns group.
		BRIEF-P WM scale	62.20 (15.56)	.52	25.27 (10.48)	1.32	31.7 (8.34)	.26	
	Communication skills at home	FOCUS 34 baseline	189.39 (39.76)	-1.52	204.05 (36.76)	-2.51	179.13 (39.11)	-1.35	Highly skewed for full sample and both groups but to a greater degree for TD

[1] Skewness: 0=perfect normality; negative skewness values indicate a clustering of scores at the high end; positive skewness values indicate clustering at the low end (except on the BRIEF-P (Gioia et al., 2003) where lower scores indicate greater degrees of executive dysfunction so positive skewness = clustering of scores at the high end. Shaded cells =highly skewed values (>1 or <-1)

[2] Raw scores on AWMA subtests represent the number of trials correct (rather than memory span)

Baseline and post-intervention mean and standard deviations for raw scores at baseline and post-intervention (per group) for full sample (*n*= 60)

Outcome	Task	Test used	Time point	RECALL (*n*= 20)	RISE Active Control (*n*= 20)	No Intervention (*n*= 20)
				Mean (SD)	**Mean (SD)**	**Mean (SD)**
Trained task	Listening recall (ELWM)	AWMA	Baseline	.47 (.77)	1.22 (1.83)	1.41 (1.66)
			Post-intervention	4.11(3.12)	5.28 (4.51)	2.35 (3.74)
	Odd one out (ELWM)	AWMA	Baseline	7.00 (3.13)	5.94 (3.11)	8.06 (4.13)
			Post-intervention	8.42 (3.16)	10.44 (3.09)	9.24 (4.49)
	Phoneme awareness	PIPA Phoneme isolation subtest*	Baseline	6.33 (5.32)	9.47 (2.97)	9.05 (4.21)
			Post-intervention	7.56 (3.64)	9.63 (3.40)	7.16 (3.85)
		PIPA Phoneme segmentation subtest†	Baseline	.30 (.21)	-	-
			Post-intervention	2.10 (.31)	-	-
Near-transfer (untrained WM)	Digit recall	AWMA	Baseline	16.58 (5.78)	19.78 (4.25)	17.59 (4.32)
			Post-intervention	19.37 (4.04)	18.61 (4.64)	18.29 (4.95)
	Block recall	AWMA	Baseline	11.05 (2.80)	10.28 (3.48)	10.76 (3.7)
			Post-intervention	11.05 (2.55)	10.56 (5.22)	9.41 (3.97)
	Counting recall	AWMA	Baseline	16.58 (5.78)	19.78 (4.25)	17.59 (4.32)
			Post-intervention	19.37 (4.04)	18.61 (4.64)	18.29 (4.95)
	Nonword recall	AWMA	Baseline	3.58 (3.61)	6.39 (2.97)	3.35 (3.37)
			Post-intervention	7.26 (2.88)	8.61(4.13)	6.65 (3.23)
Far-transfer	Auditory Attention	NEPSY-II	Baseline	18.00 (6.29)	20.05 (6.69)	21.59 (5.83)
			Post-intervention	17.47 (6.78)	21.68 (5.45)	19.82 (5.33)
	Statue	NEPSY-II	Baseline	21.32 (5.82)	22.47 (6.01)	23.72 (5.13)
			Post-intervention	26.37 (4.14)	26.47 (6.60)	23.72 (5.10)
	Language	NRDLS Comprehension Scale*	Baseline	60.56 (5.72)	61.47 (3.79)	60.35 (4.76)
			Post-intervention	62.33 (2.74)	62.95 (2.70)	61.35 (5.15)
		CELF-P† (Cumulative Raw Scores)	Baseline	55.4 (8.53)	-	-
			Post-intervention	57.00 (7.24)	-	-
	Behaviour in the classroom	BRIEF-P[4] Global Executive Composite	Baseline	60.20 (12.61)	57.55 (14.40)	68.85 (17.66)
			Post-intervention	57.45 (11.68)	50.70 (9.75)	63.45 (15.40)
		BRIEF-P Working memory scale	Baseline	27.9 (7.37)	25.85 (9.63)	33.00 (11.11)
			Post-intervention	25.55 (6.97)	22.95 (6.02)	29.80 (9.62)
	Communication skills at home	FOCUS-34 (Change score)	Post-intervention minus baseline	13.46 (21.70)	12.58 (18.38)	2.12 (10.23)

[4] Note: higher scores on the BRIEF-P [42] indicate greater degrees of executive dysfunction. A reduction in scores over time indicates improvement. For tests marked* sample (*n*= 50); for tests marked† sample (*n*= 10). For the FOCUS-34 [43] change scores of >11 points indicate significant clinical change.