Precision, Reliability and Application of the Wilkins Rate of Reading Test

James M Gilchrist [1], Peter M Allen [2], Laura Monger [2], Krithica Srinivasan [3], Arnold Wilkins [4]

1. Independent Researcher in Optometry and Vision Science, North Yorkshire, UK
2. Vision and Hearing Sciences Research Centre, Anglia Ruskin University, Cambridge, UK
3. Department of Optometry, Manipal College of Health Professions, Manipal Academy of Higher Education, Manipal, India
4. Department of Psychology, University of Essex, Colchester, UK

Running head: Wilkins Rate of Reading Test

Keywords: precision, reliability, fluency, reading rate

Commented [MR1]: Please list keywords in alphabetical order

Address for correspondence

Dr James Gilchrist
j.m.gilchrist@gmail.com

Conflict of interest

AW receives royalties on sales of the Wilkins Rate of Reading Test

Abstract

Background: The Wilkins Rate of Reading Test (WRRT) enables rapid measurement of reading speed using text passages that have no semantic content and demand minimal word recognition skills. It is suited to applications where the primary interest is in the influence of visual and ocular motor factors on reading rate.

Methods: We obtained estimates of precision and reliability of WRRT from four data samples (A-D) collected independently by the authors: A) n = 118 adults; B) n = 90 adults; C) n = 787 children; D) n = 134 children. Each participant was asked to read aloud as quickly and accurately as possible, for one minute, and results were recorded as number of words read correctly per minute (wcpm).

Results: Estimates of precision are given by the within-subjects standard deviation s_w, and reliability by the intraclass correlation coefficient for single measurements r_1. For each sample these estimates were A) s_w = 11.5 wcpm, r_1 = 0.85; B) s_w = 3.8 wcpm, r_1 = 0.98; C) s_w = 6.7 wcpm, r_1 = 0.93; D) s_w = 6.2 wcpm, r_1 = 0.94.

Conclusion: The reliability of WRRT reflects large variation in reading rate *between* individuals compared to *within*-individual variability, indicating that it is a good test for discriminating differences in reading speed between individuals. The precision of the test varies from 3.8 wcpm to 11.5 wcpm among samples, and the pooled value of 7.2 wcpm, provides a basis for setting a population-wide criterion for minimum detectable change of reading rate in individuals over time. Nevertheless, a preferable way of monitoring change in an individual would be to use a criterion determined from estimates of that individual's baseline variation in WRRT scores.

Key Points

- WRRT has been shown to be sensitive to a wide range of visual and ocular motor factors resulting from type design, display characteristics, coloured filters and visual and binocular dysfunctions.
- WRRT has good reliability for discriminating between individuals and its overall precision supports a criterion of ~22 words per minute for minimal detectable change in reading speed over time.
- WRRT shows large differences in reading speed between and within individuals, which cannot be explained in terms of text decoding and comprehension.

Introduction

The goal of reading is comprehension of text, which requires both the ability to identify words and fluency in doing so.[1,2] Word identification requires mapping of word orthography to phonology (print-to-sound decoding), recognition of whether the result constitutes a word and a decision concerning its meaning.[3,4]

Fluency involves the ability to read words quickly with natural intonation and expression (prosody), and it is regarded as a key link between word identification and comprehension.[5] Comprehension requires fast and efficient execution of the word identification process, and the ability to render the resulting stream of words with sufficient speed.[5]

Whereas single word identification may be assumed to depend primarily on orthographic-phonological decoding, prosody depends on the fluency with which a word sequence can be read. Fluency is strongly affected by certain visual and ocular motor factors.[6] The influence of these factors can be measured separately from the cognitive factors that underpin decoding. Decoding ability alone can be measured by the accuracy of single word and non-word identification, without respect to speed. Comprehension of text can be measured without regard to accuracy of specific word identification or reading rate, and fluency can be measured as the rate of reading sequential text. Ideally, fluency is measured when the decoding and comprehension demands are minimised because decoding and comprehension abilities themselves may help or hinder reading rate.[3]

Wilkins Rate of Reading Test

The Wilkins Rate of Reading Test (WRRT) was introduced to provide a test that could be used to evaluate the effects on reading speed of visual factors and interventions (notably lenses and/or filters), especially in children with reading difficulties.[7] The design of the test 'minimises the linguistic and semantic aspects of reading and maximises the visual difficulties', noting that 'many visual difficulties with reading seem to emerge when the test is presented in a long paragraph with closely spaced lines and letters.' [7] The WRRT is not a test of fluency, at least in so far as fluency includes normal prosody, because prosody is altered when words are disconnected and the text meaningless, as in this test.

INSERT FIGURE 1 ABOUT HERE

The text of the WRRT (*Figure 1*) is reproduced in a small (9 point) self-similar font (Times New Roman) with a small (4 point) space between words. The text is set as a paragraph of 10 lines 72.5 mm wide, 33.4 mm high, with an interline space of 3.15 mm. The letters have an x-height of 1.6 mm and a width that averages 1.53 mm. The text consists of 15 high-frequency words, the same 15 words on each of the 10 lines, but in a different random order. Although one word in the passage may cue another neighbouring word with which it is commonly associated (e.g., cat-dog), this association is random and will be similar overall from one version of the test to another. The test can be tackled both by adults and by children who have only a modest reading vocabulary.

Concepts of precision and reliability

Measures of human performance or ability can serve two purposes: i) to enable monitoring of changes in ability *within* individuals over time, and ii) to reveal differences in ability *between* individuals. In relation to reading ability, measurement of change within-individuals is of particular importance for monitoring the development of reading ability in children, and the deterioration of reading ability in older adults experiencing loss of function such as visual or cognitive impairment. In both cases, repeated measurements over a period of time serve not only to reveal the pattern of change but also to demonstrate whether interventions are of benefit in helping to improve development of an individual's reading ability or slow its decline. On the other hand, measuring differences between individuals enables identification (diagnosis) of those whose reading ability is substantially lower than their peers, and provides evidence to support the introduction of interventions aimed at improvement. The statistical requirements of an effective test for the two very different purposes just described are, respectively, precision and reliability.

Precision is the general term for random variation between repeated measurements from the same individual, which occurs inevitably as all measurements incorporate some degree of error. Estimates of precision take account only of the variation *within*-subjects, not that *between*-subjects. In order to be effective for monitoring change within-individuals over time, a measure should be precise; that is, the test-retest variation due to random measurement error should be low. The terms precision, repeatability and reproducibility all relate to within-subject variation, that is, the consistency of repeated measurements.[8] Reliability, on the other hand, takes account of variation both within- and between-subjects. Reliability is the degree to which variation between-subjects exceeds that within-subjects.[9] When between-subjects variation is large compared to that within-subjects, then test reliability is high and scores from different individuals are likely to indicate real difference between them rather than the effects of measurement error. Conversely when between-subjects variation does not greatly exceed that within-subjects, then test reliability is low and differences in scores between individuals may reflect the effects of measurement error rather than true differences. Reliability may be thought of as a measure of *discriminability*, as it emphasises the principle of being able to use the test to discriminate reliably between different individuals.

Recognition that precision is a measure of variability within-individuals, while reliability involves both within- and between-individual variability, means that we can expect two different scenarios in practice. The first relates to tests that are intended to be used in a single population, in which the degree of variation between individuals is assumed to be fairly constant. In this case, the better the precision of a test, the better will be its reliability; in other words, reliability will be determined by precision. The second scenario relates to tests employed in a number of different populations, each with its own degree of variation between individuals. In this case, it is quite possible for a test having good precision to have good reliability in one population and poor reliability in another. This highlights the importance of making a clear distinction between the concepts of precision and reliability, and the need for test evaluation to be undertaken in different populations as appropriate.

Although the Wilkins Rate of Reading Test (WRRT) has now been in existence for 25 years, there are only limited data on its precision and reliability. Also, beyond its use by optometrists and others for assessments related to visual stress,[10] the test remains largely unknown and its potential for more general use as a measure of reading ability unrecognised outside this area of application. Our aim here is to address these issues. We first present data that enable estimates of the precision and reliability of WRRT in schoolchildren and young adults, and later we consider its application in a variety of contexts.

Methods

Participants and data collection

Participants were recruited and assessed by the authors in their respective locations, giving four separate samples.

Adult participants in Sample A recruited by JG from the undergraduate student population attending the University of Bradford, Bradford, UK and were tested by one of two student assistants who had been trained in the use of the WRRT but were not involved in study design, data analysis or authoring. The sample consisted of 68 women and 52 men aged 18-40 years (mean 21.8 years). Participants in Sample B were also undergraduate students recruited by LM and PA from Anglia Ruskin University, Cambridge, UK: 63 women and 37 men aged 17-31 years (mean 21.4 years). Sample C were children recruited from nine different schools located in the Udupi Taluk region of India and were tested by KS. The sample consisted of 431 boys and 368 girls aged 7-16 years (mean 11.7 years). The children in Sample D were recruited by AW from a school in Norwich, UK and were obtained as part of a study that has been published previously.[11] There were 82 girls and 57 boys aged 9-12 years (mean 10.5 years). Recruitment and participation of subjects was achieved in compliance with relevant local/institutional requirements for ethical approval. In the case of the children in Samples C and D, parental consent for participation was obtained.

All participants provided two measures of reading rate with the WRRT. These test and retest measures provide all the data to be presented in the analysis that follows below. In each of Samples B, C and D, the collection of WRRT measurements was undertaken as part of investigations of the effects of coloured overlays on reading rate. For these samples, therefore, the WRRT data to be presented here are those taken without the use of any coloured overlay but obtained within a testing sequence that interleaved WRRT measurements with and without coloured overlays. In the case of Samples B and D, the WRRT test sequence in relation to use of coloured overlays was *with-without-without-with*, while for Sample C the WRRT test sequence involving overlays was allocated to subjects randomly as either *with-without-without-with* or *without-with-with-without*. The tests were presented in immediate succession, typically less than five minutes apart. For Sample A there was no use of coloured overlays at any stage of the data collection.

Participants in all samples, including the children in Sample C whose native language was not English, demonstrated that they were able to read (recognise and pronounce) the 15 words included in the WRRT prior to testing. Participants in Sample D read the entire passage, while those in the three other samples read for one minute. The passage length (150 words) is such that the time difference between reading the whole passage and reading for one minute is typically small.

The test was scored by noting the errors on a score sheet comprising an enlarged version of the text, and by measuring the total time taken to read the passage. From these measurements reading rates were calculated as words correct per minute (wcpm). Tests were not audio-recorded. Each application of the test in a clinic/practice setting takes approximately one minute and the test is easy to score in situ. Data collection replicated, as far as possible, what would typically be done in practice, the aim being to obtain estimates of precision and reliability that would reasonably apply in a typical practice setting.

All participants who would normally use a refractive correction for their academic or schoolwork were corrected for this study, otherwise no refractive correction was given. The test conditions for all samples were controlled by fixing the viewing distance at ~40 cm, with lighting conditions adjusted to give a glare-free illuminance level on the task of 500-1000 lux, resulting in task background luminance between 70 and 100 cd/m^2. For Samples A and B the lighting was a tungsten-halogen desk lamp adjusted to give the luminance described above, with ambient room lighting provided by a ceiling-mounted 'warm white' fluorescent lamp. For Sample C, natural daylight was available, while for Sample D the lighting was fluorescent with magnetic ballast. For all samples, selection of which of the four standard WRRT passages to use was randomised and each participant was presented with a different passage on test and retest (see *Figure 1*).

Statistical Approach

As described in the introduction above, the principal aim of data analysis was to obtain estimates of WRRT precision and reliability. The statistical principles are set out in a number of articles by Bland and Altman[12-15] and elaborated in greater detail in several texts.[9,16,17] Analysis was carried out using the open statistical application jamovi (jamovi.org).

Precision

The underlying assumption, consistent with classical test theory,[9] is that each subject has a *true* reading rate, which is estimated by each individual measurement. The best estimate of true reading rate will be the average of a number of repeated measurements and, assuming that the true rate remains constant (at least over the relatively short time periods that apply here), then repeated measurements of the same subject may be assumed to vary randomly around the true value due to measurement error. Thus, the standard deviation of repeated measurements on any individual subject will provide an estimate of measurement error. If such an estimate is obtained from a number of participants, then its value will vary, so evaluation of measurement error involves calculating the common within-subjects standard deviation in the sample,[13] which we denote s_w. This value is the estimated precision of the test, also called the standard error of measurement.[16] Note that this means that the precision of a test is expressed in the units of measurement of the test. Small values of s_w indicate small amounts of measurement error, which corresponds to good precision. An important caveat for the use of s_w as an overall estimate of measurement error in a sample of subjects is that there should be no evidence that the size of s_w is systematically related to the level of performance on the test. This can be checked by plotting the standard deviation of repeated measurements (or their absolute difference, in the case of test and retest measures) against their mean.[15]

Reliability

The correlation between test and retest scores is a measure of the reliability of the test. This is discussed by Bland and Altman[14] who recognise its interpretation not as a measure of the amount of measurement error but of the ability of a measure to discriminate between individuals: ’The correlation coefficient (between one test and the next across participants) can be used to compare measurements of different quantities. The measures with the highest correlation between repeated measurements would discriminate best between individuals.’ Bland and Altman[14] point out that the correct approach is to use the intraclass correlation coefficient, which estimates the degree to which the variation in measurements between-subjects exceeds that within-subjects. Note that the reliability of a test is expressed by a value between 0 and 1 with no units of measurement. This is consistent with the interpretation of reliability as a correlation between test and retest measurements.

Analysis

Analysis of each data sample comprised four stages:

1) Data were trimmed to remove extreme values from distributions between-subjects (test-retest means) and within-subjects (test-retest differences), in order that estimates of WRRT precision and reliability should not be influenced by atypical data. Trimming was carried out using a robust z-score procedure based on the median absolute deviation.[18] Values with an absolute z-score greater than 3.5 were eliminated from further analysis. The value with highest z-score was eliminated, the statistics were re-calculated, and this procedure repeated until no more values met the criterion for elimination. Although we consider that removal of extreme values is justified for this analysis, some may argue that even extreme values are representative of their populations and should be retained. For this reason, although we focus our analysis on trimmed data, we also include statistics of the untrimmed data for comparison.

2) Calculation of descriptive statistics for between-subjects and within-subjects distributions: mean, standard deviation, median, range and assessment of normality.

3) Assessment of the association between test-retest variation and mean reading rate.[15]

4) Assessment of systematic test-retest bias that may be due to learning or fatigue effects, together with estimation of precision and reliability.

The last stage, assessment of test-retest bias and estimation of precision and reliability, involved repeated-measures analysis of variance (ANOVA), in which the repeated measures were test and retest reading rates for each participant. This approach is widely endorsed, [16,17,19,20] and full details of the ANOVA and necessary computations are given by Winer.[17] Its foundation is the principle that the total variance in the data can be partitioned into between-subjects and within-subjects components, which ANOVA expresses as measures of mean-squared variation (*Table 1*).

Table 1. Partitioning of variance in repeated-measures ANOVA (see Winer [17])

Source of Variation	**degrees of freedom (*df*)**	**Mean Squares (*MS*)**	***F***
Between Subjects Effects			
Residual	$n-1$	MS_R	
Within Subjects Effects			
Between Measurements	$k-1$	MS_C	MS_C / MS_E
Residual	$(n-1)(k-1)$	MS_E	

In *Table 1*, the expressions for degrees of freedom (*df*) are for samples with *n* subjects (participants) and *k* repeated measurements per subject. In test-retest studies then $k = 2$. The denotation of mean squares values (*MS*) follows the example of Koo and Li.[20] Note that the partitioning of data variance shown here allows analysis to be conducted on the basis of both one-way and two-way ANOVA models.[16,17,19]

The choice of ANOVA model must be appropriate for the study being undertaken. Koo and Li [20] discuss this in detail and for studies involving test-retest reliability (as is the case here) they recommend use of a two-way, mixed-effects, absolute agreement model, designated Model 2 by Shrout and Fleiss.[19] Under this model, the reliability estimate takes account of both random error (precision or consistency) and systematic error (bias or agreement) between test and retest measurements. Here, however, we use Shrout and Fleiss Model 3, described by Koo and Li [20] as ’two-way, mixed-effects, consistency’, which ignores systematic bias, because our intention is to include Bland-Altman analysis of test-retest repeatability, which assesses agreement and consistency separately,[12] and the only intraclass correlation model fully consistent with this analysis is that designated by Shrout and Fleiss[19] as intraclass correlation coefficient (ICC) (3,1). Under this model, precision (standard error of measurement) is estimated by the square root of the within-subjects residual mean square value, i.e., $s_w = \sqrt{MS_E}$, and the reliability of a single measurement is estimated by intraclass correlation coefficient ICC (3,1) which is calculated thus: $r_1 = (MS_R - MS_E)/(MS_R + (k-1)MS_E)$.[19,20] The F value in *Table 1*, given by MS_C/MS_E is used to assess bias, as it evaluates the mean difference between test and retest measurements, giving a result equivalent to the paired t-test ($F = t^2$).

Results

Data trimming and descriptive statistics

A number of atypically high or low reading rates, and atypically high test-retest differences were identified and eliminated from each sample using the trimming procedure described previously.[18]

The percentages of the data removed from Sample A to D were, respectively: 1.7%, 10.0%, 1.5% and 3.6%. All the following results are for trimmed data samples unless indicated otherwise.

Figure 2 shows distributions of mean reading rates, and descriptive statistics are in *Table 2*.

INSERT FIGURE 2 ABOUT HERE

Table 2. Descriptive statistics of test-retest means

Sample	Trimmed n	Mean $\bar{x}$	SD s_x	Min - Max (Range)	Median	Skewness	Kurtosis
A	118	183.5	28.3	114 - 257 (143)	187	-0.036	0.167
B	90	159.2	23.8	104 - 224 (120)	156	0.144	-0.172
C	787	111.1	25.4	40 - 194 (154)	110	0.361	0.125
D	134	94.5	25.7	32 - 168 (136)	95.8	0.138	0.212

Mean reading rates differed significantly among the four samples in the study: $F(3,1125) = 387.8, p < 0.001, \eta_\mathrm{p}^2 = 0.508$ and, as expected, mean reading rates for adults (Samples A and B) were higher than those for children (Samples C and D): $F(1,1127) = 989.3, p < 0.001, \eta_\mathrm{p}^2 = 0.467$.

Descriptive statistics of the distributions of test-retest differences are given in *Table 3*. As will be shown later, test-retest statistics in this form may be used to assess the agreement (bias) and repeatability (precision) in repeated measurements.[12] The mean difference $\bar{d}$ is a measure of agreement, and the standard deviation of differences s_d is a measure of repeatability. The 95% limits of repeatability given in *Table 3* are calculated from the standard deviation of differences thus: $\bar{d} \pm 1.96 \times s_d$.

Table 3. Descriptive statistics of test-retest differences

Sample	Trimmed n	Mean $\bar{d}$	SD s_d	Lower 95% Limit	Upper 95% Limit	Skewness	Kurtosis
A	118	-1.58	16.26	-33.45	30.30	-0.150	0.250
B	90	-0.06	5.38	-10.60	10.50	-0.419	0.061
C	787	1.00	9.53	-17.67	19.68	-0.076	0.470
D	134	1.66	8.78	-15.55	18.88	0.054	-0.070

Test-retest differences and mean reading rates

Statistics in *Table 4* confirm the lack of any systematic association (proportionality) between absolute test-retest differences and means. [15]

Table 4. Association of absolute test-retest difference and mean

Sample	Trimmed *n*	Kendall's Tau	Kendall's p	Regression Slope	Regression Slope SE	Regression R^2
A	118	0.007	0.907	0.013	0.034	0.001
B	90	-0.004	0.961	-0.006	0.014	0.002
C	787	0.057	0.020	0.023	0.009	0.009
D	134	0.025	0.678	0.011	0.018	0.003

Bias, precision and reliability

Figure 3 shows (Bland-Altman) plots of test-retest differences against mean reading rates.[12]

INSERT FIGURE 3 ABOUT HERE

Solid horizontal lines in *Figure 3* indicate mean difference $\overline{d}$ (test-retest bias) and dashed lines indicate the 95% limits of repeatability as shown in *Table 3*. The standard deviation of test-retest differences is a measure of repeatability (precision) and may be used to determine the within-subjects standard deviation (standard error of measurement) thus: $s_w = s_d/\sqrt{2}$.

Estimates of test-retest bias $\overline{d}$ and precision (expressed in terms of the standard deviation of test-retest differences) s_d for each sample are presented in *Figure 3* and *Table 3*, in the context of Bland-Altman analysis.[12] These, along with estimates of reliability may also be obtained by two-way, repeated-measures ANOVA (Shrout & Fleiss, Model 3) described previously. *Table 5* shows evaluation of test-retest bias for each sample, and *Table 6* provides a summary of precision and reliability estimates and their 95% confidence intervals.[19]

Table 5. Evaluation of test-retest bias

Sample	***F***	***p***	η_p^2
A	1.109	0.295	0.009
B	0.010	0.922	<0.001

C	8.735	0.003	0.011
D	4.809	0.030	0.035

Table 6. Precision and reliability estimates, and confidence limits

	Precision				Reliability		
Sample	Bland-Altman $s_d\ (= s_w\sqrt{2})$	$s_w \approx \hat{\sigma}_w$	$\hat{\sigma}_w$ lower 95% CL	$\hat{\sigma}_w$ upper 95% CL	$r_1 \approx \hat{\rho}_1$ ICC (3,1)	$\hat{\rho}_1$ lower 95% CL	$\hat{\rho}_1$ upper 95% CL
A	16.26	11.50	10.20	13.19	0.848	0.695	0.927
B	5.38	3.80	3.32	4.46	0.975	0.952	0.987
C	9.53	6.74	6.42	7.09	0.932	0.681	0.987
D	8.78	6.21	5.55	7.06	0.943	0.873	0.975

Finally, for completeness, *Tables 7 and 8* give descriptive statistics and estimates of precision and reliability obtained from the original untrimmed data samples.

Table 7. Descriptive statistics (untrimmed data)

		Between-Subjects (Test-Retest Means)				Within-Subjects (Test-Retest Differences)			
Sample	Untrimmed n	Mean $\bar{x}$	SD s_x	Min-Max (Range)	Median	Mean $\bar{d}$	SD s_d	Lower 95% Limit	Upper 95% Limit
A	120	183.5	30.2	95 - 267 (172)	187	-1.54	16.15	-33.19	30.11
B	100	160.7	29.3	73 - 283 (210)	156	-0.47	7.67	-15.50	14.56
C	799	111.4	26.4	40 - 240 (200)	110	0.62	11.19	-21.31	22.55
D	139	95.4	25.9	32 - 168 (136)	96.5	1.45	10.41	-18.95	21.86

Table 8. Precision and reliability estimates, and confidence limits (untrimmed data)

	Precision				Reliability		
Sample	Bland-Altman $s_d\ (= s_w\sqrt{2})$	$s_w \approx \hat{\sigma}_w$	$\hat{\sigma}_w$ lower 95% CL	$\hat{\sigma}_w$ upper 95% CL	$r_1 \approx \hat{\rho}_1$ ICC (3,1)	$\hat{\rho}_1$ lower 95% CL	$\hat{\rho}_1$ upper 95% CL
A	16.15	11.42	10.13	13.08	0.867	0.728	0.937
B	7.67	5.42	4.76	6.30	0.966	0.933	0.983

C	11.19	7.91	7.54	8.32	0.914	0.609	0.984
D	10.41	7.36	6.59	8.35	0.923	0.826	0.967

Discussion

WRRT reading rates and variation

The distributions of mean test and retest reading rates (*Figure 2*) and their descriptive statistics (*Table 2*) show that, as expected, adults (Samples A and B) read more quickly on average than children (Samples C and D). The standard deviations of the four samples are similar and their values show a general finding of large variation in reading rates between-subjects, with an overall average range of 138 wcpm between the slowest and fastest readers, and considerable overlap in the reading rates of adults and children. The trimming of our data to remove extreme values means that the reading rate distributions reported here will be representative of the large majority of individuals of similar age. Some, but not many young adults read the WRRT at rates of less than 100-120 wcpm, most read from 150 to 200 wcpm (approximately the inter-quartile range) and some read more quickly. Similarly, most children towards the end of primary school (age 10-11 years) read around 90 to 125 wcpm, but some children manage only 60 to 80 wcpm or less. Such children, the 'slow readers', are those who may have some specific difficulty, visual or otherwise, and in whom some form of intervention may be therefore beneficial to aid development of their reading rate and fluency.

Bias, precision and reliability of the WRRT

As discussed previously, estimation of test precision using the overall within-subject standard deviation s_w assumes there is no association between participants' variability and their mean scores. In other words, the random (measurement) error in the test should not depend upon whether a subject is a slow or a fast reader. *Table 4* shows statistics of the association between absolute test-retest differences and mean reading rates for participants in each of the four samples in the study.[15] In general these show very low correlations and shallow regression slopes, confirming that there is no evidence that the magnitude of test-retest difference depends upon reading rate. Therefore, we conclude that using within-subjects standard deviation s_w as an estimate of overall test precision in each sample is justified. The implications of this will be discussed later.

Table 5 shows the evaluation of test-retest bias obtained from ANOVA using the single-measurement, two-way, repeated-measures, mixed-effects, consistency model.[19,20] Note that the *F* statistics and corresponding *p*-values assess bias by comparing test and retest measurements. This is equivalent to using the paired t-test. For Samples A and B we note that *p*-values are relatively large, confirming no systematic bias between test and retest measures, whereas $p < 0.05$ for Samples C and D suggesting there is evidence of test-retest bias that may be the result of practice or fatigue effects. However, in all samples the effect size, indicated by partial eta-squared (η_p^2), is small so we conclude that the evidence for bias in all samples is weak and that measurements with the WRRT appear overall not to be susceptible to learning or fatigue effects (see also Bland-Altman plots, *Figure 3*).

Table 6 gives estimates of WRRT precision, that is the within-subjects standard deviation (or standard error of measurement) s_w and 95% confidence limits. Precision may be calculated in two ways that give equivalent results. First, from the standard deviation of test-retest differences s_d, we calculate the standard error in the conventional way by dividing the standard deviation by the square-root of the number of repeated measurements per subject. Since the number of measurements $k = 2$, therefore $s_w = s_d/\sqrt{2}$. The second approach to calculation of s_w is directly from the two-way ANOVA mean squares, as described previously.

The variation in estimated WRRT precision between samples is worthy of discussion. There is a remarkable similarity in the $s_w$values for the two samples of children (C and D), given the marked differences in their locations and characteristics. On the other hand, adult samples (A and B) seem distinctly different from one another, with Sample A showing much poorer repeatability. This difference is seen very clearly in *Figure 3*. The most likely explanations for the variation in Sample A are that: i) the two different, inexperienced student assistants responsible for data collection may have introduced variance in how they instructed participants and/or recorded measurements - although in our experience the latter seems less likely, and/or ii) participants themselves may have been less assiduous in following instructions on how to approach reading of the passages. In general, this interpretation would indicate that WRRT is (unsurprisingly) susceptible to the effects of instruction and the reader's response to instruction. Whatever the reason, this difference in repeatability given by Samples A and B - which were otherwise comparable in being composed of young adults of similar age and background - has implications for the use of sample precision estimates to set general criteria for change, which we discuss next.

An essential use of WRRT precision estimates is to determine the limits within which repeated measurements are expected to vary due to random measurement error alone, so that a criterion can be set for the required effect of any intervention that purports to bring about a true change in reading rate - that is, the minimum detectable change. To apply this principle to Sample A, for example, we take the estimated $s_w = 11.50$ wcpm as the basis of a criterion for change and then, depending upon how strict the criterion needs to be, set a multiple of s_w as the 'threshold' value that must be exceeded. A well-known approach[12] is to set this criterion based on the standard deviation of test-retest differences, typically using $1.96 \times s_d$ or (for ease of calculation) $2 \times s_d$. When this is expressed in terms of s_w then the corresponding values are $2.77 \times s_w$ and $2.83 \times s_w$. If, once again, we approximate for ease of calculation, then a pragmatic criterion for minimum detectable change would be $3 \times s_w$ (using this approximation with the precision estimates here increases the criterion value by only 1 or 2 wcpm). On this basis, *Table 9* shows precision estimates and criteria for minimum detectable change for each sample in the study. In addition, if a more demanding criterion is required (i.e., a higher threshold), then we might use the same principle applied to the upper confidence limit of estimated precision (last two columns of *Table 9*).

Table 9. Estimates of precision and possible criteria for minimum detectable change

Sample	Precision $s_w \approx \hat{\sigma}_w$	Criterion for Minimum Detectable Change $3 \times s_w$	Upper 95% CL of Precision $\hat{\sigma}_{w.UL}$	Strict Criterion for Minimum Detectable Change $3 \times \hat{\sigma}_{w.UL}$
A	11.50	~ 35 wcpm	13.19	~ 40 wcpm
B	3.80	~ 11 wcpm	4.46	~ 13 wcpm
C	6.74	~ 20 wcpm	7.09	~ 21 wcpm
D	6.21	~ 19 wcpm	7.06	~ 21 wcpm

An important issue in the use of minimum change criteria is the scope of application. We have seen that estimates of WRRT precision in this study range from $s_w =$ 3.8 to 11.5 wcpm across four samples of participants, and thus *Table 9* shows how we might set different criteria for change in different populations according to their respective sample estimates of test precision. In practice, however, this might be considered unwieldy and practitioners may seek a more global approach under which the same change criterion can be used for all, or at least that there should be as few criteria as possible; for example, one criterion for children and another for adults. Resolution of this question will require further research, but here we recognise two extreme possibilities: the first involving adoption of a single 'universal' change criterion for all populations and applications of the WRRT, and the second requiring that the criterion should be set for every individual based on that person's unique variability on repeated measurement.

If a single, universal criterion for change is desirable then our data can contribute to the debate on what this value should be. The estimated WRRT precision obtained by combining data from all samples is $s_w = 7.2$, with population estimate (95% CI) of 6.9 <= $\hat{\sigma}_w$ <= 7.5. Based on this, and the principle represented in *Table 9*, then a single criterion of ~22 wcpm could apply as a general rule. Although this approach is appealing, and our data provide a basis for setting a suitable value, inspection of *Figure 3* shows that it is certainly not optimal. Here we see that not only does s_w vary between samples, but also that test-retest differences vary to a much greater extent between individuals within samples. The extent of this variation is such that, in each sample, there are individuals who exhibit test-retest variation much lower than the overall sample s_w, while other individuals exhibit much larger variation. For this reason, we believe that a strong case can be made, in every situation where monitoring of individual change is of concern, that the criterion for change should be based on an estimate of the baseline variation of that individual, rather than on a single, generalised population estimate.

Lastly, on the matter of criteria for change, there is a question of whether these should be expressed as a number of words correct per minute (wcpm) as in the examples discussed above, or as percentage change from the initial, baseline reading rate. Current practice is generally to express the change criterion in percentage terms, and a criterion of 15% has been proposed previously.[21] A caveat in the use of percentage change, however, is that the same change criterion of 22 wcpm applied to the adult samples (combined mean reading rate 171 wcpm) represents a change of ~13%, whereas when applied to our samples of children it represents a change of 21%. However, use of a criterion expressed as wcpm, rather than as percentage change, is justified by our finding - in all four samples - that the magnitude of test-retest variation is independent of mean rate (*Table 4*). In general, therefore, we favour expressing the criterion for change in words correct per minute rather than as a percentage of the initial value. This is not to say that a change in reading rate cannot usefully be expressed as percentage change, only that the criterion for minimum detectable change should be expressed and applied in wcpm.

Table 6 gives estimates of WRRT reliability, that is the intraclass correlation coefficient (ICC) for single measurements r_1 - Shrout and Fleiss ICC (3,1) - and 95% confidence limits.[19] Reliability coefficients evaluate the degree to which variation between-subjects exceeds that within-subjects, and thereby indicate the ability of a test to discriminate individual differences. Using the guidance on interpretation proposed by Koo and Li, [20] taking account of the confidence limits of the ICC estimate, our results show that the WRRT has moderate to excellent reliability in the populations sampled.

The relatively high reliability of the WRRT reflects large variation in rate of reading between individuals, which is striking in all the samples we examined (*Figure 2* and *Table 2*). Some children read faster than the average adult and, conversely, some adults read slower than the average child. It has previously been noted that in children who have similar scholastic attainment in reading, the variation in reading rate from slowest to fastest is more than a factor of 3-times, both in 7 year-olds[11] and 13 year-olds.[22] In the present study the variation in children's reading rate is even larger; whereas the highest reading rate in adults (Samples A and B) is ~2-times greater than the lowest, in children (Samples C and D) those with the highest reading rate were ~5-times faster than those with the lowest. Investigation of the source of this extreme variation may provide clues as to some of the causes of reading difficulty.

To conclude this section, we note that comparison of the results of analysis of trimmed and untrimmed data reveals only small differences in summary statistics and parameter estimates across the four samples (see *Tables 2, 3* and 6 versus *Tables 7* and *8*). Mean reading rates, between- and within-subjects, differ in trimmed and untrimmed samples by less than 1 wcpm on average, and corresponding standard deviations differ by only 1 to 2 wcpm. Estimates of precision in trimmed and untrimmed samples differ by less than 1 wcpm on average, while estimates of reliability differ by less than 0.01 wcpm.

Applications of the WRRT

There are many tests of reading ability, some aimed at children who are acquiring reading skills and others aimed at skilled adult readers whose reading ability is somehow impaired, perhaps by dyslexia, loss of vision or cognitive decline. Some of these tests use passages of meaningful text to assess comprehension, word identification accuracy (decoding ability) and/or reading rate (fluency),[22,23] while others use isolated words and sometimes non-words to assess decoding accuracy or efficiency (i.e., rate).[24] Unlike the WRRT, however, none of these tests attempts to separate assessment of reading rate/fluency from assessment of decoding ability and/or comprehension, and the consequence is that the influences of cognitive and language skills that underpin decoding and comprehension are confounded with the influences of visual and ocular motor skills and speed of processing/naming, which are important in determining rate of reading.

Here we have shown that the statistical properties of the WRRT support its use for monitoring reading rate change within individuals over time, and also for assessing differences in reading rate between individuals. The WRRT can be used with people of any age, including young children having limited word knowledge and vocabulary and, by minimising or eliminating the influences of decoding ability and comprehension, the WRRT provides a measure of reading rate that should be more sensitive than other tests to the influences of visual and ocular motor factors.

Although the WRRT has been most widely used to assess the effects of a particular form of visual intervention (coloured overlays) on reading rates in children,[11,25,26] it has also been used in other contexts in which primary interest is the effect on reading of visual and/or ocular motor factors. For example, in previous studies, reading rate on the WRRT has been shown to be affected by aspects of typography such as the spatial periodicity of text, font size (x-height) and font design in reading schemes for children.[27,28] Other researchers have used WRRT to assess reading rates and interventions in cases of visual asthenopia,[29] age-related macular degeneration,[30,31] dry eye disease[32] and binocular vision anomalies,[33] and to assess whether individuals using 3D displays may be susceptible to visual fatigue due to the demands of such displays on binocular visual and ocular motor functions.[34] Given the simple principles of design it is easy to create in languages other than English, and versions have been developed in a variety of European and Asian languages.

Finally, and more generally, we note that the WRRT is in effect a test of rapid automatised naming (RAN), in which the stimuli happen to be automatised words rather than the letters, digits, etc., that are common in some applications of RAN. Although it has not previously been presented as a RAN test, considering the WRRT from this perspective greatly broadens its potential scope of application, as it is widely acknowledged that RAN performance is an important predictor of reading attainment.[35]

References

1. Verhoeven, J., & Perfetti, C. (2008). Advances in text comprehension: Model, process and development. *Applied Cognitive Psychology, 22* (3) 293-301. doi: 10.1002/acp.1417
2. Garcia, J. R., & Cain, K. (2014). Decoding and reading comprehension: A meta-analysis to identify which reader and assessment characteristics influence the strength of the relationship in English. *Review of Educational Research,* 84 (1), 74–111. doi: 10.3102/0034654313499616
3. Hess, A. M. (1982). An analysis of the cognitive processes underlying problems in reading comprehension. *Journal of Reading Behavior*, 14 (3), 313–333. doi: 10.1080/10862968209547458
4. Kendeou, P., van den Broek, P., Helder, A. & Karlsson, J. (2014) A cognitive view of reading comprehension: implications for reading difficulties*. Learning Disabilities Research and Practice* 29 (1), 10-16. doi: 10.1111/ldrp.12025
5. Bashir, A.S., & Hook, P.E. (2009). Fluency: A key link between word identification and comprehension. *Language, Speech, and Hearing Services in Schools*, 40 (2), 196–200. doi:10.1044/0161-1461(2008/08-0074).
6. Reichle, E.D., Rayner, K., & Pollatsek, A. (2003). The E-Z Reader model of eye movement control in reading: Comparisons to other models. *Behavorial and Brain Sciences*, 26 (4), 445-76; discussion 477-526. doi: 10.1017/s0140525x03000104
7. Wilkins, A.J., Jeanes, R.J., Pumfrey, P.D., & Laskier, M. (1996). Rate of Reading Test: its reliability, and its validity in the assessment of the effects of coloured overlays. *Ophthalmic and Physiological Optics*, 16 (6), 491– 497. doi: 10.1046/j.1475-1313.1996.96000282.x
8. ISO 5725-2:2019 Accuracy (trueness and precision) of measurement methods and results — Part 2: Basic method for the determination of repeatability and reproducibility of a standard measurement method. https://www.iso.org/standard/69419.html.
9. Allen, M.J., & Yen, W.M. (1979). Introduction to measurement theory. Monterey (CA): Brooks/Cole. ISBN: 978-1-57766-230-3
10. Wilkins, A.J. (1995). *Visual stress*. Oxford: Oxford University Press. doi: 10.1093/acprof:oso/9780198521747.001.0001

11. Scott, L., McWhinnie, H., Taylor, L., Stevenson, N., Irons, P., Lewis, L., Evans, M., Evans, B., & Wilkins, A. J. (2002). Coloured overlays in schools; orthoptic and optometric findings. *Ophthalmic and Physiological Optics*, 22 (2), 156– 165. doi: 10.1046/j.1475-1313.2002.00009.x
12. Bland, J.M., & Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327 (8476), 307-310.
13. Bland, J.M., & Altman, D.G. (1996). Measurement error. *BMJ*, 312 (7047), 1654. doi: 10.1136/bmj.312.7047.1654
14. Bland, J.M., & Altman, D.G. (1996). Measurement error and correlation coefficients. *BMJ*, 313 (7048), 41-2. doi: 10.1136/bmj.313.7048.41
15. Bland, J.M., & Altman, D.G. (1996). Measurement error proportional to the mean. *BMJ*, 313 (7049), 106. doi: 10.1136/bmj.313.7049.106
16. Dunn, G. (1989). Design and Analysis of Reliability Studies: The Statistical Evaluation of Measurement Errors. Oxford University Press, Edward Arnold, London.
17. Winer, B.J. (1971) Statistical Principles in Experimental Design. McGraw-Hill, Kogakusha, Tokyo.
18. Iglewicz B., Hoaglin D.C. (1993). How to Detect and Handle Outliers. *ASQC Basic References in Quality Control, Vol.16.* ASQC Quality Press, Milwaukee, WI. ISBN 0-87389-247-X

Commented [MC2]: Needs completing?

19. Shrout, P.E., Fleiss, J.L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychol Bull*, 86 (2), 420–428. doi: 10.1037/0033-2909.86.2.420
20. Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2), 155–163. doi: 10.1016/j.jcm.2016.02.012
21. Wilkins, A.J., Allen, P.M., Monger, L.J., Gilchrist, J.M. (2016). Visual stress and dyslexia for the practising optometrist. *Optometry in Practice*, 17 (2), 103-112.
22. Neale, M. D. (1999). *Neale Analysis of Reading Ability (NARA)* 3rd ed. Melbourne: ACER Press, Australian Council for Educational Research.
23. Snowling, M. J., Stothard, S. E., Clarke, P., Bowyer-Crane, C., Harrington, A., Truelove, E., & Hulme, C. (2009). *York Assessment of Reading for Comprehension (YARC)*: London, GL Assessment.
24. Torgeson, J. K., Wagner, R. K., & Rashotte, C. A. (1999). *Test of Word Reading Efficiency (TOWRE)*. Pro-ed.
25. Jeanes, R., Busby, A., Martin, J., Lewis, E., Stevenson, N., Pointon, D. & Wilkins, A. (1997). Prolonged use of coloured overlays for classroom reading. *British Journal of Psychology*, 88 (4), 531-548. doi: 10.1111/j.2044-8295.1997.tb02656.x
26. Wilkins, A.J., Lewis, E., Smith, F., & Rowland, E. (2001). Coloured overlays and their benefit for reading. *Journal of Research in Reading*, 24 (1), 41-64. doi: 10.1111/1467-9817.00132
27. Wilkins, A., Cleave, R., Grayson, N., & Wilson, L. (2009). Typography for children may be inappropriately designed. *Journal of Research in Reading*, 32 (4), 402– 412. doi: 10.1111/j.1467-9817.2009.01402.x

28. Hughes, L.E. & Wilkins, A.J. (2000). Typography in children's reading schemes may be suboptimal: Evidence from measures of reading rate. *Journal of Research in Reading*, 23 (3), 314–324. doi: 10.1111/1467-9817.00126

29. Yammouni, R., & Evans, B.J.W.Is reading rate in digital eyestrain influenced by binocular and accommodative anomalies? *Journal of Optometry*, 2021; 14:229-239. doi: 10.1016/j.optom.2020.08.006

30. Eperjesi, F., Fowler, C. W., & Evans, B. J. (2004). The effects of coloured light filter overlays on reading rates in age-related macular degeneration. *Acta Ophthalmologica Scandinavica*, *82*(6), 695-700. doi: 10.1111/j.1600-0420.2004.00371.x

31. Ridder, W. H., Yoshinaga, P., Comer, G., & Ridder, S. (2017). Wilkins Reading Rates in Early and Intermediate AMD Compared to Age Matched Normal Patients. *Investigative Ophthalmology & Visual Science*, *58*(8), 3273-3273.

32. Ridder III, W. H., Zhang, Y., & Huang, J. F. (2013). Evaluation of reading speed and contrast sensitivity in dry eye disease. *Optometry and Vision Science*, *90*(1), 37-44. doi: 10.1097/OPX.0b013e3182780dbb

33. O'Leary, C.I., & Evans, B.J.W. (2006). Double-masked randomised placebo-controlled trial of the effect of prismatic corrections on rate of reading and the relationship with symptoms. *Ophthalmic and Physiological Optics*, 26 (6), 555-565. doi: 10.1111/j.1475-1313.2006.00400.x

34. Lambooij ,M., IJsselsteijn, W.A., Fortuin, M.F., Evans, B.J.W., & Heynderickx, I. (2010). Measuring visual discomfort associated with 3D displays. *J of the SID*, 18: 931–943.

35. Vander Stappen, C., & Van Reybroeck, C. (2018). Phonological awareness and rapid automatized naming are independent phonological competencies with specific impacts on word reading and spelling: an intervention study. *Front. Psychol.*, 9, 320. doi: 10.3389/fpsyg.2018.00320

FIGURE LEGENDS

Figure 1. The Rate of Reading Test (from Wilkins et al.[7])
Figure 2. Distributions of Test-Retest Mean Reading Rates
Figure 3. Bland-Altman Plots: Test-Retest Differences vs Means

come see the play look up is cat not my and dog for you to the cat up dog and is play come you see for not to look my you for the and not see my play come is look dog cat to up dog to you and play cat up is my not come for the look see play come see cat not look dog is my up the for to and you to not cat for look is my and up come play you see the dog my play see to for you is the look up cat not dog come and look to for my come play the dog see you not cat up and is up come look for the not dog cat you to see is and my play is you dog for not cat my look come and up to play see the

see the look dog and not is you come up to my for cat play not up play my is dog you come look for see and to the cat look up come and is my cat not dog you see for to play the my you is look the dog play see not come and to cat for up for the to and you cat is look up my not dog play see come you look see and play to the is cat not come for my up dog come not to play look the and dog see is cat up you for my and is for dog come see the cat up look you play my not to dog you cat to and play for not come up the see look my is the come to up cat my see dog you not look is play and for

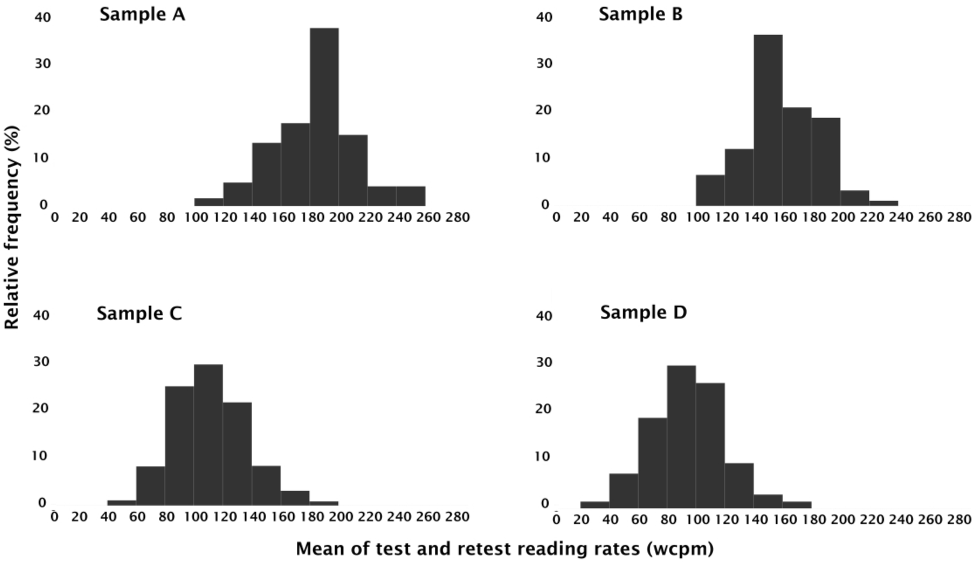

Figure 2. Distributions of Test Retest Mean Reading Rates

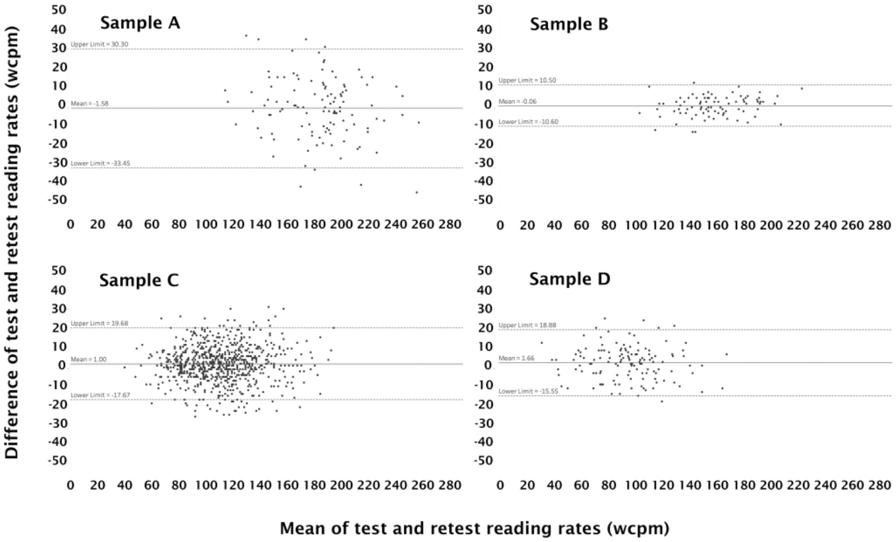

Figure 3. Bland-Altman Plots: Test-Retest Differences vs Test Retest Means