ORIGINAL ARTICLE

Eye-Gaze Training Does Not Impact Performance on a Nonverbal Test of Reading Abilities for Children with Special Educational Needs

Sharon Arnold[1] · Phil Reed[1]

© The Author(s) 2024

Abstract

Many individuals in special education have difficulty with language, and little is known about their reading abilities due to challenges presented for assessment. A nonverbal word recognition test based was developed by Measuring the word recognition abilities of children who are (2019), which proved more sensitive than traditional reading tests. The current Experiment 1 replicated that result for children in special education, and noted that the modified reading test was especially sensitive for those who do not score well in traditional versions of reading and language assessments. However, removing the motor response required by the digitised response, in favour of using eye-tracking responses, did not aid assessment of reading performance. Experiment 2 tested the notion that it could be that participants unfamiliar with use of eye gaze tracking facilities may not respond well to such a test. Reading scores for two groups of pupils in special education were compared: a group who took part in 'primer activities' designed to encourage familiarity with eye gaze tracking, and a group who received no such intervention. There was no impact of eye gaze training on the results of the eye-gaze modulated modified nonverbal reading test, suggesting that manual responses are the most effective way of assessing word recognition.

Keywords Reading · Nonverbal · Special needs

It can be difficult to gain an accurate picture of the literacy abilities of children with special educational needs who are non-verbal or minimally verbal, and developing assessment tools is a pressing need to allow accurate reflection of their abilities and appropriate planning of educational provision (Tager-Flusberg et al., 2017).

✉ Phil Reed
    p.reed@swansea.ac.uk

[1] Department of Psychology, Swansea University, Singleton Park, Swansea SA2 8PP, U.K.

🖄 Springer

To facilitate the assessment of reading abilities of such children, Arnold and Reed (2019, 2024) developed and employed a digitised version of the New Salford Reading Test (NSRT), which utilised a multiple-choice response on a touch screen facility to replace the traditional verbal response requirement. Once the requirement for verbalisation was eradicated, word recognition scores on the modified version of the NSRT (Arnold & Reed, 2019), as well as listening comprehension scores (Arnold & Reed, 2024), were improved. The conclusion was that this modified digital version of the NSRT was an effective form of assessment for children in special educational settings, who do not score well in traditional forms of reading assessment.

A limitation to the new nonverbal version of the reading task was that, whether due to physical limitation, or comorbidity, some children (especially those with ASD) struggled to develop pointing skills to complete the nonverbal version of the task (see also Paquet et al., 2016; Shetreat-Klein et al., 2014). Eye-gaze tracking has the benefit of requiring neither verbal response nor pointing ability, and so it has been suggested that the use of eye-tracking facility as a tool could form part of a solution for gathering assessment data with this population (Plesa-Skwerer et al., 2016; Tager-Flusberg & Kasari 2013).

The aim of the current series of studies was initially, in Experiment 1, to re-evaluate the use of the modified version of the NSRT compared to the traditional paper version, and to assess its relative sensitivity for those with a range of verbal and reading abilities. Experiment 1, additionally, compared the results of the traditional paper and modified NSRT to those from a version using eye-gaze as a response. Experiment 2 developed this investigation further by assessing the degree to which eye-gaze training prior to undertaking the eye-gaze version of the reading assessment would aid assessment of reading.

Experiment 1

The first experiment was conducted to replicate the results reported by Arnold and Reed (2019) with respect to the advantage of a digitised version of the NSRT compared to traditional paper versions. It also sought to establish if this advantage would be greater as a function of poorer verbal ability. In addition, it sought to establish whether a further modified version of the digitised reading test that did not rely on pointing, but used eye movements to track responses, would confer any additional advantage for those with poor verbal ability.

Method

Participants

Participants were recruited from special schools in south Wales. Special school attendance is deemed appropriate by a Local Authority when the nature of pupils' needs is so severe or complex as to make unlikely suitable educational progress in a mainstream setting. All children attending such schools have a statement of special edu-

cational needs following assessment by an Educational Psychologist. Of 22 special schools contacted, 9 agreed to become involved in the study. The special schools were given the remit of identifying pupils who, in the opinion of their teachers and ongoing classroom assessment procedures, could possibly be readers. Participation in the study was based on pupil availability, but had the following inclusion criteria: the pupil had a statement of special educational needs; the pupil be attending the special school only (and not also attending mainstream); the pupil could possibly be a reader according to teacher assessment; the pupil be aged between 5 and 16 years old; and the pupil be a native English speaker. There were no explicit exclusion criteria, other than not fulfilling the above inclusion criteria. Ethical permission for the research was gained from the University Psychology Department Ethics Committee.

A total of 93 participants (82 male; 11 female) were recruited. All participants had a statement of special educational needs. There were 61 pupils with an additional diagnosis of ASD. The diagnosis was made by a paediatrician who was independent from this study, and based on DSM criteria, psychometric testing, and clinical judgment. The reaming 32 participants had a learning disability, or Down Syndrome, diagnosed independently by a paediatrician using the above criteria. The mean age of the sample was 136.12 (SD$\pm$36.18 months). The sample mean composite IQ was 56.97 ($\pm$13.22), with a verbal IQ of 56.94 ($\pm$14.75), and a perceptual reasoning IQ of 61.71 ($\pm$14.01). The mean BPVS score of the participants was 48.73 ($\pm$35.23).

Materials

Wechsler Abbreviated Scale of Intelligence (WASI; Wechsler, 2003) was used to determine IQ scores through administering four subtests: vocabulary and similarities which make up the verbal element of the test; and block design and matrix reasoning, which make up the nonverbal element. In terms of reliability, the alpha of the revised scale was recorded at 0.81 (Wechsler, 2003).

New Salford Reading Test (NSRT; McCarty & Lallaway, 2012) consists of three forms, each made up of 17 sentences. The forms are equally graded in terms of word reading difficulty. The words increase in difficulty as the student reads through the test. On the sixth error, the test is stopped, and the reading age determined by cross-referencing the 6th incorrectly read word on the scoring form with the reading age cited below it. Included in the sentences are words that are 'not counted'. Words that are not counted, are words that are not used to determine a reading age but have been inserted into the test to provide the sentence with structure. There are a total of 114 counted words in the test. The New Salford Reading Test is designed for use in schools that would likely deliver the test year on year to the same students. Therefore, the forms are alternated when delivered to avoid practise effects. Therefore, three forms in total are available, labelled as 'a', 'b', and 'c'. Only one form, therefore, was needed to be delivered to each participant. The forms are equally graded. In terms of reliability, the alpha of the revised scale was 0.97 for this sample.

Modified Word Recognition Test. The NSRT was modified as described by Arnold and Reed (2019, 2024). The counted words were taken from the test as a context-free word recognition items presented as a multiple-choice test. The words were presented on PowerPoint slides, for use with either a desktop computer, iPad, or tablet.

Multiple-choice was a necessary adjustment to facilitate non-verbal responses. On each slide, there was an equal number of incorrect answers available as there were correct. As an example, if the sentence in the NSRT contained 4 counted words, the slide would contain those four counted words with four incorrect words, so a total of 8 words on the slide. To ensure that the incorrect answers had the same grade of difficulty as the correct answers, these words were also taken from the NSRT. Each slide represented one sentence of the test, therefore there were 17 slides in total. There is a total of 114 counted words in the test. To score the test, the same NSRT score sheet which was used for the paper version of the test was used for the modified version. Therefore, on the sixth error, the test was stopped, and the reading age determined by cross-referencing the 6th incorrectly read word on the scoring form with the reading age cited below it. In terms of reliability, the alpha of the revised scale was 0.96 for this sample.

Modified word recognition test with eye gaze facility. The test materials used for the eye gaze word recognition test were the same as those used for the modified test. However, as opposed to touch screen facility, a Tobii Eye Tracker (4 C) was connected via USB port to a school laptop computer. This enables the Eye Tracker software provided with the device to track the gaze of the eyes i.e. any eye movements will move the cursor on the screen, tracking head and eye simultaneously. The Tobii Eye Tracker (4 C) is provided with 'track status' software which is used to determine the optimum distance and angle for the user and a brief calibration test to ensure the eye gaze is being tracked accurately. In terms of reliability, the alpha of the revised scale was.98 for this sample.

Procedure

All assessments were conducted within the school environment as part of the pupils' routine programme of study by the first researcher who, as a qualified teacher, has many years of experience teaching children with special needs. Assessments were carried out over a period of three days, sometimes two, if the participant was particularly able with the tests. Usually, these days were consecutive, however, where this was not possible (due to attendance for example), all tests were carried out within a two-week window. In the first test sitting with each pupil, IQ scores were obtained by the researcher who is experienced in administering the WASI.

The reading tests were then administered by the researcher, in a second and third sitting over the course of the two days. All participants experienced the same conditions, however, the order in which they did the traditional paper-based test (NSRT) format, or one of the two modified versions, was randomised. Participants were also divided as equally as possible in terms of which reading test forms (a, b, or c) would be completed as which version of the test. All tests were administered in the same environment as one another. In the interests of promoting 'good practice,' there was nothing in either test to communicate to the participant whether their answers were correct or incorrect.

For the NSRT version of the test, participants were asked to read the sentences aloud. Whenever the participant misread a word, a note was made on the form, and once the 6th error was made the test was terminated. In line with procedures for the

NSRT, if a participant was unable to read the first word for which a reading age was shown on the record sheet, a score of 0 was recorded (McCarty & Lallaway, 2012, 2012, p.9). If a participant read a word incorrectly but spontaneously corrected the mistake, this was not counted as an error. However, if a participant was unable to produce a word after 6 s, the researcher would supply the correct word and record this as an error.

The modified version of the test was carried out with the multiple-choice options being shown on the screen and the NSRT recording sheet being completed in the same way as it had been for the NSRT version; the rules for determining errors were also the same. The counted words were presented as context-free word recognition items using PowerPoint slides. Depending on the space available in each school, the test was usually delivered using a desktop or laptop computer with a touch screen facility, or an iPad or tablet no smaller than an iPad 2 model (1024×768). Starting at the first sentence on the form, the researcher would give the instruction "show me [word from the sentence]". The participant was expected to then point to what they believed to be the corresponding words on the screen. If the participant gave no response, the researcher would change the instruction to "point to" or "where is" to ensure the participant understood the instruction. When the participant touched a word on the screen, there was no response elicited from the screen to indicate whether or not the response was correct or incorrect. In line with the guidance from the NSRT manual, if a participant identified a word incorrectly but spontaneously corrected the mistake, this was not counted as an error. If a participant was unable to identify a word after 6 s, the researcher pointed to the correct word and moved on with the test. Once the participant made their 6th error the test was stopped. The errors were marked on the NSRT scoring sheet which determined the 'reading age' of the participant. In line with procedures for the NSRT, if a participant was unable to read the first word for which a reading age was shown on the record sheet, a score of 0 was recorded (NSRT; (McCarty & Lallaway, 2012). As the aim of the current study was to explore context-free word recognition, scores were recorded as word recognition scores (WRS) and not reading ages.

For the modified version of the test with eye gaze facility, the tracking status box was open on the screen, and participants' seating adjusted to ensure that they were sat at the correct angle and distance from the screen. The calibration test was then run to ensure that the participant's gaze was being accurately tracked. Following this, the PowerPoint with the multiple-choice options were presented on the screen and the NSRT recording sheet was completed in the same way as it would be for the NSRT version; the rules for determining errors were also the same. The participant was instructed to 'only look' at the word on the screen that they believed to be the words on the screen which corresponded with those being read out by the researcher. If the participant gave no response, the researcher would change the instruction to "where is" to ensure the participant understood the instruction. The words were selected using a 'dwell time' with the duration set using the eye tracker settings facilities to 800 milliseconds. This meant that after the participant rested their gaze on a word a circle would begin to appear and then be drawn on the tile. The tab would then enlarge slightly and then revert to size and a small sound 'click' sound was emitted from the laptop, indicating that the word had been selected. These effects were achieved using

basic PowerPoint animation effects. The sound emitted was the same whether the answer was correct or not. A correct mark was allocated when participants correctly selected the target word. However, if a participant was unable to detect a target word, or fixated on the incorrect word, this was counted as an error. In line with the NSRT procedures, the test was stopped after the participant made a sixth error. No feedback was given regarding the accuracy of the response in both parts of the study, but all participants were praised and thanked for their participation.

Results and Discussion

Figure 1 shows the mean reading scores for the three versions of the test. Inspection of these data reveals that the modified NSRT gave the higher reading scores compared to the non-modified NSRT and the eye-gaze version of the modified NSRT. The latter two versions dd not differ from one another. A one-way repeated measures analysis of variance (ANOVA) conducted on these data revealed a significant effect of test, $F(2,184)=35.00$, $p<.001$, $\eta^2_p=0.276$. Paired t-tests, with a Bonferroni correction ($p=.05/3=0.016$), revealed that the modified version of the NSRT produced a higher reading score than the traditional paper version, $t(92)=8.46$, $p<.001$, $d=0.887$, and the eye gaze version of the modified NSRT, $t(92)=6.29$, $p<.001$, $d=0.652$, but that while the eye gaze version produced a numerically higher readings score than the paper version, this difference was not statistically reliable, $t(92)=1.86$, $p=.033$, $d=0.193$.

The Pearson correlations between the three version of the test revealed strong and significant positive relationships between the obtained reading scores: NSRT and MNSRT, $r=.633$, $p<.001$; NSRT and EG-MNSRT, $r=.649$, $p<.001$; MNSRT and EG-MNSRT, $r=.741$, $p<.001$. In addition, the relationships between the scores on the three versions of the reading tests and the measures of intellectual function (composite, verbal, and perceptual reasoning), as well as the BPVS, were examined, and these are shown in Table 1. Applying a Bonferroni correction ($p=.05/12=0.004$), all of the measures correlated with the reading test scores from each test, with the exception of perceptual IQ.

Table 2 presents the results of an analysis of the relationship between the difference between the paper version of the NSRT and the digital (MNSRT) and the eye gaze MNSRT. To calculate this the reading score obtained form the paper test (NSRT)

Fig. 1 Experiment 1. Mean reading scores on the three versions of the test. NSRT=traditional paper version of the New Salford Reading Test. MNSRT=digitised version of the NSRT. EG-MNSRT=eye gaze version of the modified NSRT. Error bars=95% confidence intervals
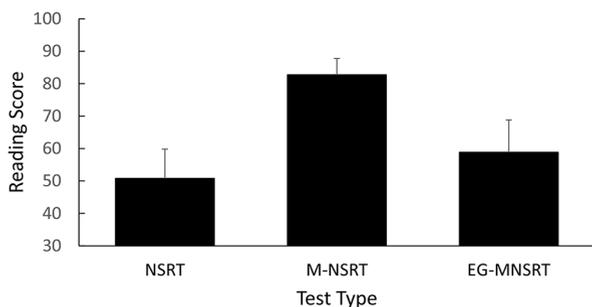
Table 1 Experiment 1: Pearson correlations between three reading test scores and the measures of IQ (Wechsler) and reading (BPVS) NSRT=traditional paper version of the New Salford Reading Test. MNSRT=digitised version of the NSRT. EG-MNSRT=eye gaze version of the modified NSRT

	Composite IQ	Verbal IQ	Perceptual IQ	BPVS
NSRT	0.454***	0.516***	0.284	0.586***
MNSRT	0.375***	0.304***	0.280	0.343***
EG-MNSRT	0.389***	0.374***	0.258	0.320***
Differences		NSRT>MNSRT		NSRT>MNSRT; NSRT > EG-MNSRT

***$p<.001$

Table 2 Experiment 1: Pearson correlations between the difference between the reading test scores obtained on the paper and digitised, and paper and eye gaze, versions of the NSRT and measures of IQ (Wechsler) and reading (BPVS) NSRT=traditional paper version of the New Salford Reading Test. MNSRT=digitised version of the NSRT. EG-MNSRT=eye gaze version of the modified NSRT

	Composite IQ	Verbal IQ	Perceptual IQ	BPVS
MNSRT– NSRT	0.308**	−0.440***	−0.159	0.513***
EG-MNSRT - NSRT	−0.027	−0.115	0.002	−0.272
Differences		NSRT>MNSRT		NSRT>MNSRT; NSRT > EG-MNSRT

***$p<.001$

was subtracted from the MNSRT and from the EG-NSRT scores. Higher scores mean that the obtained reading score was greater in the modified versions. Applying a Bonferroni correction ($p=.05/8=0.006$), the difference between the modified and the paper versions was not related to perceptual IQ measures, but it was significantly negatively related to composite and verbal IQ, and to the reading scores obtained in BPVS. The difference between the paper and eye gaze versions was not related to any of the measures. This suggests the impact of the modified version of the NSRT is greatest for those children who do not score well on traditional tests of verbal or reading ability.

With respect to the sensitivity of the modified test to detected reading ability, these data replicate those reported by Arnold and Reed (2019). Reading scores were higher for pupils on the MNSRT than for the NSRT. Moreover, the impact of the MNSRT was greatest for pupils who do not score well on traditional paper versions of verbal or reading ability. However, these data suggest that the eye gaze modification did not produce any greater improvement in scores than the digitised version that relied on pointing.

A disadvantage of eye gaze tracking can be data loss. This is when the eye tracking device fails to report the eye gaze position. This can be a calibration issue, which is usually easy to fix by recalibrating the device, or it can occur when a participant is looking outside the tracking area, usually a computer screen. This is more likely to occur when a participant has difficulty in understanding that the responses on the screen are being performed through their eye gaze.

Experiment 2

Although there were no statistically reliable effects of the eye gaze modification to the digitised NSRT in Experiment 1, the eye gaze tracking for participants with lower verbal abilities did improve word recognition. However, scores were most improved with access to a touch screen device. One possible reason that was considered for this. was a lack of familiarity on the part of participants with the eye-gaze facility which could cause data loss to occur. Given this, the simple aim of Experiment 2 was to evaluate whether eye-gaze training prior to use in the digitised reading test would improve word recognition scores for nonverbal children. Word recognition scores for two groups of participants were compared: one took part in 'primer activities' designed to encourage familiarity with eye-gaze tracking; and the other received no such intervention. If familiarity with the equipment allows participants to gain understanding that their eye-gaze was the required response, and perhaps allows them to understand the instructions more clearly, data loss would be less likely, and word recognition scores for the primer group would be improved.

Method

Participants

Potential participants were recruited as described in Experiment 1. Ethical permission for the research was gained from the University Psychology Department Ethics Committee. Thirty-three participants (29 male; 4 female) were recruited, and all had a statement of special educational needs. There were 25/33 pupils with an additional diagnosis of ASD (13 nonverbal; 12 verbal) as described in Experiment 1. The reaming 8 participants had a learning disability, diagnosed independently by a paediatrician using the above criteria, but who did not have ASD.

Participants were randomly assigned to either a 'primer' or 'non-primer' group. There were 17 participants in the non-primer group, with a mean chronological age of 141 months (SD ± 40.86; range = 59–186); and 16 participants in the primer group, who had a mean chronological age of 143 months (± 34.94; range 80–181). Table 3 shows the group-mean scores for both verbal and perceptual reasoning IQ for the non-primer and primer groups. Independent t-tests were performed on all of these scores revealed no significant differences between the groups on these measures, all ps > 0.3.

Table 3 Experiment 2. Group mean verbal and perceptual reasoning IQ scores (Weschler) for the verbal and nonverbal pupils in the primer and non-primer groups

	Verbal IQ			Perceptual Reasoning		
	Mean	SD	Range	Mean	SD	Range
Primer n=16	71.06	24.00	45–106	73.06	18.83	45–119
Non-Primer n=17	72.87	22.55	45–108	80.65	23.37	45–100

Materials

The psychometric tests were as described in Experiment 1, as was the EG-MNSRT procedure.

Eye gaze games. On-screen games were presented to the primer group prior to the word recognition test. There were two online eye-gaze games ('Colourful Caterpillar' and 'Catch Me' from 'tobiidynavox.com'). Each required participants to use their eye movement to move, target, or select, objects in the game. The first criterion for choosing these games was that they were designed specifically for use with the Tobii Eye Tracker. This was done to lessen the possibility of any 'off-putting glitches', or problems, when they were used with the pupils. Therefore, games from tobiidynavox.com were selected. The second criterion was simplicity of engagement. As it was likely that the range of cognitive abilities for the participants was going to vary, games were chosen that were based on simple cause-and-effect rules. Thirdly, the games were chosen to be colourful and/or appealing to the eye, and to appear to have an element of fun. Both games were located in the 'screen engagement' section of games on the site, and have been designed to encourage familiarity and engagement with the use of an eye-tracking facility. For the options, both games were set on 'classic' and 'easy' mode.

'Colourful Caterpillar' (http://www.tobiidynavox-webgames.com/sensory-game/sensory-game.html#settings) displays colourful circles on the screen, linked together to give the impression of a caterpillar. Once the eye-tracker equipment is attached to the screen, the caterpillar can be moved with eye-gaze. 'Catch Me' (http://www.tobiidynavox-webgames.com/catchme/catchme.html#1) involves a cartoon character which 'hides in the dark' of the screen, until it is found by the eye gaze of the player.

As the word recognition task involved the ability to dwell on a particular area of the screen, we created a PowerPoint to replicate this task. The aim was to prepare participants for the task, and for the researcher to get some idea of a participant's ability to use eye gaze, as opposed to pointing. On the screen, were a series of colourful balloons, which could be 'popped', if eye gaze rested on it for 800ms (the same amount of time used for the word recognition test to determine if a participant was 'selecting' a particular word).

Procedure

Individuals were randomly assigned to either a 'primer' or 'non-primer' group. There were three parts to the study for all participants: an initial assessment using the psychometric tools; a baseline exposure to the reading test; and a follow-up exposure to the reading test. The initial assessments for each pupil involved the researcher familiar with the WISC-IV and BPVS administering these tests. These assessments were conducted over a period of two to three days. Usually these days were consecutive, but where this was not possible (e.g., for illness, or absence from school) all assessments were carried out within a two-week window.

Following this initial assessment, the baseline and follow-up reading tests were conducted on the same day using the EG_MNSRT procedures as described in Experiment 1. After the baseline assessment, the primer group took part in the eye-gaze

games. They played the two online games, presented in 10-min slots. However, if a participant appeared to be particularly adept at moving the caterpillar around the screen, or revealing the cartoon character, using their gaze, the game could be stopped early to prevent boredom. Participants were then asked to complete a PowerPoint activity which required them to pop the balloons by resting their eyes on the image. If a participant was unable to pop the balloons, the two online games were repeated to reinforce the concept of using eye-gaze as a cause-and-effect tool. The time in which each participant demonstrated an awareness of this concept (i.e. being able to rest their gaze on the image long enough to pop the balloon), varied from 6 min to 12 min, with an average of 8 min. To control for the time spent with the researcher for the primer group, the non-primer group spent time playing action games on an iPad for approximately 8 min. During these activities, the researcher was seated next to the participants to observe progress and engagement. All participants then repeated the word recognition test again, as in the baseline phase, but the form administered was different. So, if a participant has been administered 'form a' at baseline, they would identify words from form 'b' or 'c' post-intervention.

Results

Figure 2 shows the mean reading scores for both intervention groups at baseline and post-intervention. Contrary to the hypothesis, means at post-intervention were marginally lower baseline for both groups. A two-factor mixed-model analysis of variance (ANOVA), with test (baseline and post-intervention) as a within-subject factor, and intervention group (primer and non-primer) as between-subject factor, revealed no significant main effects or interaction, all $Fs<1$.

To test for the effect of verbal IQ the sample was split at the median verbal IQ score to create a lower verbal IQ group ($N=18$, mean verbal IQ$=53.6\pm10.1$), and a higher verbal IQ group ($N=15$, mean 93.93 ± 11.46). The top panel of Fig. 3 shows the group-mean reading scores for the primer and non-primer group at baseline and post-intervention for lower and higher verbal IQ groups. The higher verbal IQ group had higher reading score at baseline and post-intervention than the lower verbal IQ group. A three-way mixed-model ANOVA (verbal IQ x intervention x test) revealed a significant main effect of verbal IQ, $F(1,31)=9.08$, $p=.005$, $\eta^2_p=0.226$, but no other significant main effects nor interactions, all $Fs<1$.


Fig. 2 Experiment 2: Group-mean reading scores for the primer and non-primer group at baseline and post-intervention test. Error bars=standard deviations
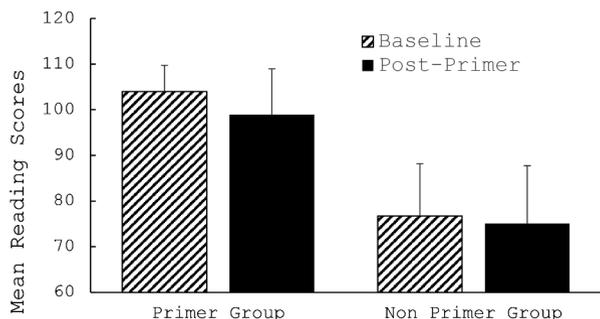
Fig. 3 Experiment 2: Top panel = Group-mean reading scores for baseline and post-intervention tests for higher verbal IQ (VIQ) and lower verbal IQ (VIQ) groups. Bottom panel = Group-mean WRS Scores for baseline and post-intervention tests test for higher and lower perceptual reasoning (PRIQ) groups exposed to the primer and non-primer manipulations. Error bars = standard deviations
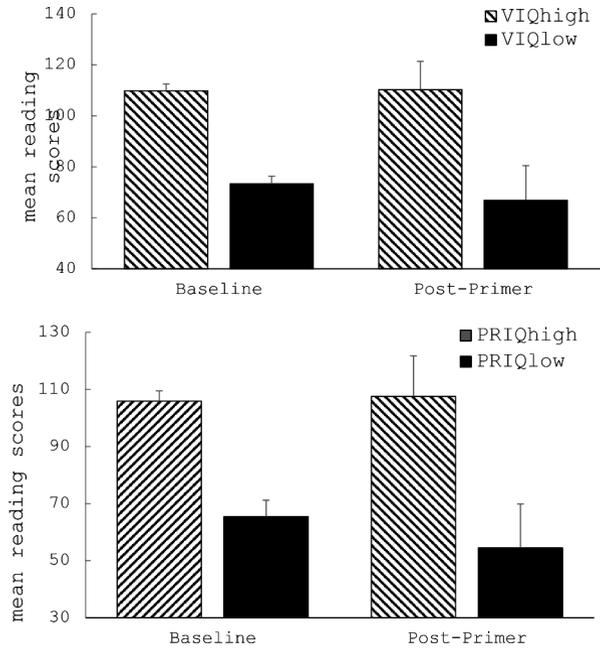
Fig. 4 Group-mean Scores for baseline and post-intervention tests for lower and higher BPVS Groups, for both intervention types. Error bars = standard deviations
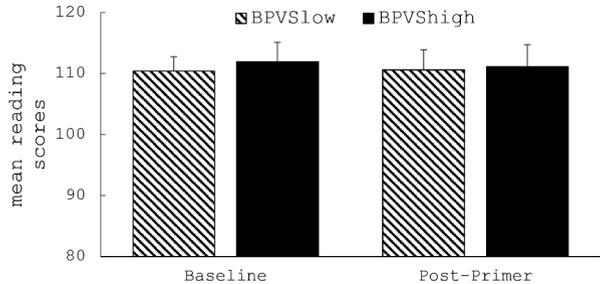
The sample also was split at the median perceptual reasoning IQ: (lower: $N = 13$, mean perceptual reasoning IQ = 53.77 ± 7.19; higher: $N = 20$, mean perceptual reasoning IQ = 91.7 ± 11.22). Figure 3 (bottom panel) shows mean reading scores for primer and non-primer groups, at baseline and post-intervention. The higher perceptual reasoning group had higher reading scores than the lower perceptual reasoning group in both tests. A three-factor mixed-model ANOVA (perceptual IQ x intervention x test) revealed a significant main effect of perceptual IQ, $F(1,31) = 13.27$, $p < .001$, $\eta^2_p = 0.300$, but no other significant main effects or interactions, all $ps > 0.10$.

The sample was split at the median to create a lower BPVS group ($N = 14$, mean BPVS = 23.28 ± 28.12), and a higher BPVS group ($N = 14$, mean BPVS = 88.5 ± 18.05). There was very little difference between WRS for these groups (Fig. 4). A three-factor mixed-model ANOVA (BPVS x intervention type x test) revealed no significant main effects or interactions, all $Fs < 1$.

The results of the current study demonstrated no significant differences between the reading performance of those who did and did not participate in eye-gaze training. Further analysis showed no impact of diagnosis, verbal IQ, perceptual reasoning IQ, or receptive vocabulary ability. Taken together, these results imply that eye-gaze may not be a particularly helpful response to use in the current version of the reading test. A limitation of the current study was the amount of time given to familiarise with the use of eye-gaze, and a more longitudinal approach may yield better results. Although participants engaged well with the games, it may be that these particular games were not reinforcing the hoped-for understanding that eye-gaze was the cause of on-screen effects. Alternatively, participants may have developed that understanding when playing the games but did not transfer to the word recognition assessment. Additionally, if the time spent looking at the screen is important, variables such as duration with gaze within the screen should also be compared. If such information is available, using the variables in the analysis may enable the study to achieve its objectives better.

General Discussion

The current series of studies evaluated the use of a modified version of the NSRT in assessing reading for children with special educational needs. The results demonstrated that removing the verbal requirement from this test produced higher reading scores (Experiment 1), and that this effect was most pronounced for those with poor verbal IQ and reading ability as measured by traditional assessments. These results replicate and extend those reported by Arnold and Reed (2019, 2024). However, the removal of the need to produce a pointing response, and the substation of this by eye-gaze monitoring did not help the participants demonstrate any latent reading abilities. Neither did provision of the delivery of a primer activity designed to familiarise participants with the use of eye-gaze tracking equipment (Experiment 2).

The current Experiment 2 was premised on the notion that a lack of familiarity with eye-gaze equipment could in part account for the results from Experiment 1. However, an intervention designed to promote this ability was not effective in improving performance on the eye-gaze reading test. Moreover, where a participant is able to use pointing as a form of response, it is easier for the assessor to know if the participant has not provided a response (as in their finger hasn't made contact with the screen), or has provided a wrong answer, by touching an area of the screen that represents an incorrect answer.

Nevertheless, if the prospects of children in special education who struggle with language are to be improved, it is important that new methods of assessment that could lead to more successful interventions are developed. This could be particularly important for younger children or those who may, as a result of motor deficits, not have developed pointing skills. As well as children who have a physical disability, motor deficits are also commonly associated with children who have autism (Paquet et al., 2016; Shetreat-Klein et al., 2014). Another comorbid issue for children on the autism spectrum can be deficits in joint attention. This means that although some children will appear to be very capable with the use of touch screens, manipulating

an iPad or tablet, for example, they won't necessarily have developed the ability to use pointing to make requests or as a form of response. For such children to achieve their true potential therefore, there has to be a desire to develop forms of assessment that can account for such barriers. This can only be achieved through a willingness to experiment with methods of observation that have the potential to be truly inclusive. Only then will educational settings be making the kind of reasonable adjustments that all non-verbal children should be entitled to.

In summary, the current study suggests that the modified touch-screen version of the test created by Arnold and Reed (2019) is the most effective format for children with additional learning needs, as long as they have the ability to utilise pointing as a form of response.

Data availability Data are available on request from the authors.

Declarations

Research Involving Human Participants and/or Animals Ethical approval was given by the Ethics Committee of the Psychology Department of the University in accordance with the Helsinki agreement.

Informed Consent All participants gave their informed consent for participation. Caregivers of the participants gave their informed consent for their children's participation. The Headteachers of the schools gave their informed consent for their pupils' participation.

Conflicts of Interest The authors report no conflicts of interest.

References

Arnold, S., & Reed, P. (2019). Measuring the word recognition abilities of children who are both verbal and nonverbal with ASD using a traditional paper -based and a novel digital test format *British Journal of Special Education*.

Arnold, S., & Reed, P. (2024). Measuring the comprehension abilities of children who are both verbal and nonverbal with ASD Extending the analysis of a novel digital test format *Good Autism Practice*.

McCarty, C., & Lallaway, M. (2012). *New Salford Reading Test*. Hodder & Stoughton.

Paquet, A., Olliac, B., Golse, B., & Vaivre-Douret, L. (2016). Current knowledge on motor disorders in children With autism spectrum disorder (ASD). *Child Neuropsychology*, *22*(*7*), 763–794.

Plesa Skwerer, D., Jordan, S. E., Brukilacchio, B. H., & Tager-Flusberg, H. (2016). Comparing methods for assessing receptive language skills in minimally verbal children and adolescents With autism spectrum disorders. *Autism*, *20*(*5*), 591–604.

Shetreat-Klein, M., Shinnar, S., & Rapin, I. (2014). Abnormalities of joint mobility and gait in children with autism spectrum disorders. *Brain and Development*, *36(2)*, 91–96.

Tager-Flusberg, H., & Kasari, C. (2013). Minimally verbal school-aged children with autism spectrum disorder The neglected end of the spectrum. *Autism Research*, *6*(*6*), 468–478.

Tager-Flusberg, H., Skwerer, P., Joseph, D., Brukilacchio, R. M., & Decker, B. J. Eggleston, B., & Yoder, A. (2017). Conducting research with minimally verbal participants with Autism spectrum disorder. *Autism*, *21*(*7*), 852–861.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children*, *4th Edn*. San Antonia, TX: PsychCorp.