Scoring Reading Parameters: An Inter-Rater Reliability Study Using The MNREAD Chart

Karthikeyan Baskaran, Antonio Filipe Macedo, Yingchen He, Laura Hernandez-Moreno, Tatiana Queirós, J Stephen Mansfield, Aurélie Calabrèse

PLOS ONE

Scoring Reading Parameters: An Inter-Rater Reliability Study Using The MNREAD Chart

--Manuscript Draft--

Manuscript Number:	
Article Type:	Research Article
Full Title:	Scoring Reading Parameters: An Inter-Rater Reliability Study Using The MNREAD Chart
Short Title:	Inter-Rater Reliability Of The MNREAD Acuity Chart
Corresponding Author:	Aurélie Calabrèse Aix-Marseille Universite Marseille, FRANCE
Keywords:	low vision; reading performance; reading test; MNREAD acuity chart; inter-rater reliability; computer-based scoring algorithms
Abstract:	Purpose: First, to evaluate inter-rater reliability when human raters estimate the reading performance of visually impaired individuals using the MNREAD acuity chart. Second, to evaluate the agreement between computer-based scoring algorithms and compare them with human rating. Methods: Reading performance was measured for 101 individuals with low vision, using the Portuguese version of MNREAD. Seven raters estimated the maximum reading speed (MRS) and critical print size (CPS) of each individual MNREAD curve. MRS and CPS were also calculated automatically for each MNREAD curve using two different algorithms: the original standard deviation method (SDev) and a non-linear mixed effects (NLME) modeling. Intra-class correlation coefficients (ICC) were used to estimate absolute agreement between raters and/or algorithms. Results: Absolute agreement between raters was excellent for MRS (ICC = 0.97; 95%CI [0.96, 0.98]) and good for CPS (ICC = 0.77; 95%CI [0.69, 0.83]). For CPS inter-rater reliability was poorer among less experienced raters (ICC = 0.70; 95%CI [0.57, 0.80]) compared to experienced ones (ICC = 0.82; 95%CI [0.57, 0.80]). Absolute agreement between the two algorithms was excellent for MRS (ICC = 0.96; 95%CI [0.91, 0.98]). For CPS, the best possible agreement was good and for CPS defined as the print size sustaining 80% of MRS (ICC = 0.77; 95%CI [0.68, 0.84]). Conclusion: For MRS, inter-rater reliability is excellent, even considering the possibility of noisy and/or incomplete data collected in low-vision individuals. For CPS, inter-rater reliability is lower, which may be problematic, for instance in the context of multicenter studies or follow-up examinations. Setting up consensual guidelines to deal with ambiguous datasets may help improve reliability. While the exact definition of CPS should be chosen on a case-by-case basis depending on the clinician or researcher's motivations, evidence suggests that estimating CPS as the smallest print size sustaining about 80% of MRS would increase inter-rater reliability.
Order of Authors:	Karthikeyan Baskaran
	Antonio Filipe Macedo
	Yingchen He
	Laura Hernandez-Moreno
	Tatiana Queirós
	J. Stephen Mansfield
	Aurélie Calabrèse
Opposed Reviewers:	
Additional Information:	
Question	**Response**
Financial Disclosure	AFM and LHM were supported by FCT (COMPETE/QREN) grant reference

Enter a financial disclosure statement that describes the sources of funding for the work included in this submission. Review the submission guidelines for detailed requirements. View published research articles from *PLOS ONE* for specific examples. This statement is required for submission and **will appear in the published article** if the submission is accepted. Please make sure it is accurate. **Unfunded studies** Enter: *The author(s) received no specific funding for this work.* **Funded studies** Enter a statement with the following details: • Initials of the authors who received each award • Grant numbers awarded to each author • The full name of each funder • URL of each funder website • Did the sponsors or funders play any role in the study design, data collection and analysis, decision to publish, or preparation of the manuscript? • **NO** - Include this sentence at the end of your statement: *The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.* • **YES** - Specify the role(s) played. * typeset	PTDC/DPT-EPI/0412/2012 in the context of the Prevalence and Costs of Visual Impairment in Portugal: a hospital based study (PCVIP-study). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.
Competing Interests Use the instructions below to enter a competing interest statement for this submission. On behalf of all authors, disclose any competing interests that could be perceived to bias this work—acknowledging all financial support and any other relevant financial or non-financial competing interests. This statement **will appear in the published article** if the submission is accepted. Please make sure it is accurate. View published research articles	I have read the journal's policy and the authors of this manuscript have the following competing interests: JSM receives royalties from the sales of MNREAD Acuity Charts

from *PLOS ONE* for specific examples. **NO authors have competing interests** Enter: *The authors have declared that no competing interests exist.* **Authors with competing interests** Enter competing interest details beginning with this statement: *I have read the journal's policy and the authors of this manuscript have the following competing interests: [insert competing interests here]* * typeset	
Ethics Statement Enter an ethics statement for this submission. This statement is required if the study involved: • Human participants • Human specimens or tissue • Vertebrate animals or cephalopods • Vertebrate embryos or tissues • Field research Write "N/A" if the submission does not require an ethics statement. General guidance is provided below. Consult the submission guidelines for detailed instructions. **Make sure that all information entered here is included in the Methods section of the manuscript.**	The study protocol was reviewed by the ethics committee for Life Sciences and Health of the University of Minho (REF: SECVS-084/2013) and was conducted in accordance with the principles of the Declaration of Helsinki. Written informed consent was obtained from all participants. The study was registered with the Portuguese data protection authority with the reference 9936/2013 and received approval number 5982/2014.

Format for specific study types **Human Subject Research (involving human participants and/or tissue)** • Give the name of the institutional review board or ethics committee that approved the study • Include the approval number and/or a statement indicating approval of this research • Indicate the form of consent obtained (written/oral) or the reason that consent was not obtained (e.g. the data were analyzed anonymously) **Animal Research (involving vertebrate animals, embryos or tissues)** • Provide the name of the Institutional Animal Care and Use Committee (IACUC) or other relevant ethics board that reviewed the study protocol, and indicate whether they approved this research or granted a formal waiver of ethical approval • Include an approval number if one was obtained • If the study involved *non-human primates*, add *additional details* about animal welfare and steps taken to ameliorate suffering • If anesthesia, euthanasia, or any kind of animal sacrifice is part of the study, include briefly which substances and/or methods were applied **Field Research** Include the following details if this study involves the collection of plant, animal, or other materials from a natural setting: • Field permit number • Name of the institution or relevant body that granted permission	
Data Availability Authors are required to make all data underlying the findings described fully available, without restriction, and from the time of publication. PLOS allows rare exceptions to address legal and ethical concerns. See the PLOS Data Policy and FAQ for detailed information.	Yes - all data are fully available without restriction

A Data Availability Statement describing where the data can be found is required at submission. Your answers to this question constitute the Data Availability Statement and **will be published in the article**, if accepted. **Important:** Stating 'data available on request from the author' is not sufficient. If your data are only available upon request, select 'No' for the first question and explain your exceptional situation in the text box. Do the authors confirm that all data underlying the findings described in their manuscript are fully available without restriction?	
Describe where the data may be found in full sentences. If you are copying our sample text, replace any instances of XXX with the appropriate details. • If the data are **held or will be held in a public repository**, include URLs, accession numbers or DOIs. If this information will only be available after acceptance, indicate this by ticking the box below. For example: *All XXX files are available from the XXX database (accession number(s) XXX, XXX.).* • If the data are all contained **within the manuscript and/or Supporting Information files**, enter the following: *All relevant data are within the manuscript and its Supporting Information files.* • If neither of these applies but you are able to provide **details of access elsewhere**, with or without limitations, please do so. For example: *Data cannot be shared publicly because of [XXX]. Data are available from the XXX Institutional Data Access / Ethics Committee (contact via XXX) for researchers who meet the criteria for access to confidential data.* *The data underlying the results presented in the study are available from (include the name of the third party*	All XXX files are available from the XXX database (accession number(s) XXX, XXX.)

and contact information or URL). • This text is appropriate if the data are owned by a third party and authors do not have permission to share the data. * typeset	
Additional data availability information:	Tick here if the URLs/accession numbers/DOIs will be available only after acceptance of the manuscript for publication so that we can ensure their inclusion before publication.

Aurelie Calabrese, PhD

Post-doctoral Associate

Aix-Marseille University - Centre St Charles - Pôle 3C
3 place Victor Hugo - case D - 13331 Marseille cedex 3

aurelie.calabrese@univ-amu.fr

Date: January 31st 2019

Dear members of the Editorial Board,

Please find enclosed our manuscript entitled: "**Scoring Reading Parameters: An Inter-Rater Reliability Study Using The MNREAD Chart**", by Karthikeyan Baskaran, Antonio Filipe Macedo, Yingchen He, Laura Hernandez-Moreno, Tatiana Queirós, J. Stephen Mansfield, and Aurélie Calabrèse, which we would like to submit for publication as an research article in PLoS ONE.

The primary goal of this work is to evaluate inter-rater reliability when human raters estimate reading performance using the MNREAD acuity chart. Our motivation for this study was the lack of evidence that different extraction methods used by different raters would lead to comparable estimates of reading performance, which is especially relevant in the context of multicenter studies, or when looking at follow-up data. Our results demonstrate excellent inter-rater reliability for the Maximum Reading Speed (i.e. the fastest that one can read when print size is not limiting) and good inter-rater reliability for the Critical Print Size (i.e. the print size for which reading speed is maximum). Our work also provides further tips and instructions on how to score noisy and/or incomplete MNREAD data. These tips may serve as a starting point to help clinicians and researchers reduce variability.

We confirm that this manuscript has not been published elsewhere and is not under consideration by another journal. All Authors have approved the manuscript and agree with submission to PLoS ONE.

Authors report no conflict of interest, except for JSM who receives royalties from the sales of MNREAD Acuity Charts.

We appreciate your consideration of publication of this paper.

Sincerely,
Aurélie Calabrèse, PhD

Full title: Scoring Reading Parameters: An Inter-Rater Reliability Study Using The MNREAD Chart.

Short title: Inter-Rater Reliability Of The MNREAD Acuity Chart.

Authors: Karthikeyan Baskaran[1], Antonio Filipe Macedo[1, 2], Yingchen He[3], Laura Hernandez-Moreno[2], Tatiana Queirós[4], J. Stephen Mansfield[5], Aurélie Calabrèse[6,7]

Affiliations:

[1] Department of Medicine and Optometry, Linnaeus University, Kalmar, Sweden

[2] Low Vision and Visual Rehabilitation Lab, Department and Center of Physics—Optometry and Vision Science, University of Minho Braga, Braga, Portugal

[3] Department of Ophthalmology & Visual Neurosciences, University of Minnesota, Twin Cities, United States

[4] Serviço de Oftalmologia, Hospital de Braga, Braga, Portugal.

[5] Department of Psychology, SUNY College at Plattsburgh, Plattsburgh, New York, United States

[6] Aix-Marseille University, Marseille, France

[7] Laboratoire de Psychologie Cognitive, CNRS, Marseille, France

Corresponding author:

Aurélie Calabrèse (AC)

aurelie.calabrese@univ-amu.fr

3 place Victor Hugo - 13003 Marseille, France

Word count:

Abstract = 300 words

Text = 3740 words

Intro = 1025; Methods = 1010; Results = 580; Discussion = 1011; Conclusion = 114

References = 30; Figures = 5; Tables = 4

Funding:

This work was partially supported by FCT (COMPETE/QREN).

Declarations of interest:

Author JSM receives royalties from the sales of MNREAD Acuity Charts.

Abstract

Purpose: First, to evaluate inter-rater reliability when human raters estimate the reading performance of visually impaired individuals using the MNREAD acuity chart. Second, to evaluate the agreement between computer-based scoring algorithms and compare them with human rating.

Methods: Reading performance was measured for 101 individuals with low vision, using the Portuguese version of MNREAD. Seven raters estimated the maximum reading speed (MRS) and critical print size (CPS) of each individual MNREAD curve. MRS and CPS were also calculated automatically for each MNREAD curve using two different algorithms: the original standard deviation method (SDev) and a non-linear mixed effects (NLME) modeling. Intra-class correlation coefficients (ICC) were used to estimate absolute agreement between raters and/or algorithms.

Results: Absolute agreement between raters was excellent for MRS (ICC = 0.97; 95%CI [0.96, 0.98]) and good for CPS (ICC = 0.77; 95%CI [0.69, 0.83]). For CPS inter-rater reliability was poorer among less experienced raters (ICC = 0.70; 95%CI [0.57, 0.80]) compared to experienced ones (ICC = 0.82; 95%CI [0.57, 0.80]). Absolute agreement between the two algorithms was excellent for MRS (ICC = 0.96; 95%CI [0.91, 0.98]). For CPS, the best possible agreement was good and for CPS defined as the print size sustaining 80% of MRS (ICC = 0.77; 95%CI [0.68, 0.84]).

Conclusion: For MRS, inter-rater reliability is excellent, even considering the possibility of noisy and/or incomplete data collected in low-vision individuals. For CPS, inter-rater reliability is lower, which may be problematic, for instance in the context of multicenter studies or follow-up

examinations. Setting up consensual guidelines to deal with ambiguous datasets may help improve reliability. While the exact definition of CPS should be chosen on a case-by-case basis depending on the clinician or researcher’s motivations, evidence suggests that estimating CPS as the smallest print size sustaining about 80% of MRS would increase inter-rater reliability.

Introduction

Reading difficulty is a major concern for patients referred to low-vision centers [1]. Therefore, most Quality-of-Life questionnaires assessing the severity of vision disability contain one or more items on subjective reading difficulty [2-5]. However, substantial discrepancy has been observed between self-reported reading difficulty and measured reading speed [6]. For this reason, reading performance should be evaluated objectively to serve as a reliable outcome measure in clinical trials, multisite investigations or longitudinal studies. To assess, for instance, the success of vision rehabilitation techniques, surgical procedures or ophthalmic treatments, measures of reading ability should be obtained using standardized tests with demonstrated high repeatability.

Among the standardized tests available, the MNREAD acuity chart can be used to evaluate reading performance for people with normal vision or low vision in clinical and research environments [7]. In brief, the MNREAD chart measures four parameters that characterize how reading performance changes when print size decreases: the maximum reading speed (MRS), the critical print size (CPS), the reading acuity (RA) and the reading accessibility index (ACC) [8]. The reading acuity and reading accessibility index are clearly defined by the number of reading errors made at small print sizes and the reading speeds for a range of larger sizes. In the original MNREAD manual, provided with the chart, MRS and CPS are defined as follows: “The critical print size is the smallest print size at which patients can read with their maximum reading speed. […] Typically, reading time remains fairly constant for large print sizes. But as the acuity limit is approached there comes a print size where reading starts to slow down. This is the critical print size. The maximum reading speed with print larger than the critical print size is the maximum reading speed (MRS).” In short, values for MRS and CPS depend on the location of the flexion point in the curve of reading speed versus print size (Fig 1). In normally sighted individuals, for whom the MNREAD curve usually exhibits a standard shape (Fig 1-A), the above definitions may be sufficient to extract MRS and CPS confidently by inspecting the curve. However, they can be difficult to determine, especially for readers with visual impairments, who may experience visual field defects (e.g. ring scotoma; Fig 1-B) or the use of multiple fixation sites (i.e. PRL; Fig 1-C) [9]. In such cases, the noisy and/or incomplete dataset resulting from atypical visual function may be inconsistent with the assumption that people will read at a fairly constant speed until font size compromises their ability to identify words and MNREAD curves may take an unusual shape (Fig 1-D). If so, subjective decisions (e.g. ignoring outliers) must be made by the individual analysing the data (referred to as the "rater" in the present work, as opposed to the "experimenter" who recorded the data). For this reason, MRS and CPS estimates may be considered highly sensitive to inter-rater variability.

Fig 1: MNREAD curve examples.

In an attempt to reduce variability and unify the process of curve information extraction, alternative scoring methods have been proposed. According to these "simpler" scoring rules, MRS equals either the single largest reading speed [10] or the mean of the three largest reading speeds [11]. Nonetheless, a criterion must be chosen for the CPS (smallest print size supporting reading speed at either: 90% of MRS, 85%, 80%, etc.) but there is no general agreement on the appropriate criterion to use. Overall, open discussions on how to score MNREAD parameters optimally still persist in the literature [12]and the choice of scoring method constitutes an additional factor contributing to inter-rater variability.

Another approach to reduce variability is to fit the MNREAD curve and estimate its parameters using automated algorithms [13]. In the present work, we will focus on two of these methods. The first one has been described by the MNREAD creators [14,15] and is used in the MNREAD iPad app [16]. It is also the most widely used in the literature [11,17,18]. In short, it determines the CPS as the smallest print size that supports reading speeds that are not significantly different from the reader's maximum reading speed; we will refer to it as the standard deviation method (SDev). The second method, especially recommended with large but incomplete datasets, estimates the critical print size from smooth curve-fit to the MNREAD data using non-linear mixed effects (NLME) modeling [19]; we will refer to it as the NLME method. Both methods are described in the Methods section. Despite the advantage of these algorithms in operationalizing the estimation of the MNREAD parameters, they present two major drawbacks: (1) they may not be easily accessible in clinical environments, (2) they may fail to provide satisfactory measures with noisy or small and incomplete datasets, necessitating further human inspection of the curves for validation.

The Repeatability of the MNREAD chart measures has been assessed before in low vision populations. Overall, studies have reported good intra and inter-session reliability [11,17,18,20], as well as good repeatability across multiple testing sites and experimenters [21]. But to our knowledge, variability of the MNREAD estimates scored by different raters from the same dataset has not been evaluated. This question of inter-rater variability is especially relevant (1) in the context of multicenter studies, where data are scored by different raters with different levels of expertise, (2) when comparing results from different studies performed by different groups, or (3) when looking at follow-up data involving different raters.

We have investigated the reliability of CPS and MRS estimates for MNREAD data collected from participants with visual impairments. First, we evaluated the inter-rater reliability among raters (Analysis 1). Second, we evaluate agreement between the NLME and SDev algorithms (Analysis 2). Third, we evaluated agreement between raters and the two algorithms (Analysis 3).

Methods

Participants

Data from 101 participants with visual impairment were selected from a larger dataset, originally collected to study the prevalence and costs of visual impairment in Portugal (PCVIP-study) [22,23]. Only participants whose visual acuity in the better eye was 0.5 decimal (0.3 logMAR) or worse and/or whose visual field was less than 20 degrees were selected for the present study. Among them, only the participants who read at least five sentences on the MNREAD chart with their "presenting reading glasses" were included. The study protocol was reviewed by the ethics committee for Life Sciences and Health of the University of Minho (REF: SECVS-084/2013) and was conducted in accordance with the principles of the Declaration of Helsinki. Written informed consent was obtained from all participants. The study was registered with the Portuguese data protection authority with the reference 9936/2013 and received approval number 5982/2014.

MNREAD Data

Reading performance was measured for each participant using the Portuguese version of the MNREAD acuity chart [24]. Reading distance was adjusted for each participant and chosen according to his/her near visual acuity. Participants were asked to read the chart aloud as fast and accurately as possible, one sentence at a time, starting from the largest print size. For each sentence, reading time and number of misread words were recorded and reported on a score sheet by the experimenter. Data were then transferred into a digital file and further processed in R [25]. For each individual test, a corresponding MNREAD curve was plotted using the mnreadR package [26] to display log reading speed as a function of print size (see S1 Appendix for all 101 curves). Because the shape of the curve can influence visual estimation of the reading parameters, reading speed was plotted using a logarithmic scale so that reading speed variability (which is proportional to the overall measure of reading speed) was constant at all speeds [14].

Raters' visual scoring

Seven raters were recruited to estimate the MRS and CPS of each individual MNREAD curve. Since inter-rater reliability may be influenced by raters' prior experience with the MNREAD chart, we included raters with different levels of expertise in MNREAD parameters estimation. Each rater gave a self-rated score of expertise (on a 5 point scale from 0 = 'no previous to experience' to 4 = 'top expertise'), both before and after rating all the MNREAD curves, to account for the amount of practice gained during the study. Each rater was provided with S1 Appendix, containing the 101 MNREAD curves to score. Raters were instructed to follow the standard guidelines provided with the MNREAD chart instructions (see Introduction). However, coming from patients with impaired vision, many of the curves had noisy or incomplete data, which potentially made it difficult to estimate the MRS and CPS. In such cases, we provided more detailed instructions to the raters. These detailed instructions are available in S2 Appendix.

Algorithms' automated scoring

MRS and CPS were also calculated automatically for each 101 datasets using two algorithm-based estimations: the 'standard deviation' method and non-linear mixed effects modeling. The standard deviation method (SDev) uses the original algorithm described in [14] and [15] to estimate the MNREAD parameters. This algorithm iterates over the data searching for an optimal reading speed plateau, from which MRS and CPS will be derived. To be considered optimal, a plateau must encompass a range of print sizes that supports reading speed at a significantly faster rate (1.96 × standard deviation) than the print sizes smaller or larger than the plateau range (Fig 2). MRS is estimated as the mean reading speed for print sizes included in the plateau and CPS is defined as the smallest print size on the plateau. In most cases, several print-size ranges can qualify as an optimal plateau and the algorithm chooses the one with the fastest average reading speed. In the present work, the standard deviation method estimation was performed using the curveParam_RT () function from the mnreadR R package.

Fig 2: Example of the standard deviation algorithm calculation on a typical dataset.

On *iteration 1* (dark blue), the algorithm selects the first two sentences as *plateau 1* (1.3 and 1.2 logMAR) and calculates a selection criterion for this plateau. Criterion $_{\text{plateau 1}}$ = mean (reading speed $_{\text{plateau 1}}$) – 1.96 x standard deviation (reading speed $_{\text{plateau 1}}$) = 60.5 - 1.96 × 2.1 = 56.3 wpm. The point adjacent to *plateau 1* (1.1 logMAR) was read at 60 wpm, which is faster than criterion $_{\text{plateau 1}}$, indicating that this point belongs to the optimal plateau. A second iteration is then launched (light blue) with *plateau 2* now encompassing the first three sentences and a new criterion calculation. Criterion $_{\text{plateau 2}}$ = 60.3 - 1.96 × 1.5 = 57.3 wpm. Among the points adjacent to *plateau 2,* there is still a value higher than this criterion (59 wpm at 0.9 logMAR), so the algorithm continues to iterate one sentence at a time, including 1.0 logMAR in *plateau 3* and 0.9 logMAR in *plateau 4*. The calculations stop with *plateau 4,* for which selection criterion is higher than any remaining points (criterion $_{\text{plateau 4}}$ = 44.7 wpm). MRS is estimated as 57.2 wpm and CPS as 0.9 logMAR.

The non-linear mixed effects (NLME) modeling method is particularly suited for incomplete datasets from individuals with reading or visual impairment [19]. The NLME model uses parameter estimates from a larger group (101 datasets here) to allow suitable curve fits for individual datasets that contain few data points. In the present work, we used an NLME model with a negative exponential decay function, as described in details in [19], where a single estimate of MRS can yield several measures of CPS depending on the definition chosen (e.g. print size required to achieve 90% of MRS, 80% of MRS, etc.). Therefore, five values of CPS were estimated, i.e. 95%, 90%, 85%, 80% and 75% of MRS. NLME modeling and parameters estimation were performed using the nlmeModel () and nlmeParam () functions from mnreadR.

Statistical Analysis

In all three analyses, intra-class correlation coefficient (ICC) was used to assess absolute agreement between raters and/or algorithms [27]. This reliability index (ranging from 0 to 1; 1 meaning perfect agreement) is widely used in the literature in test-retest, intra-rater, and inter-rater reliability analyses [28]. In the present work, ICC values estimate the variation between two or more methods (whether raters or algorithms) in scoring the same data by calculating the absolute agreement between them. For each analysis, the appropriate ICC form (dependent on research design and assumptions) was chosen by selecting the correct combination of “model”, “type” and “definition”, as detailed in Table 1 [29]. ICC values were calculated using SPSS statistical package and limits of agreement were visualized with Bland-Altman plots. Following guidelines from [28], ICC values and their 95% confidence intervals (95% CI) were interpreted as showing: *“poor agreement”* if less than 0.5; *“moderate agreement”* if comprised between 0.5

and 0.75; *"good agreement"* if comprised between 0.75 and 0.9 and *"excellent agreement"* if greater than 0.9.

Table 1: Details of the ICC form chosen for Analyses 1, 2 and 3

	Intra-class correlation coefficient (ICC) form		
	Model	Type	Definition
Analysis 1 Agreement among the 7 raters	2-way random effects Both raters & curves are considered as selected randomly from a larger population	Single rater Each rater is compared against all others	Absolute agreement
Analysis 2 Agreement between the 2 automated algorithms	2-way mixed-effects Raters are fixed & curves are considered as selected randomly from a larger population	Single measurement	Absolute agreement
Analysis 3 Agreement between raters and automated algorithms	2-way mixed effects	Mean of *7* raters	Absolute agreement

Results

Analysis 1: Agreement between raters (221 words)

For MRS, ICC value was 0.97 (95% CI [0.96, 0.98]), indicating excellent agreement between raters (Fig 3). For CPS, ICC value was 0.77 (95% CI [0.69, 0.83]), suggesting good agreement between raters. We hypothesized that the weaker agreement for CPS could be attributed to the difference in raters’ expertise level. These scores, both before and after evaluating the 101 MNREAD curves, are reported in Table 2. Prior to rating, one rater had no previous experience in rating MNREAD curves (TQ), three raters considered themselves intermediate raters (LM, AM and KB), two raters scored themselves as advanced raters (SM and YH) and one rater reported to be an expert rater (AC). Among the less experienced raters (score 0-2), CPS estimation reliability was only moderate (ICC = 0.70; 95% CI [0.57, 0.80]). Among the most experienced raters (score 3-4), it was good (ICC = 0.82; 95% CI [0.57, 0.80]). Interestingly, three raters (43%) considered that their expertise improved (TQ, LM and AM), whereas the remaining four (57%) did not report any change in their expertise level (KB, SM, YH and AC).

Table 2: Self-reported score of expertise for our 7 raters

Raters		TQ	LM	AM	KB	SM	YH	AC
Self-reported score of expertise	Prior rating	0	2	2	2	3	3	4
	After rating	1	3	3	2	3	3	4

Score of expertise in rating low-vision MNREAD data before and after rating the 101 curves (0 – no prior experience, 1 – novice, 2 – intermediate, 3 – Advance, 4 – Expert).

Fig 3: Box and whisker plots of estimated MRS (left) and CPS (right), grouped by raters and sorted in ascending order of expertise level (from 0 to 4). Boxes represent the 25th to 75th percentiles and whiskers range from min to max values. Medians (lines) and means (cross) are also represented.

Analysis 2: Agreement between automated algorithms (245 words)

For MRS, the ICC value of absolute agreement between SDev and NLME methods was 0.96 (95% CI [0.91, 0.98]), showing excellent agreement. Contrary to the SDdev method, for which a single MNREAD test yields only one estimate for MRS and one estimate for CPS, the NLME method can generate several measures of CPS depending on the reading-speed criterion chosen to define the CPS (e.g. print size required to achieve 90% of MRS, 80% of MRS, etc.). Therefore, for each of the 101 MNREAD datasets, we estimated five values of CPS with NLME (corresponding to: 95%, 90%, 85%, 80% and 75% of MRS) and measured agreement between SDev and NLME for each of them. The results are reported in Table 3. The strongest agreement between the two automated methods was found for the 80% criterion, and was good, with an ICC value of 0.77 (95% CI [0.68, 0.84]). Additionally, limits of agreement between the two algorithms were estimated using Bland – Altman plots for both MRS and CPS (Fig 4). For MRS, the average difference (i.e. bias) between the SDev method and the NLME model was 5.8 wpm (i.e. 4.5%), with 95% limits of agreement of 11.4 wpm (i.e. 10%). For CPS (defined as 80% of MRS, which showed the best agreement between methods), bias was 0.031 logMAR with 95% limits of agreement of 0.06 logMAR (1 step unit being 0.1 logMAR). Overall, we concluded that no significant difference could be observed between the two automated algorithms.

Table 3: Absolute agreement (ICC values and their 95 % confidence intervals) between CPS values estimated with the SDev method and the NLME model for five different definitions of CPS.

	ICC value	95% CI	Absolute agreement
95% CPS	0.56	[0.10, 0.77]	Moderate
90% CPS	0.70	[0.53, 0.81]	
85% CPS	0.76	[0.66, 0.83]	Good
80% CPS	**0.77**	**[0.68, 0.84]**	
75% CPS	0.76	[0.62, 0.84]	

Best agreement is highlighted in grey.

Fig 4: Bland – Altman plots showing agreement between SDev method and NLME model for both MRS (left) and CPS (right). x-axes represent the mean estimate for both methods; y-axes represent the estimate difference between SDev method and NLME model. Dashed lines show the mean difference (i.e. bias) and the dotted lines represent the 95% CI of limits of agreement (i.e. confidence limits of the bias, defined as the mean difference ± 1.96 times the standard deviation of the difference).

Analysis 3: Agreement between raters and automated algorithms (139 words)

For MRS, absolute agreement between raters (k = 7) and automated algorithms was found to be excellent for both the SDev method (ICC = 0.96; 95% CI [0.88, 0.98] and the NLME model (ICC = 0.97; 95% CI [0.95, 0.98]). For CPS, agreement between raters and the SDev method was only moderate (ICC = 0.66; 95% CI [0.3, 0.80]), whereas agreement between raters and the NLME model was ‘good’ for CPS defined as 90% of MRS (ICC = 0.83; 95% CI [0.76, 0.88] - Table 4 shows the ICC values for each of the five CPS definitions). Overall, the NLME model showed better agreement with the raters than the SDev method for both reading parameters. Fig 5 shows the MRS and CPS obtained by the automated algorithms and the 7 raters.

Fig 5: Box and whisker plots showing the median and average MRS (left panel) and CPS (right panel) from the two algorithms and the mean of raters. The box represents 25th to 75th percentile with median line and the + sign represents the mean and the whiskers represent minimum to maximum.

Table 4: Absolute agreement (ICC values and their 95 % confidence intervals) between CPS values estimated by the raters and with the NLME model for five different definitions of CPS.

	ICC value	95% CI	Absolute agreement
95% CPS	0.78	[0.61, 0.87]	
90% CPS	**0.83**	**[0.76, 0.88]**	Good
85% CPS	0.79	[0.55, 0.71]	
80% CPS	0.72	[0.18, 0.88]	Moderate
75% CPS	0.66	[0.02, 0.87]	

Best agreement is highlighted in grey.

Discussion (1001 words)

In this project we investigated *i)* the agreement between raters for MNREAD parameters extracted from reading curves (Analysis 1), *ii)* the agreement between SDev and NLME automated methods extracting reading parameters from raw data (Analysis 2) and *iii)* the agreement between raters and automated methods (Analysis 3).

Our first main result was that inter-rater reliability can be classified as excellent for MRS (ICC of 0.97) and good for CPS (ICC of 0.77). Because they are lower than 1, these agreement indexes reveal the existence of discrepancies when extracting MNREAD parameters visually from reading curves. Whilst the variability for MRS can be considered residual, the CPS estimation may be questionable. On average, the range of difference in CPS estimates was 0.19 logMAR (i.e. almost 2 lines on a logMAR chart), implying that the variability among raters can be considered clinically significant and potentially problematic, for example when CPS is used to prescribe optimal magnifying power. To identify the underlying factors of the discrepancies observed in CPS rating, we considered whether the data itself could be involved, hypothesizing that the modest ICC value that we found (0.77) was largely due to the presence of highly noisy data. To confirm this hypothesis, we identified extreme outliers for which CPS values were three times larger than the standard deviation of the mean. A total of five curves (5%) were identified as extreme outliers (#2, #31, #58, #70 and #89 in S1 Appendix). What these curves have in common is: the lack of a clear plateau and/or the lack of a clear drop point. After removing these five outliers, the resulting ICC value for CPS improved to 0.82 (95%CI [0.76, 0.87]. This increased value suggests that, to increase inter-rater reliability, ambiguous cases of noisy data should be discussed before final estimates of CPS are reached. Therefore, the advice for our fellow researchers is to inspect our 5 ambiguous samples and define how to deal with such cases on an individual basis whilst maintaining consistency in data extraction. The tips provided in S2 Appendix on how to score ambiguous data can serve as a starting point. When possible, measurements should be repeated to help interpret problematic data.

We also found that for CPS inter-rater reliability was poorer among less experienced raters compared to experienced ones. We speculate that this tendency may be related to both the lack of experience in administrating and rating the test that would lead more naïve raters to follow strictly the definitions of CPS and MRS. Taking the example of curve #2 (see S1 Appendix), raters SM and AC (self-reported expertise scores of 3 and 4) estimated CPS to be 0.7 logMAR (MRS = 68 wpm, both) whilst TQ and KB (self-reported expertise score of 0 and 2) estimated CPS to be 1.3 and 1.1 logMAR (MRS = 85 and 75 wpm, respectively). In this case, the more experienced raters (SM and AC) may have decided to ignore the outlier initial data point, assuming that this measure resulted from experimental noise.

Our second main result is the excellent agreement between the two automated methods for MRS. Regarding CPS estimation, the NLME method provides more flexibility over the SDev method, since it allows to determine CPS for different levels of MRS. For instance a higher, more conservative criterion, can be chosen for fluent reading while a lower criterion would be preferred for spot reading. However, there is no rule yet on how to set this criterion optimally to increase reliability. Our results show that the reading speed cut-off to determine CPS yielding the best reliability between methods is 80% MRS. This result resonates with conclusions from [19], who showed that agreement between NLME models using a two-limb function and an exponential decay function was greater if CPS was set at 80% MRS. On the question of test-retest reliability, [11] also reported that using a criterion of 80% yield improved repeatability of the CPS (when compared to 90%). While an optimal criterion should be chosen on a case-by-case basis depending on the clinician or researcher’s motivations, all these evidence suggest that a criterion close to 80 % would increase both inter-rater and test-retest variability.

Our third result is that raters and automated methods show excellent agreement for MRS values (ICC of 0.96 and 0.97 for the SDev and NLME respectively). The agreement for CPS was more variable. It was found to be poor for the SDev (ICC of 0.66) and good for the NLME (ICC of 0.83 with a CPS criterion set to 90% MRS). It is worth noting that ICC values were almost identical when measuring agreement between raters and agreement between algorithms for both MRS and CPS. This observation is quite interesting and somehow indicates the robustness and efficacy of human visual inspection of MNREAD curves.

The represent work presents some limitations. First, despite the relatively large sample of MNREAD data considered in the present work, it is hard to predict to what extent the different shaped curves are representative of the curves found in typical clinical practice. Second, it is likely that the new instructions helped reduce inter-rater variability, but there are no data to support this assumption. While all raters used these extended instructions, the ICC value for CPS was still low, suggesting that additional fixes should be considered to help increase reliability. It is possible to run participants through the test more than once, at least with the English version [16,30]. Repeated measures would make it easier for the rater to determine whether a measure should be considered as noise or not. Another possibility might be to pool estimates from multiple raters or in combination with curve fits. Third, the finding that 80% MRS yields the most reliable CPS using the NLME method is convenient to parameterize the curve in research studies using curve fitting. But for low vision rehabilitation the goal ought to be to enlarge text so that it can be read at the reader’s MRS, not at the 80% of the reader’s MRS.

Conclusions

In summary, our study shows that extraction of the maximum reading speed from MNREAD data is highly consistent across methods and researchers. It also reveals that for low-vision data, it is difficult to obtain excellent inter-rater reliability for CPS estimates. Future studies, such as rehabilitation interventions aiming at improving reading ability in people with low vision, can now follow the advices and instructions resulting from our investigation. Using a standard set of instructions and criteria to analyze reading curves may help increase the reliability of the results. Additional ways to improve inter-rater reliability should also be considered, e.g. use the curve fits, collect multiple runs per participant or combine the estimates of multiple raters.

Acknowledgments

References

1. Brown JC, Goldstein JE, Chan TL, Massof R, Ramulu P, Low Vision Research Network Study Group (2014) Characterizing functional complaints in patients seeking outpatient low-vision services in the United States. Ophthalmology 121: 1655-62.e1.

2. Frost NA, Sparrow JM, Durant JS, Donovan JL, Peters TJ, Brookes ST (1998) Development of a questionnaire for measurement of vision-related quality of life. Ophthalmic Epidemiol 5: 185-210.

3. Hart PM, Chakravarthy U, Stevenson MR, Jamison JQ (1999) A vision specific functional index for use in patients with age related macular degeneration. Br J Ophthalmol 83: 1115-20.

4. Mangione CM, Lee PP, Gutierrez PR, Spritzer K, Berry S, Hays R (2001) Development of the 25-list-item national eye institute visual function questionnaire. Archives of Ophthalmology 119: 1050-1058.

5. Massof RW, Hsu CT, Baker FH, Barnett GD, Park WL, Deremeik JT, et al. (2005) Visual Disability Variables. I: The Importance and Difficulty of Activity Goals for a Sample of Low-Vision Patients. Archives of Physical Medicine and Rehabilitation 86: 946 - 953.

6. Friedman SM, Munoz B, Rubin GS, West SK, Bandeen-Roche K, Fried LP (1999) Characteristics of discrepancies between self-reported visual function and measured reading speed. Salisbury Eye Evaluation Project Team. Invest Ophthalmol Vis Sci 40: 858-64.

7. Mansfield JS, Ahn SJ, Legge GE, Luebker A (1993) A new reading-acuity chart for normal and low vision. Ophthalmic and Visual Optics/Noninvasive Assessment of the Visual System Technical Digest, (Optical Society of America, Washington, DC., 1993.) 3: 232--235.

8. Calabrèse A, Owsley C, McGwin G, Legge GE (2016a) Development of a Reading Accessibility Index Using the MNREAD Acuity Chart. JAMA Ophthalmol 134: 398-405.

9. Macedo AF, Nascimento SMC, Gomes AOS, Puga AT (2007) Fixation in patients with juvenile macular disease. Optom Vis Sci 84: 852-8.

10. Finger RP, Charbel Issa P, Fimmers R, Holz FG, Rubin GS, Scholl HPN (2009) Reading performance is reduced by parafoveal scotomas in patients with macular telangiectasia type 2. Invest Ophthalmol Vis Sci 50: 1366-70.

11. Patel PJ, Chen FK, Da Cruz L, Rubin GS, Tufail A (2011) Test-retest variability of reading performance metrics using MNREAD in patients with age-related macular degeneration. Invest Ophthalmol Vis Sci 52: 3854-9.

12. Rubin GS (2013) Measuring reading performance. Vision Res 90: 43-51.

13. Cudeck R, Harring R Jeffrey (2010) Developing a random coefficient model for nonlinear repeated measures data. In S.-M. Chow, E. Ferrer, & F. Hsieh (Eds.). The Notre Dame series on quantitative methodology. Statistical methods for modeling human dynamics: An interdisciplinary dialogue (pp. 289-318). New York, NY, US: Routledge/Taylor & Francis Group..

14. Legge GE (2007) Psychophysics of reading in normal and low vision. Mahwah: NJ & London: Lawrence Erlbaum Associates.

15. Mansfield JS, Legge GE, Bane MC (1996) Psychophysics of reading XV - Font effects in normal and low vision. Invest Ophthalmol Vis Sci 37: 1492-501.

16. Calabrèse A, To L, He Y, Berkholtz E, Rafian P, Legge GE (2018a) Comparing performance on the MNREAD iPad application with the MNREAD acuity chart. J Vis 18: 8.

17. Virgili G, Cordaro C, Bigoni A, Crovato S, Cecchini P, Menchini U (2004) Reading acuity in children: evaluation and reliability using MNREAD charts. Invest Ophthalmol Vis Sci 45: 3349-54.

18. Subramanian A, Pardhan S (2009) Repeatability of reading ability indices in subjects with impaired vision. Invest Ophthalmol Vis Sci 50: 3643-7.

19. Cheung SH, Kallie CS, Legge GE, Cheong AM (2008) Nonlinear Mixed-Effects Modeling of MNREAD Data. Invest Ophthalmol Vis Sci 49: 828-35.

20. Subramanian A, Pardhan S (2006) The repeatability of MNREAD acuity charts and variability at different test distances. Optom Vis Sci 83: 572-6.

21. Calabrèse A, Cheong AMY, Cheung S, He Y, Kwon M, Mansfield JS, et al. (2016c) Baseline MNREAD Measures for Normally Sighted Subjects From Childhood to Old Age. Invest Ophthalmol Vis Sci 57: 3836-43.

22. Macedo AF, Ramos PL, Hernandez-Moreno L, Cima J, Baptista AMG, Marques AP, et al. (2017) Visual and health outcomes, measured with the activity inventory and the EQ-5D, in visual impairment. Acta Ophthalmol 95: e783-e791.

23. Ramos PL, Santana R, Moreno LH, Marques AP, Freitas C, Rocha-Sousa A, et al. (2018) Predicting participation of people with impaired vision in epidemiological studies. BMC Ophthalmol 18: 236.

24. Tamaki Monteiro de Castro C, Kallie CS, Salomão SR (2005) [Development and validation of the MNREAD reading acuity chart in Portuguese]. Arq Bras Oftalmol 68: 777-83.

25. R Core Team (2018) R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. , https://www.R-project.org/.

26. Calabrèse A, Mansfield JS, Legge GE. (2017) mnreadR, an R package to analyze MNREAD data.

27. Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. Psychol Bull 86: 420-8.

28. Koo TK, Li MY (2016) A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. J Chiropr Med 15: 155-63.

29. McGraw KO, Wong SP (1996) Forming inferences about some intraclass correlation coefficients.. Psychological Methods 1: 30-46.

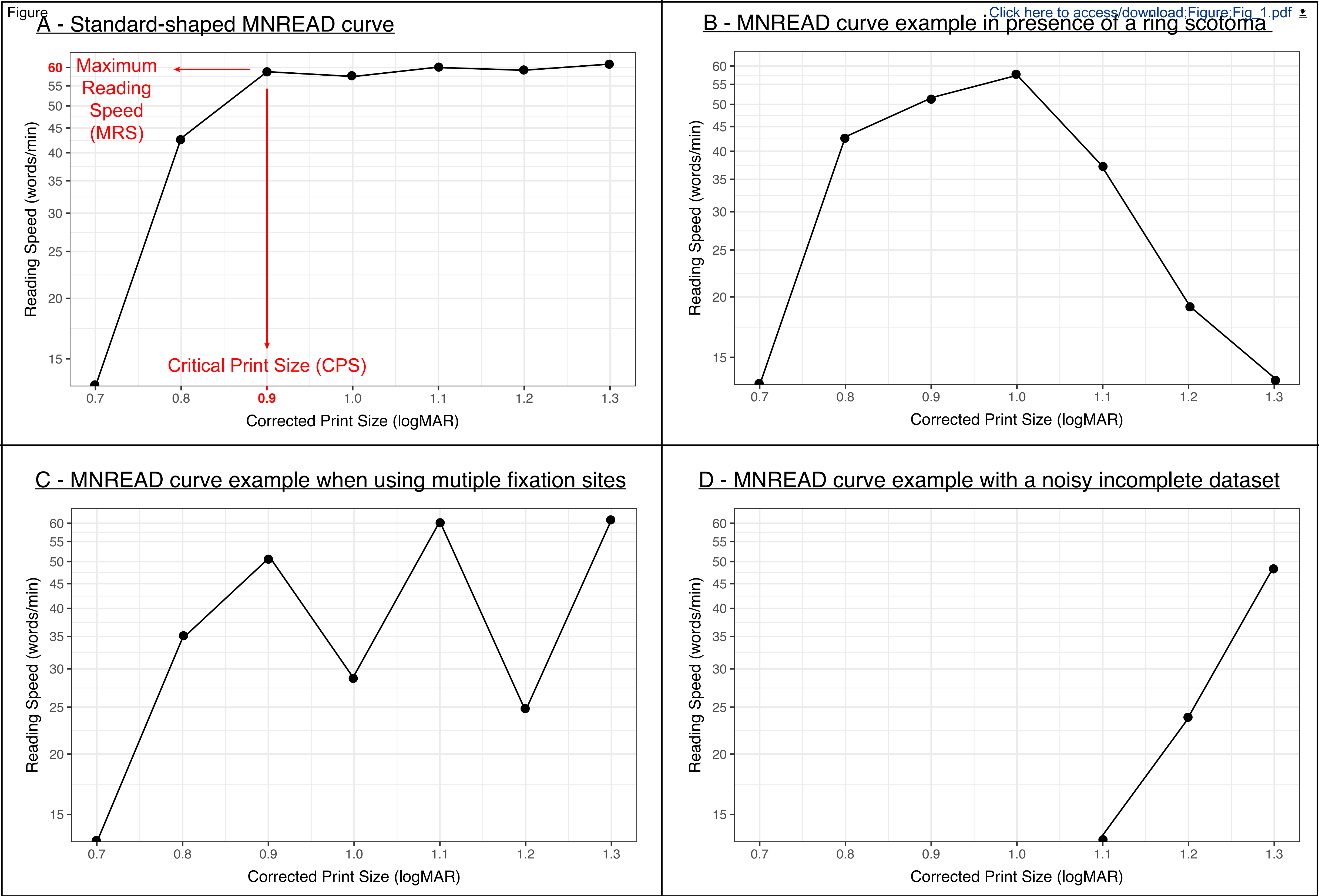
30. Mansfield JS, Atilgan N, Lewis A, Legge GE (in press) Extending the MNREAD sentence corpus: Computer-generated sentences for measuring vision for reading. Vision Research.

Supporting information captions

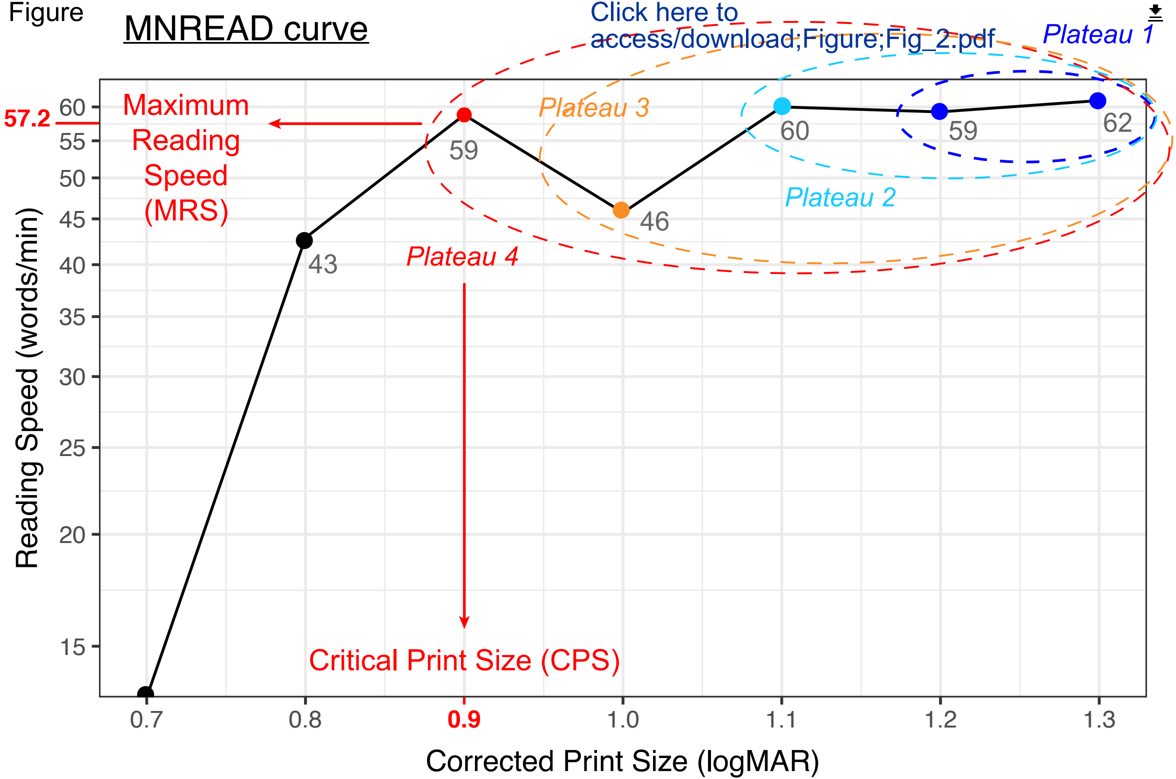
S1 Appendix. Individual MNREAD curves from the 101 MNREAD measurements.

S2 Appendix. Detailed scoring instructions provided to the raters.

Figure

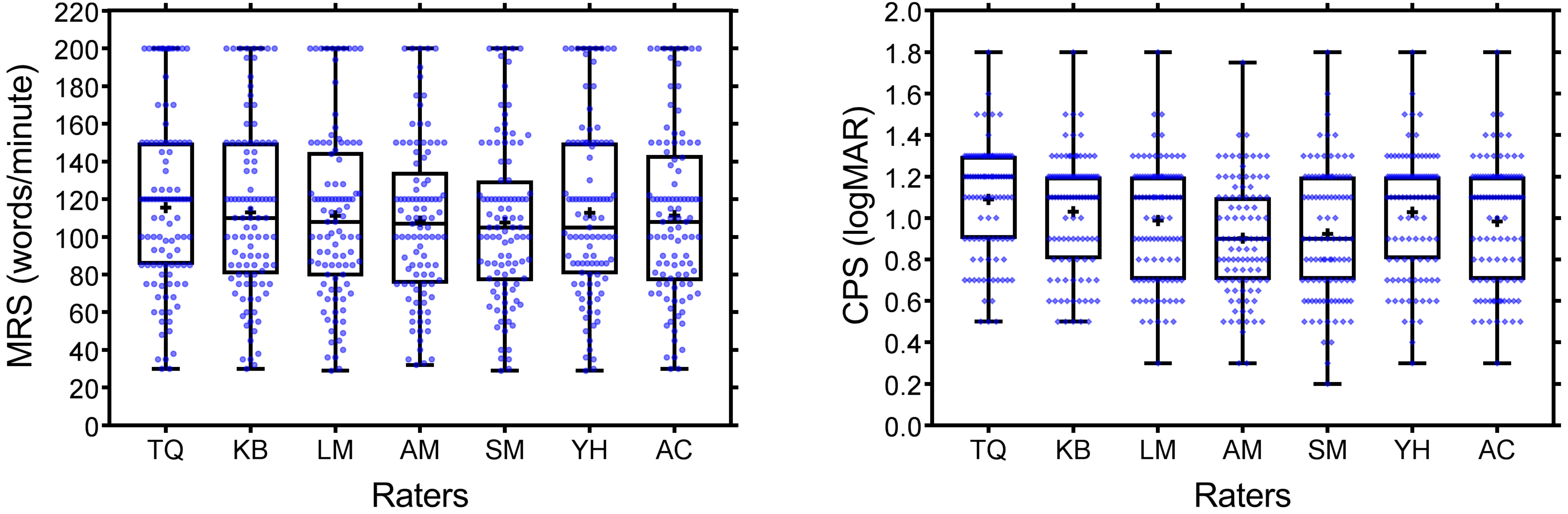

A - Standard-shaped MNREAD curve
Maximum Reading Speed (MRS)
Critical Print Size (CPS)
Reading Speed (words/min)
Corrected Print Size (logMAR)
B - MNREAD curve example in presence of a ring scotoma
Reading Speed (words/min)
Corrected Print Size (logMAR)
C - MNREAD curve example when using mutiple fixation sites
Reading Speed (words/min)
Corrected Print Size (logMAR)
D - MNREAD curve example with a noisy incomplete dataset
Reading Speed (words/min)
Corrected Print Size (logMAR)

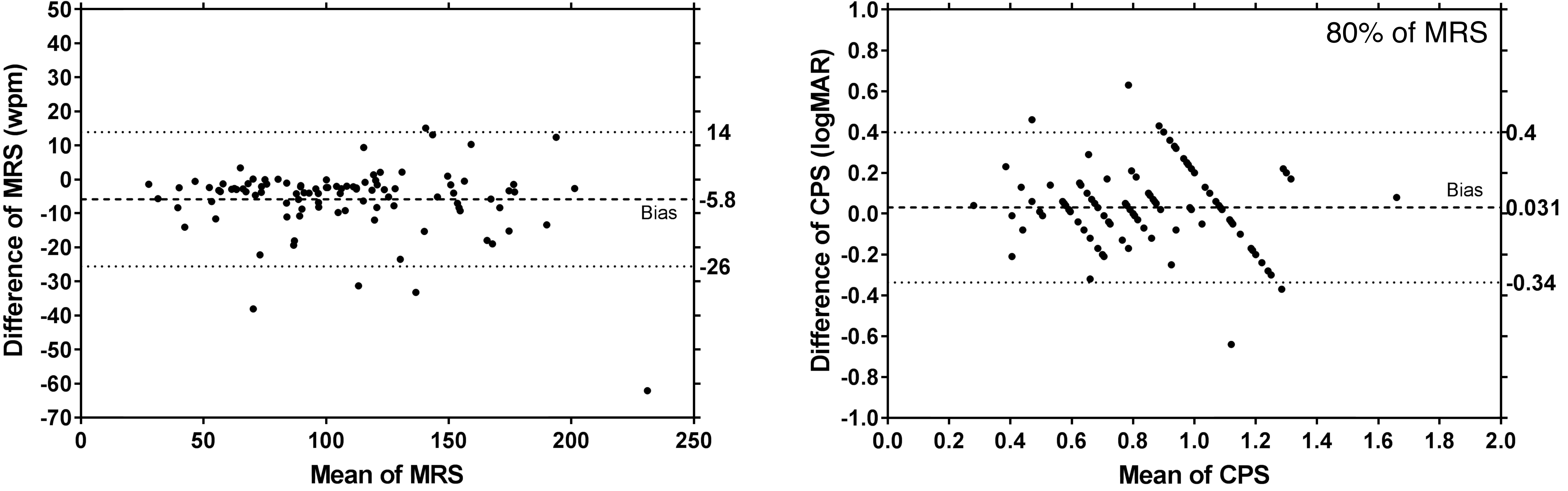
Figure


MNREAD curve
Plateau 1
Plateau 2
Plateau 3
Plateau 4
Maximum Reading Speed (MRS)
57.2
Critical Print Size (CPS)
59
43
46
60
59
62
60
55
50
45
40
35
30
25
20
15
0.7
0.8
0.9
1.0
1.1
1.2
1.3
Reading Speed (words/min)
Corrected Print Size (logMAR)

Figure

Click here to access/download;Figure;Fig_3.pdf

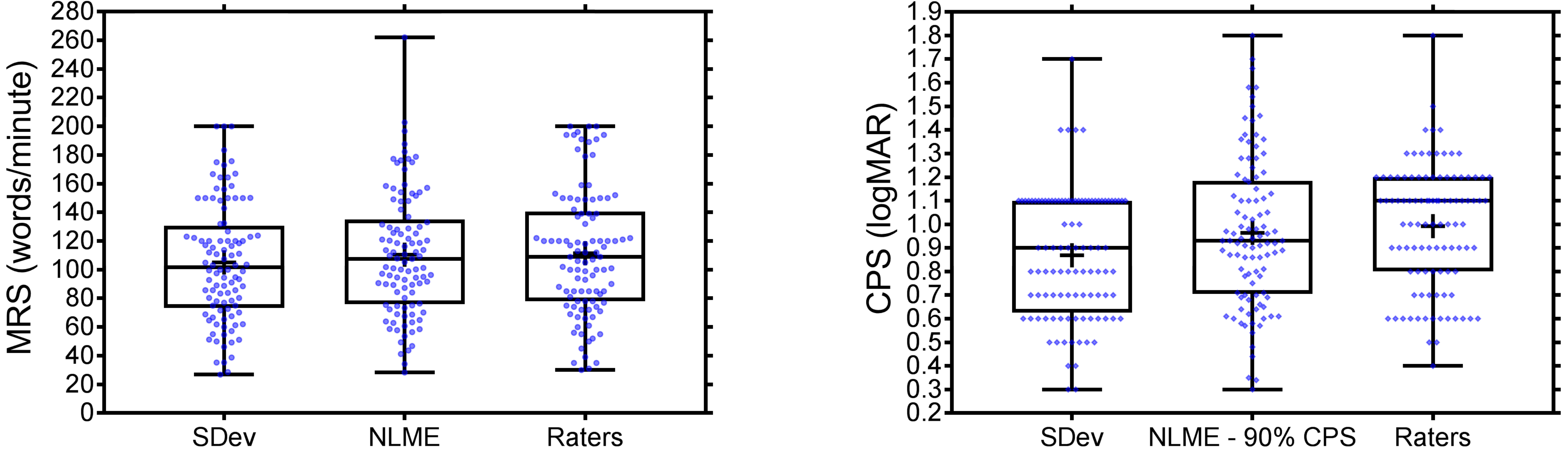




Difference of MRS (wpm)
50
40
30
20
10
0
-10
-20
-30
-40
-50
-60
-70
14
Bias
-5.8
-26
0
50
100
150
200
250
Mean of MRS
Difference of CPS (logMAR)
1.0
0.8
0.6
0.4
0.2
0.0
-0.2
-0.4
-0.6
-0.8
-1.0
80% of MRS
0.4
Bias
0.031
-0.34
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6
1.8
2.0
Mean of CPS

Figure

Click here to access/download;Figure;Fig_5.pdf



Click here to access/download
Supporting Information
S1_Appendix.pdf