Test-retest Reliability of the qReading Method in Normally Sighted Young Adults

Timothy G. Shepard, PhD,[1] Zhong-Lin Lu, PhD,[2,3,4] and Deyue Yu, PhD, FAAO[1]*

SIGNIFICANCE: We recently developed a novel Bayesian adaptive method, qReading, to measure reading function. The qReading method has both the efficiency and excellent test-retest reliability in normally sighted young adults to make it an excellent candidate for future studies of its value in diagnosis and longitudinal evaluation of treatment and/or rehabilitation outcomes.

PURPOSE: A novel Bayesian adaptive method, qReading, was recently developed to measure reading function. Here we performed a systematic assessment of the test-retest reliability of the qReading method.

METHODS: The variability of five repeated measurements of the reading curve was examined in two settings: within session and between sessions. For the within-session design, we considered two subpopulations: naive observers and experienced observers. All observers were normally sighted young adults. For each set of data, in addition to examining the intrinsic precision of the qReading method (the half width of the credible interval of the posterior distribution of the estimated performance), we computed four metrics to assess repeatability: standard deviation, Bland-Altman coefficient of repeatability, correlation coefficient, and Fractional Rank Precision.

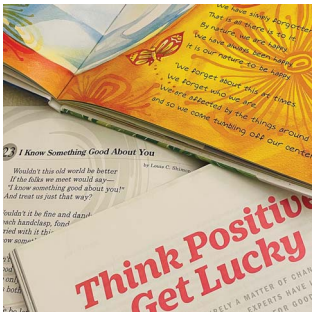
RESULTS: Extrinsic factors such as observer, time interval between repeated measures, and observer experience all contribute to the variation across measurements. Nevertheless, the four metrics consistently show that the variability across five repeated measurements is small for each set of data. This is true even without taking learning effects into account (standard deviations, ≤0.092 log10 units; Bland-Altman coefficient of repeatability, ≤0.15 $(\log10)^2$ units; correlation coefficient, ≥0.91; and Fractional Rank Precision, ≥0.81).

CONCLUSIONS: The qReading method has excellent test-retest reliability in normally sighted young adults.

Optom Vis Sci 2021;98:936–946. doi:10.1097/OPX.0000000000001754

Supplemental Digital Content: *Direct URL links are provided within the text.*

OPEN SDC



Author Affiliations:
[1]The Ohio State University College of Optometry, Columbus, Ohio
[2]Division of Arts and Sciences, NYU Shanghai, Shanghai, China
[3]Center for Neural Science and Department of Psychology, New York University, New York, New York
[4]NYU-ECNU Institute of Cognitive Neuroscience at NYU Shanghai, Shanghai, China
*yu.858@osu.edu

As one of the essential daily activities, reading allows us to obtain and exchange valuable information. Assessing reading performance is important in both clinics and laboratory because reading performance is a strong predictor of visual ability[1] and can be significantly affected by visual impairment,[2] which, in turn, has a substantial impact on the patient's quality of life.[3] It has also been shown that visual acuity, the most common functional vision endpoint, is not always sensitive to some retinopathies and their progression especially in the early stages of diseases.[4,5] Sometimes, visual acuity can be within the normal range or have little change while reading performance is impaired.[6] In situations like these, reading performance may be a more informative and useful measure.

Across a wide variety of reading materials, print size plays a crucial role in the legibility of text.[7] Most of existing reading tests measure reading speed in a range of print sizes to construct a reading curve that can be described using an exponential function (Fig. 1). Given the well-established relationship between reading speed and print size,[7] examining changes of the reading curve relative to age-matched controls provides an informative way for diagnosing non–age-related impairments in patients such as visual and cognitive impairments.

In clinics, the reading curve is typically measured using printed text such as the Bailey-Lovie Near Reading Card[8] and the MNREAD test.[9] Computerized reading tests such as the Flashcard test[10] and Rapid Serial Visual Presentation test[11] are mainly used in laboratory settings. In this study, we used the Rapid Serial Visual Presentation as the text presentation method to be consistent with our previous work.[12] In terms of testing method, one common challenge lies in the trade-off between the efficiency and the precision and accuracy of the measurements. To overcome this trade-off, we recently developed a novel Bayesian adaptive testing method, qReading.[12,13] By adopting a Bayesian adaptive testing framework, the qReading method provides efficient assessment while maintaining high accuracy of the measurements. Specifically, the qReading method uses Bayes' rule and an information-theoretic framework to select the most informative test stimulus (i.e., maximize information gain over a large selection of testing stimulus) for each test trial and exploits the functional regularity between reading speed and print size to apply the information harvested from each trial to the entire reading curve. In two studies,[12,13] we validated the qReading method with both computer simulations and psychophysical experiments. In both studies, the qReading procedure exhibited outstanding accuracy and high efficiency compared with conventional or other adaptive methods.

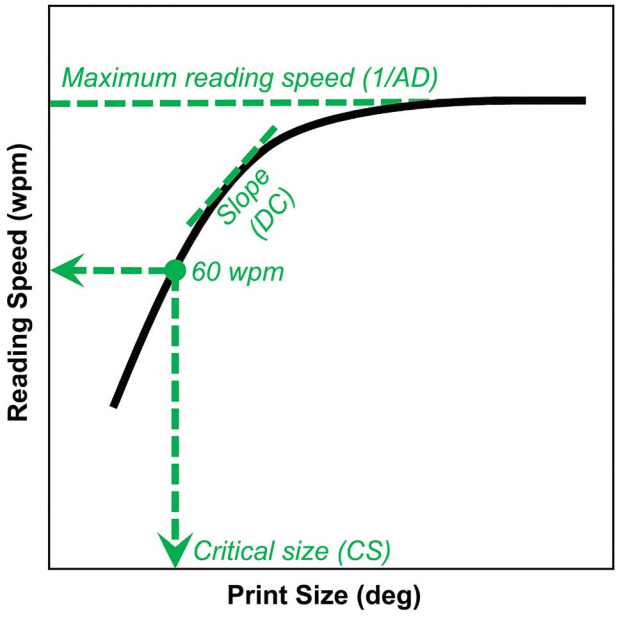


FIGURE 1. Reading speed (wpm) as a function of print size (lowercase *x*-height in degrees). The reading curve is described by an exponential function with three parameters: asymptotic performance level (AD; the threshold exposure duration corresponding to the maximum reading speed), the print size corresponding to a reading speed of 60 wpm (CS; the critical size), and slope of the function (DC; describing the changing rate of the reading curve). wpm = words per minute.

In addition to validating the qReading method, a systematic evaluation of the test-retest reliability (repeatability) across repeated measurements is also necessary before applying the method to clinical practice and research. When there is only one measurement of reading function from each observer,[12] the half width of the credible interval of the posterior distribution of the estimated performance is usually calculated as an indicator of precision (i.e., the reproducibility of the measurement of the reading function). However, the variation induced by any extrinsic factors (e.g., time interval between repeated measurements and level of experience of the observer) is not reflected in the half width of the credible interval calculation given the nature of the metric. In other words, half width of the credible interval describes the minimum test-retest variability, the intrinsic portion from the qReading method. Considering the potential applications of the qReading method in both clinics and laboratory, a more comprehensive evaluation of precision should consider both intrinsic and extrinsic factors, which, in the matter of data collection, requires obtaining more than one measurement from each observer. Although one qReading study[13] did acquire multiple runs of the qReading measurement, the data have some limitations: the qReading measurements were mingled with other reading trials, and the data are a mix of within-session and between-session repeated measures.

The goal of the present study is to perform a systematic assessment of the test-retest reliability of the qReading method in normally sighted observers. We examine the variability of repeated measurements of the reading curve in two settings: within session and between sessions. For the within-session design, we consider two subpopulations: naive observers and experienced observers. For each set of data, test-retest reliability of the method is estimated using multiple statistical measures. In addition, we analyzed the data with and without taking potential learning effects into account. We expect that the intrinsic variability of the qReading method (i.e., half width of the credible interval) remains invariant, whereas the variability due to the extrinsic factors may vary across testing conditions. Our results show that the qReading method has excellent test-retest reliability even without incorporating learning corrections.

METHODS

Observers

The present study is a detailed small-sample study. Thirteen naive, native English-speaking adults (age, 19 to 35 years) with normal or corrected-to-normal vision were divided into two groups. Seven observers participated in a one-session experiment. The other six observers participated in a five-session experiment. Statistics tests were performed to confirm that there was no significant difference in the initial reading performance between the one-session and five-session groups ($t_{11} = 1.62$, $P = .13$). The research protocol followed the tenets of the Declaration of Helsinki and was approved by the Ohio State University Institutional Review Board. Written consent was obtained from all observers after the nature and purpose of the study were explained.

Apparatus and Stimuli

The experiment was programmed and controlled using MATLAB (MathWorks, Ltd., Natick, MA), and Psychophysics Toolbox 3[14–16] and its EyeLink extensions[17] on a MacBook Pro. All stimuli (black lowercase English letters on white background with a mean luminance of 110 cd/m^2) were displayed at 10° in the lower visual field on a 32-inch Display++ (Cambridge Research Systems, Rochester, United Kingdom; resolution, 1920 × 1080 pixels; refresh rate, 120 Hz). Testing was binocular in a dark room at a viewing distance of 60 cm, which was maintained with a chin and head rest. The Courier font and its standard letter spacing were used. Eye movements were monitored monocularly at 1000 Hz with an EyeLink 1000 infrared video-based eye tracker (SR Research, Ottawa, Canada). Each testing block began with the calibration of the eye tracker using the standard nine-point fixation procedure.

Rapid Serial Visual Presentation Reading Task

The experiments used a Rapid Serial Visual Presentation reading task, identical to the paradigm used in the study by Shepard et al.[12] On each trial, a sentence was randomly drawn without replacement from a pool of 1180 meaningful sentences. Each sentence consists of 10 frequently used words in written English with word length of six or less. Words of the sentence were presented left justified on the left side of the display at 10° eccentricity in the lower visual field in a serial manner with a specified print size and exposure duration. Observers were asked to read the words aloud while maintaining stable fixation along a horizontal fixation line across the center of the display. A trial was canceled and replaced with a different sentence if fixation strayed more than ±1° vertically from the fixation line. Deviation of the fixation was detected in 4% of the trials. Note that the fixation line is not a required component in the Rapid Serial Visual Presentation method. We used it here to ensure that the normally sighted observers read using their peripheral vision. Given that each sentence contains 10 words, 10 responses (correct or incorrect) were generated from each trial. Missing words were treated as incorrect responses.

The qReading Method

The qReading method,[12,13] utilizing the Bayesian adaptive testing framework, was used to obtain the reading function (reading speed as a function of print size). The method consists of six components: a functional form of the reading curve (Fig. 1), a 3D parameter space, a 2D stimulus space, one-step-ahead search for stimulus selection for the upcoming trial, update of the joint posterior distribution (i.e., the new prior for the subsequent trial) of the reading function parameters based on observer's response, and iterations of the last two components until reaching a pre-determined criterion.

The reading curve is modeled with an exponential function (Eq. 1) with three parameters: asymptotic performance level (AD; the threshold exposure duration corresponding to the maximum reading speed), the print size corresponding to a reading speed of 60 words per minute (CS; the critical size), and the slope of the function (DC; describing the changing rate of the reading curve):

$$\log_{10}\text{Reading Speed} = \log_{10}\left(\frac{60}{\text{AD}}\right) + \log_{10}(\text{AD})e^{-\left(\frac{\log_{10}\text{Print Size}-\log_{10}\text{CS}}{\text{DC}}\right)}. \quad (1)$$

A 3D parameter space is defined to encompass all potentially observable reading curves of the observer population and testing condition. Specifically, $\log_{10}$AD ranges from –1.3 to 0 (corresponding to 1200 to 60 words per minute), $\log_{10}$CS ranges from –0.7 to 0.3 (corresponding to 0.2° to 2.0° print size), and $\log_{10}$DC ranges from –1.3 to 0.2. Within the range of each parameter, 30 values were evenly sampled during data collection.

The 2D stimulus space contains 25 log-spaced print sizes between 0.5° and 4.2° and 50 log-spaced exposure durations between 25 and 3000 milliseconds (3 to 360 frames). Reading speed for each print size was derived from a psychometric function (proportion of words read correctly vs. exposure duration) specified by the parameters of the reading curve using a criterion of 80% correct (see more detailed information in Shepard et al.[12]).

A one-step-ahead search is performed to select a combination of print size and exposure duration for the upcoming trial with the goal of optimizing the expected information gain on the reading curve. After the observer's response in each trial, the joint posterior distribution of the three parameters is updated using Bayes' rule and serves as the new prior for the upcoming trial. Because there are 10 responses (correct or incorrect) per trial, the joint posterior distribution is updated 10 times after each trial. We repeat the one-step-ahead search and update the joint posterior distribution until a pre-determined criterion has been achieved (e.g., reaching a certain number of trials).

Procedure and Data Analysis

The variability of repeated measurements of the reading curve using the qReading method was investigated in both within- and between-session designs. The observers in the one-session group completed five qReading blocks. The five-session group received one qReading block in each of the first four sessions and five qReading blocks in the fifth session and completed all sessions within 7 to 16 days. Each qReading block contained 50 Rapid Serial Visual Presentation reading trials. Each session began with 20 practice trials.

A finer grid (100 instead of 30 values for each parameter; same ranges) was adopted in the data analysis. We compared reading functions across five within-session blocks for the one-session group and for the fifth session of the five-session group (they were considered as experienced observers when reaching the fifth session). A similar comparison was performed across five between-session blocks (the block from each of the first four sessions and the first block in the fifth session) for the five-session group. In this study, we refer to the three sets of data (or conditions) by one-session within-design, five-session within-design, and five-session between-design, respectively. For each

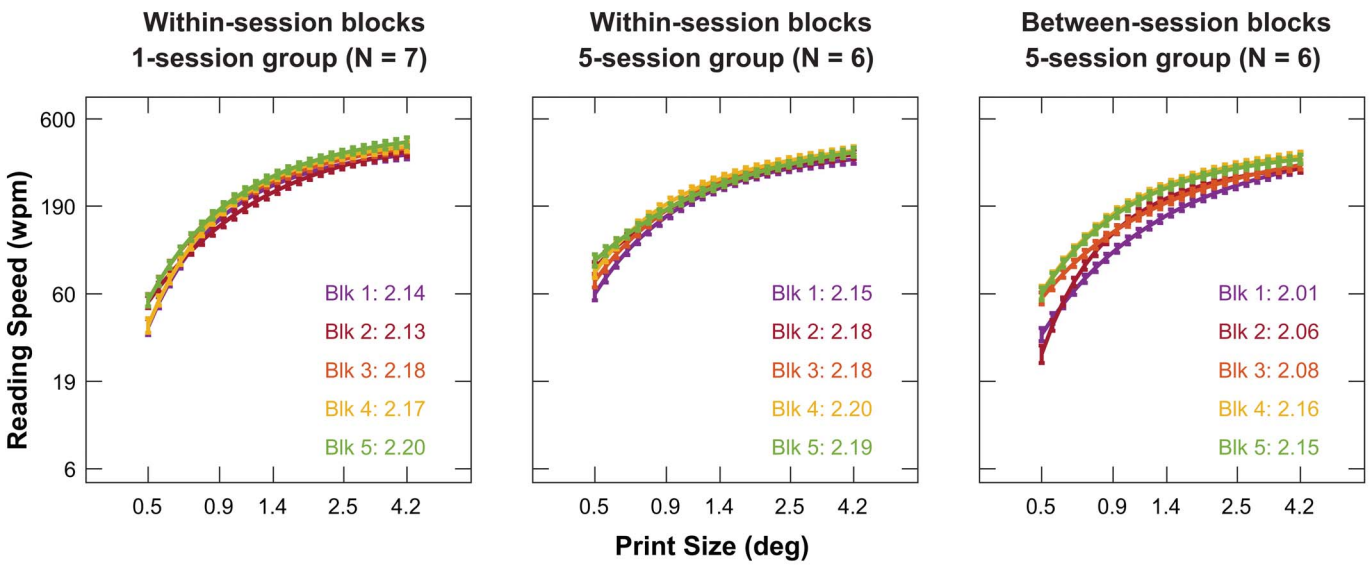


FIGURE 2. Average reading curves for the five within-session blocks of the one-session group, the five within-session blocks of the five-session group, and the five between-session blocks of the five-session group. Each curve represents the group average estimation. The error bars represent the average ± 68.2% half width of the credible intervals across observers. Average areas under the curve are listed for all five blocks in each panel.

set of data, we examined the area under the curve and half width of the credible interval across blocks[12] and computed the standard deviation, Bland-Altman coefficient of repeatability (with 95% confidence limits),[18,19] correlation coefficient, and Fractional Rank Precision[20] between blocks to evaluate the test-retest reliability of the method. We evaluated the data and report the results in two ways: with or without taking learning into account.

RESULTS

Repeatability without Learning Correction

Reading Curves

Fig. 2 shows the reading curves averaged across observers for each qReading block. The three subplots are for the five

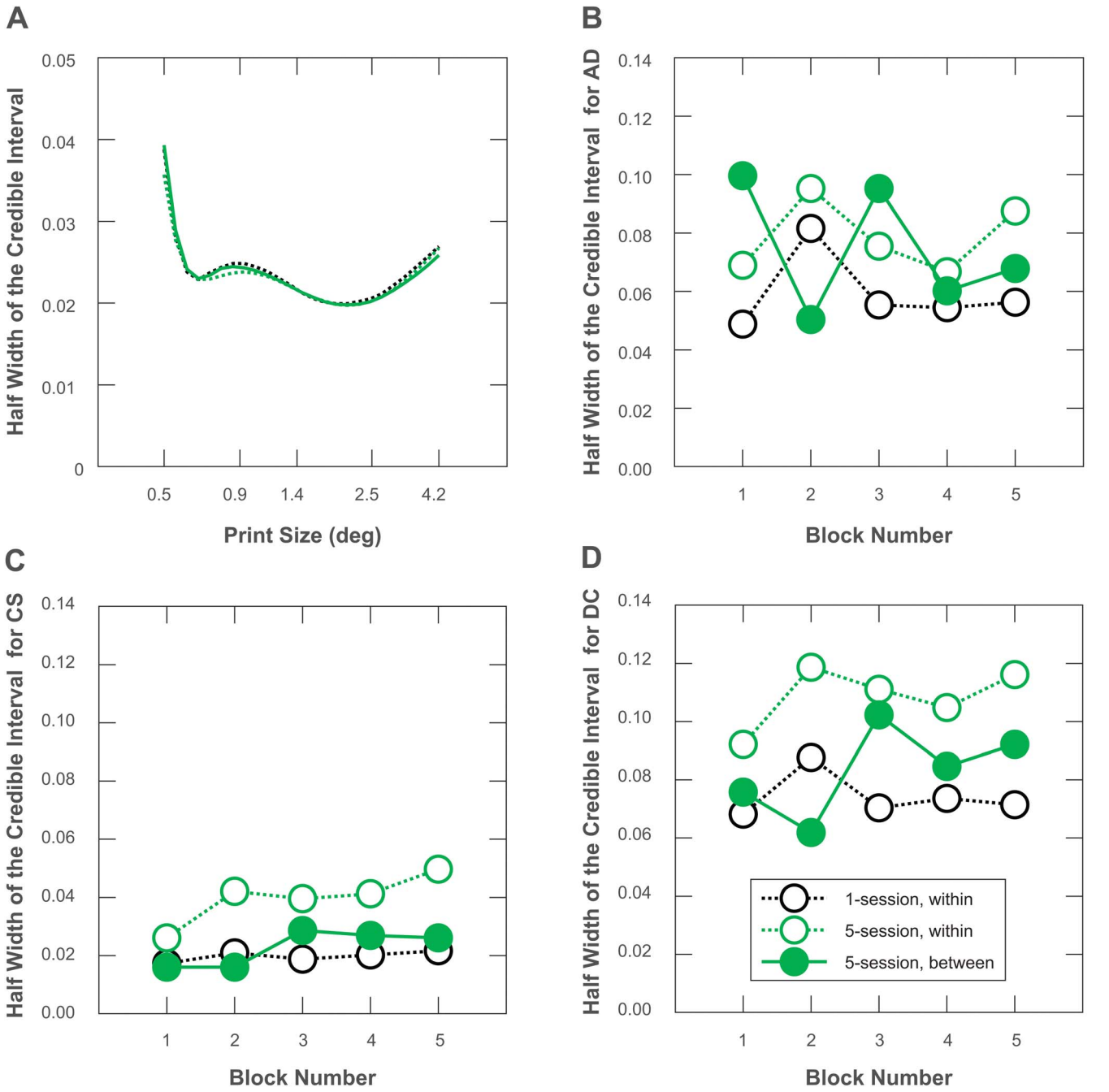


FIGURE 3. A, Half width of the credible interval (averaged across observers and blocks) as a function of print size for one-session within-design, five-session within-design, and five-session between-design conditions. B, C, and D, Half width of the credible interval (averaged across observers) as a function of block number for each condition for AD, CS, and DC. AD = asymptotic performance level; CS = critical size; DC = slope of the function.

within-session blocks of the one-session group, the five within-session blocks of the five-session group, and the five between-session blocks of the five-session group.

Area under the Curve

Area under the curve allows us to quantify the overall reading performance across a range of print sizes.[12] For each estimated reading curve, area under the curve is computed as the area enclosed by the reading curve and the horizontal line through log10(1 word per minute) in the print size range assessed in the experiment (i.e., 0.5° to 4.2°). Because both *x* and *y* axes are in the log10 scale, the unit for area under the curve is the square of log10 units. Area under the curve/(log10(4.2°) – log10(0.5°)) provides the average log10(words per minute) across the print sizes tested. Fig. 2 also includes the group average of area under the curve listed for each block. A one-way repeated-measures ANOVA was performed on each of the three data sets to examine the block effect. There was a significant block effect for one-session within-design ($F_{4,24} = 4.64$, $P = .006$) and five-session between-design ($F_{4,20} = 9.17$, $P = .0005$), but not for five-session within-design ($F_{4,20} = 1.60$, $P = .21$). The standard deviation of the areas under the curve across the five blocks was calculated from the group average. For the one-session within-design, five-session within-design, and five-session between-design conditions, the standard deviations were 0.03, 0.02, and 0.06 $(\log10)^2$ units, respectively. The corresponding coefficients of variation (i.e., the ratio of the standard deviation to the mean) were minimal (1.34%, 0.97%, and 3.02%, respectively), indicating good test-retest reliability.

Half Width of the Credible Interval

As demonstrated previously,[12] the 68.2% half width of the credible interval of the posterior distribution of the estimated performance provides a measure of precision in a single run of the qReading procedure. Half width of the credible interval typically decreases with increasing the number of trials. Here we computed the half width of the credible interval after the 50th trial of each block. First, we examined half width of the credible interval as a function of print size. The results showed that variation of half width of the credible interval was minimal across blocks and conditions. As revealed by Fig. 3A, half width of the credible interval changed with print size in a nonmonotonic fashion and was greater at smaller and some larger print sizes, which may be due to a combination of multiple factors such as the fluctuation of observer's performance across print sizes, the selected functional form of the reading function, and the distribution of testing stimuli in the stimulus space. The overall mean half width of the credible intervals for one-session within-design, five-session within-design, and five-session between-design conditions were 0.024, 0.023, and 0.023 log10 units, respectively, consistent with previous results.[12] Half width of the credible interval can also be calculated for each of the three parameters of the reading curve. Half width of the credible intervals for the parameters (Figs. 3B to D and Table 1) were small compared with the ranges of the parameters (1.3 log10 units for AD, 1 log10 unit for CS, and 1.5 log10 units for DC) in all three conditions (2 to 7%).

Standard Deviation

Test-retest reliability can be expressed in terms of standard deviation. The standard deviation across blocks can be calculated for reading performance measured at each print size for each individual observer. Fig. 4A shows the mean standard deviation (averaged across observers) as a function of print size in the three conditions. We also estimated the amount of variation due to repeated measures at the group level, standard $\text{deviation}_{\text{group}}$ (calculated from the group reading curves [i.e., the average reading curves across observers]; Fig. 4C). Standard $\text{deviation}_{\text{group}}$ should include the variations introduced by both learning and nonlearning factors. Overall, standard deviations were larger at smaller print sizes. Table 1 lists the standard deviations averaged across observers and print sizes. When plotting standard deviation and standard $\text{deviation}_{\text{group}}$ versus the number of blocks from which calculations were made, we observed similar or slightly smaller standard deviations when

TABLE 1. Half width of the credible intervals (in log10 units, averaged across observers and blocks) for the three parameters (AD, CS, and DC) of reading curve, standard deviations (mean standard deviation, standard $\text{deviation}_{\text{group}}$, standard $\text{deviation}_{\text{nonlearning}}$ in log10 units), and correlation coefficient and Fractional Rank Precision (mean ± standard deviation) with or without the correction for learning for the one-session within-design, five-session within-design, and five-session between-design conditions

	One-session within-design	Five-session within-design	Five-session between-design
AD	0.059	0.079	0.075
CS	0.020	0.040	0.023
DC	0.074	0.109	0.083
Mean standard deviation	0.063	0.052	0.092
Standard $\text{deviation}_{\text{group}}$	0.037	0.026	0.073
Standard $\text{deviation}_{\text{nonlearning}}$	0.021	0.018	0.030
Correlation coefficient			
No correction	0.96 ± 0.04	0.96 ± 0.04	0.91 ± 0.08
Correction	0.98 ± 0.02	0.98 ± 0.02	0.97 ± 0.03
Fractional Rank Precision			
No correction	0.89 ± 0.05	0.85 ± 0.05	0.81 ± 0.08
Correction	0.94 ± 0.05	0.88 ± 0.07	0.91 ± 0.07

AD = asymptotic performance level; CS = critical size; DC = slope of the function.

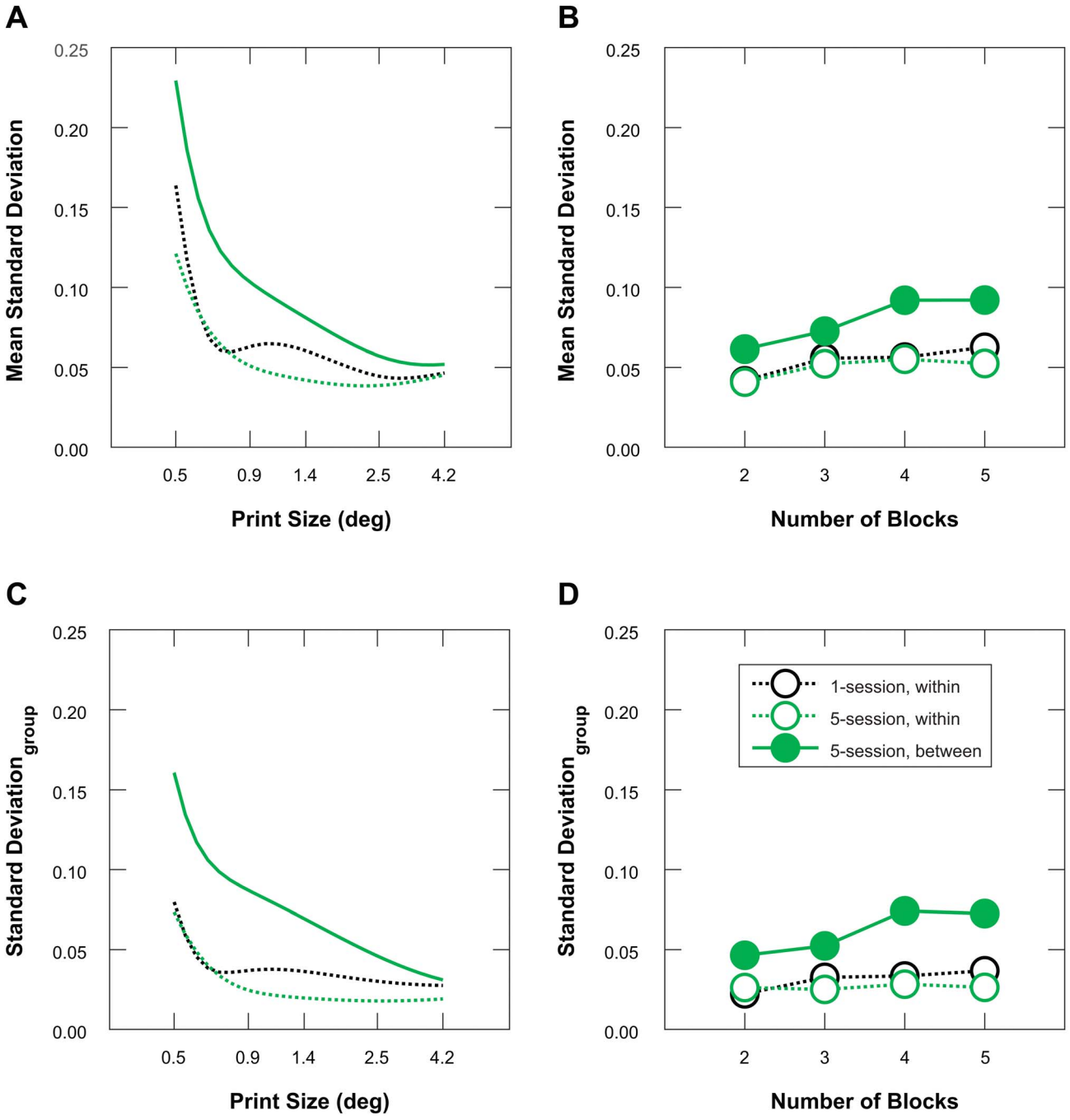


FIGURE 4. Standard deviations (in log10 units). A, Mean standard deviation (averaged across observers) as a function of print size for one-session within-design, five-session within-design, and five-session between-design conditions. B, Standard deviation versus the number of blocks from which calculations are made. C, Standard deviation$_{group}$ as a function of print size. D, Standard deviation$_{group}$ versus the number of blocks.

fewer than four blocks were considered (Figs. 4B, D). An increase of standard deviation with a larger number of blocks indicated performance change in later blocks.

Bland-Altman Plot

Bland-Altman statistics[18,19] were performed to assess the level of agreement between repeated measurements of the area under the curve of the reading curve. In Fig. 5, the Bland-Altman plot is constructed for each of the three conditions. Given five repeated measures per observer, there are 10 unique test-retest combinations (i.e., 10 data points) for each observer. In each data point, the first measurement was always taken before the second measurement during the data collection. Coefficient of repeatability was calculated based on the 10 data points for each observer and condition. The mean coefficients of repeatability for the three conditions were 0.10, 0.10, and 0.15 $(\log10)^2$ units (corresponding to 5%, 5%, and 7% differences in area under the curve), respectively.

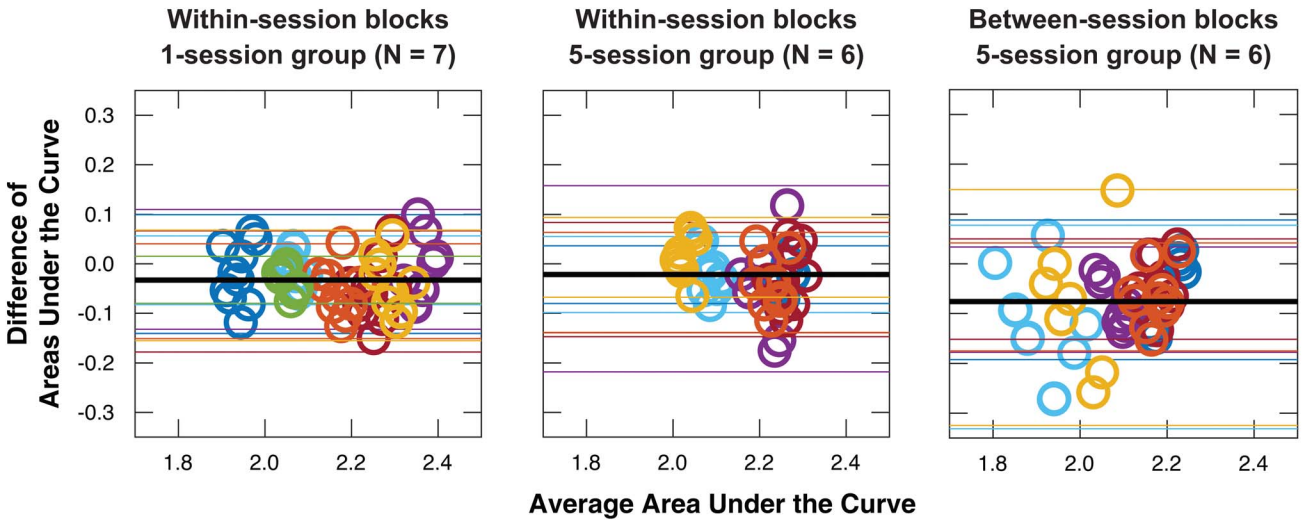


FIGURE 5. Bland-Altman plots for one-session within-design, five-session within-design, and five-session between-design conditions. Each color represents one observer. Each observer contributes 10 data points. The dashed horizontal lines indicate the individual 95% limits of agreement. The solid black line signifies the bias averaged across observers.

Similar coefficients of repeatability were obtained when pooling across observers with two test-retest combinations (sampled without replacement) per observer. Across all observers, the difference between the area under the curve measurements from different blocks always fell within the individual limits of agreement (except one data point that was right next to the borderline in the five-session between-design condition). The five-session between-design condition had higher variability overall. We also observed small biases in the three conditions (–0.03, –0.02, and –0.08 $(\log10)^2$ units); that is, the later measurements were slightly larger than the earlier measurements.

Correlation

Repeatability of the estimated reading curves can also be evaluated using correlation. Because the reading speeds on each reading curve are determined by the same exponential function and are therefore not independent, we adopted a procedure similar to Hou et al.[21] to remove the dependency from the analysis. First, we designated N log-spaced print sizes between 0.5° and 4.2° ($N = 7$ for the one-session condition; $N = 6$ for the two five-session conditions). For each print size, we acquired one test-retest data point (reading speed) by randomly selecting an observer and then randomly selecting two blocks from the same observer (both without replacement). Then, we ran correlation analysis on the N pairs of reading speeds. Lastly, we repeated the aforementioned two steps 500 times in each of the three conditions. As summarized in Fig. 6 and Table 1, the mean correlation coefficient across iterations was very high (≥0.91) in all three conditions.

Fractional Rank Precision

Fractional Rank Precision, a rank-based measure, has been recently developed to assess test-retest reliability in terms of population variability.[20] Specifically, the analysis is to identify the test-retest pair of measurements for an observer, given only the test measurement for the observer and a set of retest measurements from all observers in the group. Because test-retest reliability is described in terms of interobserver variability without resorting to absolute values, the interpretation of the result is intuitive, and the measurement is also suitable for comparing tests with outputs of different magnitude or dimensionality. However, because Fractional Rank Precision scores depend on the variability of the specific observer cohort, comparison of results is meaningful only within but not across studies.[20] To calculate the Fractional Rank Precision, first, we sort all observers' retest values by their Euclidean distance to the target observer's test value in ascending order. Then, we calculate the fractional precision $(1 – (\text{rank} – 1)/N)$ for this observer's retest measurement, where rank denotes the rank of the observer's retest measurement in the sorted sequence and N is the number of observers in the group. For instance, if the rank equals 1 (i.e., the observer's retest value has the shortest Euclidean distance to the observer's test value among all), the fractional precision would be 1, indicating a perfect test-retest identification. We repeated the aforementioned steps for each observer, and the Fractional Rank Precision was calculated as the average of all observers' fractional precisions. Randomly distributed test and retest scores would result in a Fractional Rank Precision of 0.5. Five repeated measurements from each observer provided 10 unique test-retest pairs. We calculated the Fractional Rank Precision for each pair and each of the three conditions. As shown in Table 1, the mean Fractional Rank Precision was equal to or higher than 0.81 in all three conditions.

Learning Rate

As shown in Fig. 7, there is a trend of an increasing area under the curve with the number of repeated measurements. We calculated learning rate by fitting a straight line to the data in each condition and for each observer. The learning rates in the three conditions (one-session within-design, five-session within-design, and five-session between-design) were 0.016 ± 0.009 (standard deviation), 0.011 ± 0.009, and 0.038 ± 0.014 $(\log10)^2$/block, respectively. Because difference in the initial reading performance may confound the comparison between the one-session within-design and five-session between-design conditions and the

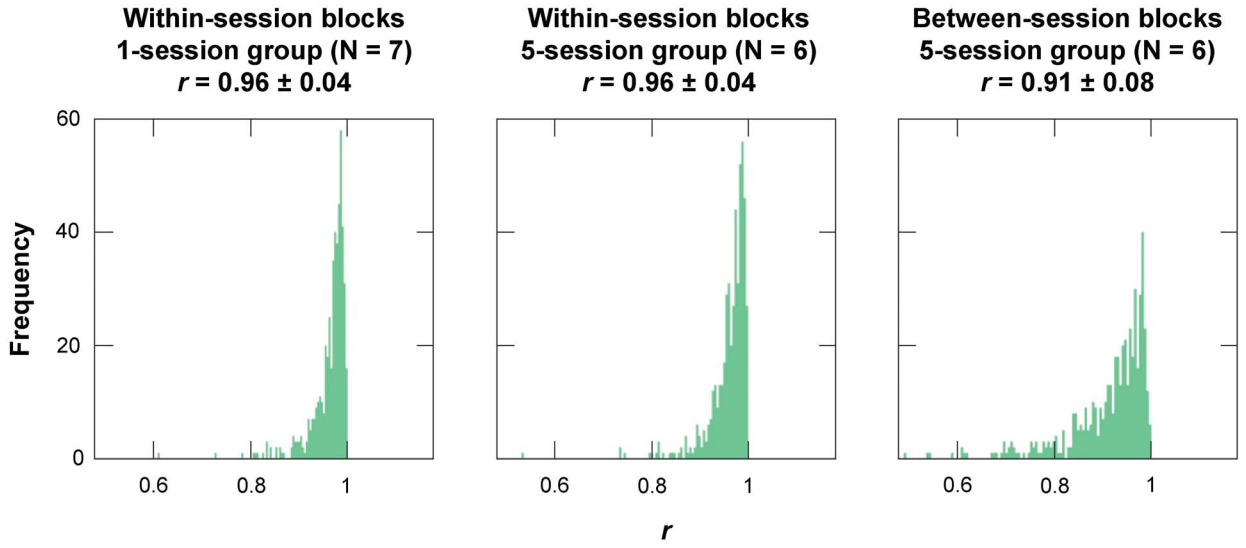


FIGURE 6. Histograms of correlation coefficients for one-session within-design, five-session within-design, and five-session between-design conditions. Each plot contains correlation coefficients calculated from 500 iterations of sampling.

comparison between the two within-design conditions, we performed statistical tests and confirmed that reading performance was equated for each of the two pairs at baseline (t_{11} = 1.62, P = .13; t_{11} = 0.12, P = .91).

We can also calculate a point-by-point learning rate across the reading curve. For instance, we can compute the learning rate for each of the 25 log-spaced print sizes between 0.5° and 4.2°. Fig. 8 shows that learning rate is essentially the same across print sizes for one-session within-design, decreases with print size until about 1° for five-session within-design, and exhibits a monotonic decrease for five-session between-design. The five-session between-design condition has the highest learning rate overall. The learning rates are less variable across observers and more similar across the three conditions at larger print sizes.

Repeatability with Learning Correction

We reassessed the variability of the repeated measurements of the reading curve after taking learning into account. Correction for learning was made to reading speed and area under the curve: Corrected value = Original value – (Block number – 1) × Learning rate. We recomputed the standard deviation, Bland-Altman coefficient of repeatability, correlation coefficient, and Fractional Rank Precision after the correction. Although individual learning rates were applied here, the results were either identical or highly similar (both qualitatively and quantitatively) to those based on a group learning rate correction.

After learning correction, the reading curves became more similar across blocks (Appendix Fig. A1, available at http://links.lww.

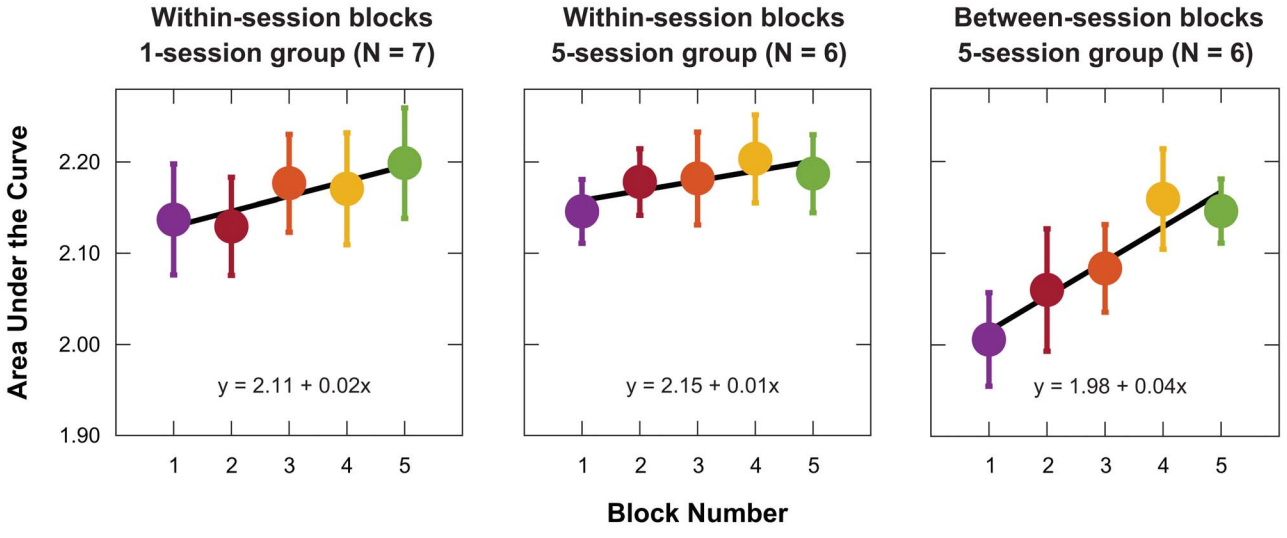


FIGURE 7. Average area under the curve as a function of block number for one-session within-design, five-session within-design, and five-session between-design conditions. The error bars denote standard error. The black lines and equations represent the best-fitted lines to the data.

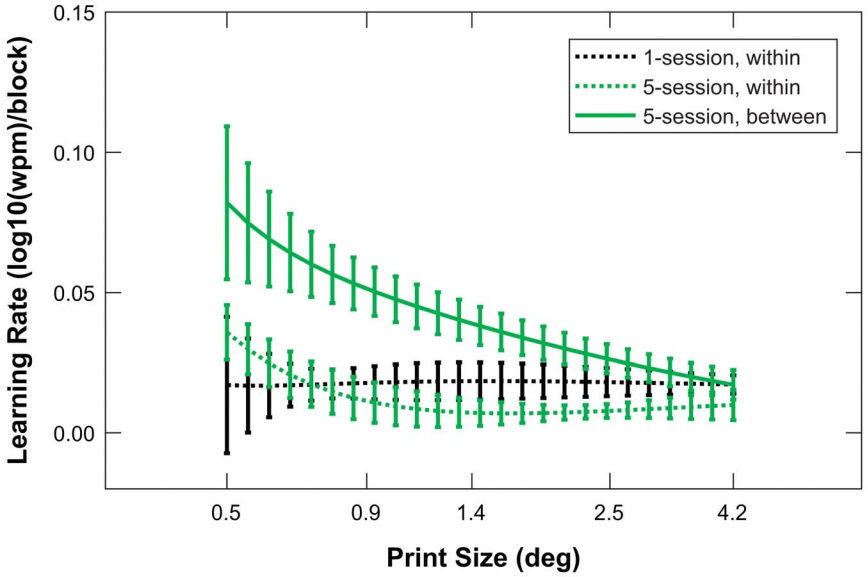


FIGURE 8. Average learning rate as a function of print size for one-session within-design, five-session within-design, and five-session between-design conditions. The error bars denote standard error.

com/OPX/A505, which shows average reading curves after the correction for learning). One-way repeated-measures ANOVAs showed no significant effect of block on the area under the curve in all three conditions. Both the standard deviations of the areas under the curve across the five blocks and the corresponding coefficients of variation were also reduced.

As expected, standard deviations became smaller with the correction for learning (Appendix Fig. A2, available at http://links.lww.com/OPX/A506, which shows standard deviations after the correction for learning). In general, standard deviations are still larger at smaller print sizes. After taking learning into account, the standard deviations remained constant for different numbers of blocks.

With the correction for learning, the biases in the Bland-Altman plots reduced to zero in all three conditions (Appendix Fig. A3, available at http://links.lww.com/OPX/A507, which shows Bland-Altman plots after the correction for learning). The coefficients of repeatability also slightly reduced. Across observers, the difference between the estimated areas under the curve from different blocks always fell within the individual limits of agreement.

When performing correlation and Fractional Rank Precision analyses on the learning-corrected data, we found higher Fractional Rank Precisions (Table 1) and slightly higher correlation coefficients with smaller standard deviations (Table 1; Appendix Fig. A4, available at http://links.lww.com/OPX/A508, which shows histograms of correlation coefficients after the correction for learning).

DISCUSSION

We performed a systematic evaluation of the test-retest reliability of the qReading method in normally sighted observers reading peripherally. By comparing repeated measurements of the reading curve, our evaluation on precision considered both the intrinsic precision of the qReading method and the variability from several extrinsic factors such as observer, time interval between repeated measures, and the level of experience of the observer. As expected, the half width of the credible interval (the intrinsic precision of the qReading method) not only was small and quantitatively consistent with the previous finding by Shepard et al.[12] but also had minimal change across multiple measurements and testing conditions. The variation observed in the test-retest reliability (e.g., the variation of standard deviation) was mostly caused by extrinsic factors. We adopted four metrics to assess the repeatability of the qReading method, including the common metrics for repeatability measurement,[22,23] standard deviation, correlation coefficient (despite being vulnerable to artifacts), and Bland-Altman coefficient of repeatability, and a newly developed metric based on concepts from information retrieval (Fractional Rank Precision[20]). All four metrics consistently showed that the qReading method had excellent test-retest reliability even without taking learning into account.

In the present study, we considered three extrinsic factors and evaluated their impact on the test-retest variability. The one-session within-design and five-session between-design conditions represented two different measurement schedules (i.e., different time intervals between repeated measures). All four repeatability metrics showed that test precision was lower when the repeated measures were spread over multiple days instead of being collected in one quick session (less than an hour). Compared with naive observers in the one-session within-design condition, the observers in the five-session within-design condition completed four qReading blocks before the within-session data collection and therefore had more experience in performing Rapid Serial Visual Presentation reading task with the qReading procedure. The results showed that the experienced group seemed to have slightly lower standard deviations but similar values in the other metrics; that is, test-retest reliability was similar or slightly better for the experienced group. As we will discuss hereinafter, learning can occur during repeated measurements, which accounts for part of the variability. Other observer-dependent factors can also contribute to test-retest variability (standard deviation$_{\text{nonlearning}}$). Both the learning and nonlearning factors led to the most variability in the five-session

between-design condition and the least variation in the five-session within-design condition.

Learning is inevitable when an observer repeatedly performs the same task for a large number of trials. Fig. 7 shows a visible trend of improvement in area under the curve with an increasing number of repeated measurements, which was also confirmed by the small negative biases (indicating that the later measurements were slightly larger than the earlier measurements) observed in Bland-Altman plots (Fig. 5). The rate or magnitude of improvement may depend on various aspects of the study design such as the initial performance level,[24] the total amount of stimulus exposure,[24] and the length of each testing session (i.e., measurement schedule).[25] To thoroughly examine the test-retest reliability of the qReading method, we estimated the learning rate for each condition and individual and then reevaluated the data with correction for learning. We found that learning had the largest impact on the five-session between-design condition (indicated by the steepest slope in Fig. 7) and the smallest influence on the five-session within-design condition. Namely, observers had faster/greater improvement when measurements were distributed over multiple days or when observers were naive. After taking learning into account, we found better test-retest reliability for all three conditions, with the five-session between-design condition benefiting the most.

When we measured each reading curve, reading speed was estimated at different print sizes between 0.5° and 4.2°. As shown in Figs. 3 and 4, smaller print sizes (corresponding to slower reading speeds) exhibited lower test-retest reliability (larger half width of the credible intervals and standard deviations), especially for those close to 0.5°. Regardless of exposure duration, smaller text is more difficult to read. Likely, the more visually challenging the text, the greater the reliance on nonvisual factors such as lexical knowledge[26] and sentence context.[27] This means that more fluctuations can be induced in the performance and the associated measures at smaller print sizes. When calculating point-by-point learning rate across the reading curve, we found higher learning rates at smaller print sizes for the two five-session conditions. However, even after taking learning into account, the standard deviations remained larger at smaller print sizes in all three conditions.

The Rapid Serial Visual Presentation involves minimal eye movements and is very different from natural reading. Future studies should consider adopting text presentation methods that require eye movements, such as page reading or self-paced reading, to better estimate eye-mediated reading performance in everyday life. As demonstrated by Arango et al.,[28] an easy modification of the qReading method will allow us to test these alternative conditions. For instance, in self-paced reading, observers reveal words of a sentence one at a time at their own pace while having continued access to the overall spatial layout of the sentence. This presentation method has been broadly used in psycholinguistic studies and recently adopted to assess low-vision reading because of its advantage of providing extract reading time of each word.[29] To apply the qReading technique to the self-paced reading, in the one-step-ahead search, we would only need to select a print size instead of a combination of print size and exposure duration for the upcoming trial. By recording the observer's response (read the word correctly or incorrectly) and the reading time of each word, we would then be able to update the joint posterior distribution of the parameters of the reading function.

Shepard et al.[12] and Hou et al.[13] validated the qReading method mainly for its accuracy and efficiency. The present study performed a systematical assessment of its precision and demonstrated high test-retest reliability of the method. Further evaluations in different age groups and patient populations are needed to examine the suitability of the method in various potential applications (e.g., diagnosis, or longitudinal assessment of disease progression and treatment and/or rehabilitation outcome). With possible poorer test-retest reliability in people with visual impairment,[30] it is especially important to perform a similar assessment among visually impaired individuals. When using the qReading method in the clinics, we may need to consider the effect of learning to further improve the precision of the method, especially for longitudinal assessments.

ARTICLE INFORMATION

Supplemental Digital Content: Appendix Figure A1, available at http://links.lww.com/OPX/A505. Average reading curves after the correction for learning. Each curve represents group average estimation. The error bars represent the average ± 68.2% half width of the credible intervals across observers. Average areas under the curve are listed for all five blocks in each panel. The standard deviations of the areas under the curve across the five blocks are 0.01, 0.01, 0.02 $(\log10)^2$ units for the one-session within-design, five-session within-design, and five-session between-design conditions, respectively. The corresponding coefficients of variation are 0.58%, 0.57%, and 0.97%.

Appendix Figure A2, available at http://links.lww.com/OPX/A506. Standard deviations (in log10 units) after the correction for learning. (A) Mean standard deviation (averaged across observers) as a function of print size for one-session within-design, five-session within-design, and five-session between-design conditions. (B) Standard deviation versus the number of blocks from which calculations are made. (C) Standard $\text{deviation}_{\text{nonlearning}}$ (the amount of variation introduced by repeated measures at the group level excluding the variation resulted from learning) as a function of print size. (D) Standard $\text{deviation}_{\text{nonlearning}}$ versus the number of blocks.

Appendix Figure A3, available at http://links.lww.com/OPX/A507. Bland-Altman plots after the correction for learning. Each color represents one observer. Each observer contributes ten data points. The dashed horizontal lines indicate the individual 95% limits of agreement. The solid black line represents the bias averaged across observers. The mean coefficients of repeatability are 0.09, 0.10, and 0.13 $(\log10)^2$ units (corresponding to 4%, 5% and 6% difference in area under the curve) for one-session within-design, five-session within-design, and five-session between-design conditions, respectively.

Appendix Figure A4, available at http://links.lww.com/OPX/A508. Histograms of correlation coefficients after the correction for learning. Each plot contains correlation coefficients calculated from 500 iterations of sampling.

Submitted: June 29, 2020

Accepted: April 6, 2021

Funding/Support: National Eye Institute (EY025658; to DY) and National Eye Institute (EY021553; to Z-LL).

Conflict of Interest Disclosure: Z-LL and DY own intellectual property rights on qReading technology. Z-LL has equity interest in Adaptive Sensory Technology, Inc. The authors were responsible for the preparation of this manuscript and the decision to submit this article for publication. Each of the authors had full access to the study data and takes full responsibility for the presentation in this article.

Author Contributions: Conceptualization: Z-LL, DY; Data Curation: TGS, DY; Formal Analysis: TGS, DY; Funding Acquisition: Z-LL, DY; Investigation: TGS; Methodology: Z-LL, DY; Project Administration: DY; Resources: DY; Software: TGS, DY; Supervision: DY; Validation: DY; Visualization: DY; Writing – Original Draft: TGS, DY; Writing – Review & Editing: TGS, Z-LL, DY.

REFERENCES

1. McClure M, Hart P, Jackson A, et al. Macular Degeneration: Do Conventional Measurements of Impaired Visual Function Equate with Visual Disability? Br J Ophthalmol 2000;84:244–50.

2. Crossland MD, Gould ES, Helman CG, et al. Expectations and Perceived Benefits of a Hospital-based Low

Vision Clinic: Results of an Exploratory, Qualitative Research Study. Vis Impair Res 2007;9:59–66.

3. Mitchell J, Wolffsohn J, Woodcock A, et al. The MacDQol Individualized Measure of the Impact of Macular Degeneration on Quality of Life: Reliability and Responsiveness. Am J Ophthalmol 2008;146: 447–54.

4. Klein R, Wang Q, Klein B, et al. The Relationship of Age-related Maculopathy, Cataract, and Glaucoma to Visual Acuity. Invest Ophthalmol Vis Sci 1995;36: 182–91.

5. Sunness JS, Rubin GS, Applegate CA, et al. Visual Function Abnormalities and Prognosis in Eyes with Age-related Geographic Atrophy of the Macula and Good Visual Acuity. Ophthalmology 1997;104:1677–91.

6. Crossland MD, Culham LE, Rubin GS. Fixation Stability and Reading Speed in Patients with Newly Developed Macular Disease. Ophthalmic Physiol Opt 2004;24: 327–33.

7. Legge GE, Bigelow CA. Does Print Size Matter for Reading? A Review of Findings from Vision Science and Typography. J Vis 2011;11:10.1167/11.5.8 8.

8. Bailey IL, Lovie JE. The Design and Use of a New Near-vision Chart. Am J Optom Physiol Opt 1980;57: 378–87.

9. Mansfield J, Ahn S, Legge G, et al. A New Reading-acuity Chart for Normal and Low Vision. Opt Soc Am Techn Digest 1993;3:232–5.

10. Legge GE, Ross JA, Luebker A, et al. Psychophysics of Reading. VIII. The Minnesota Low-vision Reading Test. Optom Vis Sci 1989;66:843–53.

11. Rubin GS, Turano K. Low Vision Reading with Sequential Word Presentation. Vision Res 1994;34: 1723–33.

12. Shepard TG, Hou F, Bex PJ, et al. Assessing Reading Performance in the Periphery with a Bayesian Adaptive Approach: The qReading Method. J Vis 2019;19:5.

13. Hou F, Zhao Y, Lesmes LA, et al. Bayesian Adaptive Assessment of the Reading Function for Vision: The qReading Method. J Vis 2018;18:6.

14. Brainard DH. The Psychophysics Toolbox. Spat Vis 1997;10:433–6.

15. Kleiner M, Brainard D, Pelli D. What's New in Psychtoolbox-3? Perception 2007;36:14.

16. Pelli DG. The Videotoolbox Software for Visual Psychophysics: Transforming Numbers into Movies. Spat Vis 1997;10:437–42.

17. Cornelissen FW, Peters EM, Palmer J. The Eyelink Toolbox: Eye Tracking with Matlab and the Psychophysics Toolbox. Behav Res Methods Instrum Comput 2002;34:613–7.

18. Bland JM, Altman DG. Statistical Methods for Assessing Agreement between Two Methods of Clinical Measurement. Lancet 1986;1:307–10.

19. Bland JM, Altman DG. Measuring Agreement in Method Comparison Studies. Stat Methods Med Res 1999;8:135–60.

20. Dorr M, Elze T, Wang H, et al. New Precision Metrics for Contrast Sensitivity Testing. IEEE J Biomed Health Inform 2018;22:919–25.

21. Hou F, Lesmes LA, Kim W, et al. Evaluating the Performance of the Quick CSF Method in Detecting Contrast Sensitivity Function Changes. J Vis 2016;16:18.

22. McAlinden C, Khadka J, Pesudovs K. Precision (Repeatability and Reproducibility) Studies and Sample-size Calculation. J Cataract Refract Surg 2015; 41:2598–604.

23. McAlinden C, Khadka J, Pesudovs K. Statistical Methods for Conducting Agreement (Comparison of Clinical Tests) and Precision (Repeatability or Reproducibility) Studies in Optometry and Ophthalmology. Ophthalmic Physiol Opt 2011;31:330–8.

24. Husk JS, Yu D. Learning to Recognize Letters in the Periphery: Effects of Repeated Exposure, Letter Frequency, and Letter Complexity. J Vis 2017;17:3.

25. Molloy K, Moore DR, Sohoglu E, et al. Less Is More: Latent Learning Is Maximized by Shorter Training Sessions in Auditory Perceptual Learning. PLoS One 2012;7:e36929.

26. Sass SM, Legge GE, Lee HW. Low-vision Reading Speed: Influences of Linguistic Inference and Aging. Optom Vis Sci 2006;83:166–77.

27. Pelli DG, Tillman KA, Freeman J, et al. Crowding and Eccentricity Determine Reading Rate. J Vis 2007; 7:20.1–36.

28. Arango T, Yu D, Lu ZL, et al. Effects of Task on Reading Performance Estimates. Front Psychol 2020; 11:2005.

29. Stolowy N, Calabrèse A, Sauvan L, et al. The Influence of Word Frequency on Word Reading Speed when Individuals with Macular Diseases Read Text. Vision Res 2019;155:1–10.

30. Subramanian A, Pardhan S. Repeatability of Reading Ability Indices in Subjects with Impaired Vision. Invest Ophthalmol Vis Sci 2009;50:3643–7.