



ORIGINAL ARTICLE

Comparison, within-session repeatability and normative data of three phoria tests

Alessio Facchin^{a,b,*}, Silvio Maffioletti^{a,c}^a IRSOO – Institute of Research and Studies in Optics and Optometry, Vinci, Italy^b COMIB, Center of Research in Optics and Optometry, University of Milano-Bicocca, Italy^c Degree Course in Optics and Optometry, University of Turin, Turin, Italy

Received 12 February 2020; accepted 26 May 2020

Available online 4 August 2020



KEYWORDS

Phoria;
Measurement;
Repeatability;
Comparison;
Method

Abstract

Purpose: Phoria measurements form a routine part of the examination of binocular vision. Various studies have compared different methods of phoria measurement and their repeatability between sessions. However, no studies have accounted for within-session repeatability, and few norms have been available to date. Our aims were to assess the short-term within-session repeatability, the agreement and the comparisons between three phoria tests and the delineation of normative data.

Method: The participants comprised 315 refractive corrected participants with a normal binocular vision, with ages between 20 and 80 years. Phoria was measured using three methods: von Graefe (VG), modified Thorington (MT) and Thorington (TH) procedures. Each measurement was taken 3 consecutive times at far and near.

Results: The repeatability between measurements was high (0.87 to 0.96), but the relationship between tests was medium (0.407 to 0.682). About the mean bias, VG show more exo values (0.02 to 0.14 at far and 0.07 to 0.14 at near) and MT and TH similar values (−0.04 to 0.08 at far and −0.1 to 0.03 at near). The best agreement between tests was for TH and MT (LoA = 2.33 at far and LoA = 4.44 at near). Normative data for non-presbyopic were provided.

Conclusion: Overall, there is a high agreement between MT and TH. Conversely, VG shows more exo values at near and shows large limits of agreement. We recommend that subjective measurements of phoria can best be quantified once using the MT or TH techniques in free space and the values compared with updated norms.

© 2020 Spanish General Council of Optometry. Published by Elsevier España, S.L.U. This is an open access article under the CC BY-NC-ND license (<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

* Corresponding author at: Institute of Research and Studies in Optics and Optometry, Vinci, Piazza della Libertà 18 50059, FI, Italy.
E-mail address: alessiopietro.facchin@gmail.com (A. Facchin).

Introduction

During a clinical evaluation of the visual system, several areas required investigation and some guidelines were developed in order to aid in the checking of all these areas.¹⁻³ During a diagnostic testing routine, it is important to evaluate and diagnose accommodative, ocular motor, and non-strabismic binocular vision problems.⁴ One of the steps in the evaluation of binocular vision disorders is the measurement of the phoria (or heterophoria), both at distance (3 m) and at near (40 cm). Phoria is defined as a "misalignment in the horizontal, vertical, or cyclo-direction that is corrected or correctable by disparity vergence".⁵

Several tests and procedures are available to clinicians for the subjective measurement of phoria, such as the estimated cover test, the cover test with prisms (subjective and objective), the Von Graefe technique (continuous and flash presentation), the Maddox rod, the Thorington and the modified Thorington (continuous and flash presentation).⁶⁻¹⁰ These tests, which are sometimes applied under different names, have some common features but they mostly differ in some form or another. All of them may be used in clinical practice. However, the usefulness of a tests needs to be determined not subjectively by clinicians but by using objective scientific criteria such as validity, repeatability, available expected values or norms, and other measurement parameters.^{11,12}

In the past, several studies have taken into account reliability, repeatability and the agreement between different phoria tests.^{8,13-17} Three main comparisons have generally been made: between test, between session, between examiner.

Some groups of studies have primarily taken into account the inter-rater reliability, showing that the modified Thorington (MT) and the Thorington (TH) test present the highest correlation between examiners and the smallest limits of agreement (LoA).^{7,8,14} when compared with the Von Graefe (VG) and Cover Test.^{17,18} Apart from that, the repeatability of phoria measurement was usually higher than 0.90 (correlation coefficient) for test-retest evaluation. Conversely, the relationship between different tests was not as high, with a large variation between studies when different procedures were used.⁷ With respect to the agreement between measurements, the Thorington and the modified Thorington (MT) techniques were the tests with lower differences, and the Von Graefe (VG) test produced more exo-values and high variations compared to other tests for the poor control of accommodation and its specific methodology of administration.¹³

However, in order to perform a comparison and to obtain a stable value of phoria, different studies break binocularity for long periods of time. Frequently, a 5-min period of monocular occlusion was used to allow the dissipation of disparity vergence.^{16,20,21} This procedure appears attractive, but the clinical relevance of phoria measurement was compromised because the length of the time that binocular vision was broken directly affects the measurement.¹³ This method was useful for pure research purposes, excluding vergence adaptation, but it does not give valid results for clinical purposes because this condition is not ecological.

Different phoria tests give different results for their specific accommodative control,²² procedure and dissociation methods. Coherently, each method should require specific norms,²³ but few norms are available.⁹ Well-established norms require the best-corrected refraction to be used. In actual fact, some cited studies have used habitual correction that does not permit the results to be generalised as normative. The control of accommodation under best-corrected refraction is required for the testing of binocular vision to define normative data.^{8,16}

In some of the studies that have been mentioned, for experimental purposes and to improve the reliability, each measurement was made typically three times.²⁴ However, the question must be asked as to whether this measurement is important under the day-to-day working conditions encountered in the clinic? Coherently, with this observation, the main question that we want to answer in this study is, "Is it necessary to measure phoria at different times during a routine clinical examination?" This could be referred to as "within-session repeatability". This is different from the classic between-session test-retest repeatability after between 24 and 48 h.¹⁶ The repeatability within session has its importance because it represents the first source of variability of measurement without any changes and also provides the baseline for the study of variability between sessions. To our knowledge, the within session repeatability has never been taken into account in phoria measurement.

Consequently, we have aimed to evaluate the short-term within-session (between measurement) repeatability, and also to consider three methods of phoria measurement using a new and robust statistical approach. Moreover, we have aimed to define normative data for each test.

Material and methods

Subjects

A total of 331 participants were initially enrolled in the study. The inclusion criteria were: the presence of normal binocular vision assessed by cover test, the absence of ocular pathology reported by the participants or by signs and symptoms observed by examiners, no use of medication or the presence of a disease that could affect accommodation and/or binocular vision, no history of refractive surgery, strabismus or amblyopia, a best corrected visual acuity (BCVA) equal to or greater than to +0.1logMAR in each eye at distance using a Sloan logMAR chart.²⁵

The exclusion criteria were the presence of phoria values that exceeded the range of TH/MT cards (+/-4 Δ at far and +10/-15 Δ at near). A total of 16 participants did not meet the inclusion criteria (11 fell outside the range of the TH/MT cards at far, and 5 presented a monocular VA lower than +0.1logMAR) and were consequently excluded. Therefore, 315 participants took part in the study. The mean age was 33.4 years, SD 13.6, range 18-80. There were 85 (i.e. 25.7%) participants over 40 years of age. Based on the same criteria, participants were split into two groups of non-presbyopic (NP, 231 participants, age 18-40) and presbyopic (P, 84 participants, age 41-80). The study was carried out in accordance with the guidelines given in the Declaration of

Helsinki and the participants gave their written consent to participate in the study.

Tests

VonGraefe (VG)

The phoropter and Risley rotary prisms were only used for this procedure. For the measurements at far (3 m), the distance refractive prescription was used. A 6 Δ base-up prism was placed in front of the right eye and a 10 Δ base-in was placed in front of left. The projected target was a vertical column of letters with a size of +0.2 LogMAR. For the measurements at near (40 cm), a 9 Δ prism was placed base-up in front of the right eye and a 15 Δ was placed base in in front of the left, together with the appropriate refractive correction. A 40 cm reduced-size Snellen table was used as a target. For both measurements, the participants were instructed to fixate the upper image and to keep the letter clear. The horizontal prism was reduced up to the two targets which were aligned (at about 1.0–1.5 Δ for one second). The amount of prism used was recorded as the phoria value.

Modified Thorington (MT)

In order to produce comparable conditions between this and the subsequent Thorington method, a similar target was used. Black-and-white phoria cards equivalent to those used for the Thorington method were used for far and near measurements. In each card, a small central hole was made so that a thin LED light could be inserted. Measurements were made at 3 m and at 40 cm. For these measurements, the relative distance prescription was placed into the trial frame, together with a multiple Maddox red cylinder. The participants were instructed to fixate the card, to keep the numbers clear and to report the number and the direction (left/right) that correspond to the red vertical line compared to the light. The number, according to the direction was recorded as the phoria value.

Thorington (TH)

Participants viewed the Facchin Phoria Card (COI vision, Milano, Italy), which is based on the Thorington technique.⁸ Two cards were used: one for 3 m (far) and one for 40 cm (near) distances. The specific prescription was placed into the trial frame. The participants were asked to look at the single “coloured row of numbers with an arrow”. At this point, a 6 Δ base-up prism was inserted in front of the right eye and the participants were asked, “Can you see two rows of numbers? The arrow for the upper row indicates a number in the lower row. Which number is it? What colour is it?” According to the construction, light blue values corresponded to exo- and pink to eso- values. The value reported by participants was recorded as the phoria value. The same procedure was applied at near.

Procedure

Independently of their habitual correction, for testing purposes, all participants were refractive correct at distance. At near, the addition for presbyopic patients was calculated taking into account their age²⁶ and by balancing positive and negative relative accommodation.⁵ The testing order (far

and near) and test sequence (VG, MT, TH) were balanced across participants.

The participants first gave their written consent to participation in the study and the general procedure was explained to them. Subsequently, the best subjective refractive corrections at distance and at near were assessed. Phoria measurement started at near or at far depending on balancing. The first test administered therefore also depended on the sequence of test used. Three measurements for each test were taken in sequence and normal binocular vision was re-established between each measurement. After three measurements for the first test using the same procedure, the second and the third tests were performed. Subsequently, the same procedure was replicated for the other distance. The whole procedure comprised a total of 18 measurements. In order to obtain a large sample of participants with different ages and to avoid the influence of a single examiner, the whole procedure was applied by eight examiners in their offices.

Statistical analyses

The different terms used in studies of the measurement properties of tests have at times been confused. In this study, in order to avoid confusion, we have used the term “repeatability” for test-retest correlation comparisons between measurement, “agreement between measurement” for the estimation of bias and the limits of agreement between the three measurements. Furthermore, we have also used the term “relationship between tests” for the correlation between tests (at first measurement) and “agreement between tests” for calculation of bias and LoA between tests (at first measurement). For each analysis, the results for the far and near tests have been discussed independently. For repeatability and convergence between tests, the Intraclass Correlation Coefficient (ICC) selected²⁷ is the two-way model single score ICC [ICC(A,1)].²⁸ The interpretation of correlation coefficients was based on five steps, each of 0.2 points between 0 and 1, with the respective subdivision: low, low to moderate, moderate, moderate to high and high.^{29–31} Since there is evidence for differences between non-presbyopic (NP) and presbyopic (P) in the agreement of tests of phoria measurement,²⁴ this between-participants factor was taken into account during the analyses.

In order to evaluate agreement, two analyses were performed. The first was focused on bias. Since the results were expressed as a combination of the mean values between measurements and tests, we have reported the results of both frequentist and Bayesian repeated measure ANOVA.³² Modern Bayesian methods and procedures provide richer information compared to the Null-hypothesis significance testing (NHST).³³ Bayesian analyses overcome the problems of p-value using a modern approach. Specifically, one advantage related to our aim is to quantify the evidence for the null hypothesis (similar tests result) compared to the whole resource used (number of participants).³⁴ A Bayes factor (BF10) of <0.33 was interpreted as positive evidence for H_0 and one of >3.0 as positive evidence for H_1 .^{34–36} Secondly, the Bland and Altman limits of agreement (LoA) were calculated and the plotted graphically. Bland-Altman limits of agree-

Table 1 Intraclass Correlation Coefficients (ICC) between the three sessions of testing for the two distance of test separately.

Distance	Test	ICC	95% CI	p-value
Far	VG	0.925	0.910 to 0.938	<0.0001
	MDX	0.908	0.890 to 0.923	<0.0001
	TH	0.871	0.846 to 0.892	<0.0001
Near	VG	0.948	0.938 to 0.957	<0.0001
	MDX	0.961	0.954 to 0.968	<0.0001
	TH	0.949	0.939 to 0.958	<0.0001

ment (LoA; $1.96 \times \text{SD}$) provide the mean value and the range of differences between the test and re-test or between different measurements.^{37,38} In order to compare tests and/or measurements, the LoA coefficients were compared. LoA corresponds perfectly to the previously named coefficient of reliability³⁹ and coefficient of repeatability¹⁴ used in other studies. Finally, a series of tests was performed to assess the normality of the data and their deviation from zero, and hence define parametric and/or non-parametric normative data. As standard practice, esophoria values were reported with a positive sign (+) and the exophoria values with a negative sign (−). Data were analysed and represented using the R statistical environment and JASP.^{40,41}

Results

Repeatability

The ICC scores produced by the three measurements for the three tests at far and near are reported in Table 1. These showed high values. The VG showed a high degree of correlation for a wide range of values compared to the other tests. In fact, the other two tests were limited in the range between $+4/-4^{\Delta}$. In addition, at near the three tests present a high degree of correlation for the three measurements examined. No significant differences between non-presbyopic (NP) and presbyopic (P) participants were found.

Relationship between tests

Firstly, we tested the relationship between all tests for far and near vision. Data are reported in Table 2. From these results there emerges a medium relationship between tests at far, and a medium to high correlation at near. However, these results are comprehensive and they cannot provide information on the specific comparison between tests. In order to accomplish this, the previous analyses between three tests were split into three comparisons between two tests. The results generally showed medium values, with the highest correlations between MT and TH. No significant differences between non-presbyopic (NP) and presbyopic (P) participants were found.

Bias

Since the bias can be calculated between measurements tests and groups, and these three factors could interact, a unique factorial approach of analysis was primarily used. For

far measurements, a repeated measure analysis of variance was performed using two within factors and one between factor: Test with three levels (VG, MT, TH), Measurement with three levels (M1, M2, M3) and Group with two levels (NP, P). Bayesian analysis was performed using the same approach.

The results at far showed a significant effect of Test [$F_{(2,626)} = 5.57$, $p < 0.005$, $\eta^2_p = 0.02$, $\text{BF}_{10} = 6.20\text{E}+20$], a significant effect of Measurement [$F_{(2,626)} = 14.96$, $p < 0.0001$, $\eta^2_p = 0.05$, $\text{BF}_{10} = 0.02$], a significant effect of Group [$F_{(1,313)} = 6.69$, $p < 0.05$, $\eta^2_p = 0.02$, $\text{BF}_{10} = 3.93$], a significant interaction Test $\times$ Group [$F_{(2,626)} = 29.45$, $p < 0.0001$, $\eta^2_p = 0.03$, $\text{BF}_{10} = 2.36\text{E}+21$], a significant interaction Measurement $\times$ Group [$F_{(2,626)} = 9.76$, $p < 0.0001$, $\eta^2_p = 0.03$, $\text{BF}_{10} = 0.08$], a significant interaction Test $\times$ Group [$F_{(2,626)} = 29.45$, $p < 0.0001$, $\eta^2_p = 0.03$, $\text{BF}_{10} = 2.19\text{E}+34$], a significant interaction between Test and Measurement [$F_{(4,1252)} = 5.75$, $p < 0.0001$, $\eta^2_p = 0.02$, $\text{BF}_{10} = 2.59\text{E}+16$] and a significant triple interaction Test $\times$ Measurement $\times$ Group [$F_{(4,1252)} = 4.42$, $p < 0.005$, $\eta^2_p = 0.01$, $\text{BF}_{10} = 5.62\text{E}+19$].

In order to show the differences in each test between measurements better, a series of one-way ANOVA was performed for each test using the factor Measurement with three levels (M1, M2 and M3). The results for non-presbyopic show a non-significant effect for VG ($p = 0.57$, $\text{BF}_{10} = 0.03$), MT ($p = 0.38$, $\text{BF}_{10} = 0.04$) and TH ($p = 0.33$, $\text{BF}_{10} = 0.05$). For presbyopic participants, the results show a significant effect for VG [$F_{(2,166)} = 15.48$, $p < 0.0001$, $\eta^2_p = 0.16$, $\text{BF}_{10} = 19678$], a significant effect for MT [$F_{(2,166)} = 3.38$, $p < 0.05$, $\eta^2_p = 0.04$, $\text{BF}_{10} = 0.82$], and a non-significant effect for TH ($p = 0.77$, $\text{BF}_{10} = 0.05$).

The same approach viewed for distance phoria was repeated at near. The results showed a significant effect of Test [$F_{(2,626)} = 46.32$, $p < 0.0001$, $\eta^2_p = 0.13$, $\text{BF}_{10} = 1.95\text{E}+67$], a significant effect of Measurement [$F_{(2,626)} = 3.37$, $p < 0.05$, $\eta^2_p = 0.01$, $\text{BF}_{10} = 0.05$], a significant effect of Group [$F_{(1,313)} = 19.81$, $p < 0.0001$, $\eta^2_p = 0.06$; $\text{BF}_{10} = 1412$]. Significant second order interaction were Test $\times$ Group [$F_{(2,626)} = 4.72$, $p < 0.05$, $\eta^2_p = 0.01$, $\text{BF}_{10} = 2.6\text{E}+70$], Measurement $\times$ Group [$F_{(2,626)} = 4.73$, $p < 0.01$, $\eta^2_p = 0.02$, $\text{BF}_{10} = 7.90$] and Test $\times$ Measurement [$F_{(4,1252)} = 3.17$, $p < 0.05$, $\eta^2_p = 0.01$, $\text{BF}_{10} = 1.13\text{E}+65$].

A series of one-way ANOVAs performed on each specific test, intended to assess the difference between measurements, showed, for non-presbyopic, a non-significant effect for VG ($p = 0.09$, $\text{BF}_{10} = 0.15$), a non-significant effect for MT ($p = 0.14$, $\text{BF}_{10} = 0.10$) and a non-significant effect for TH ($p = 0.12$, $\text{BF}_{10} = 0.02$). For presbyopic participants, the results showed a non-significant effect for VG ($p = 0.4$,

Table 2 Intraclass correlation coefficients between the three tests for the two distance separately.

Distance	Test comparison	ICC	95% CI	p-value
Far	All	0.528	0.458 to 0.593	<0.0001
	VG - MT	0.589	0.494 to 0.667	<0.0001
	MT - TH	0.588	0.511 to 0.656	<0.0001
	TH - VG	0.407	0.297 to 0.503	<0.0001
Near	All	0.615	0.520 to 0.691	<0.0001
	VG - MT	0.650	0.522 to 0.740	<0.0001
	MT - TH	0.682	0.617 to 0.738	<0.0001
	TH - VG	0.517	0.317 to 0.653	<0.0001

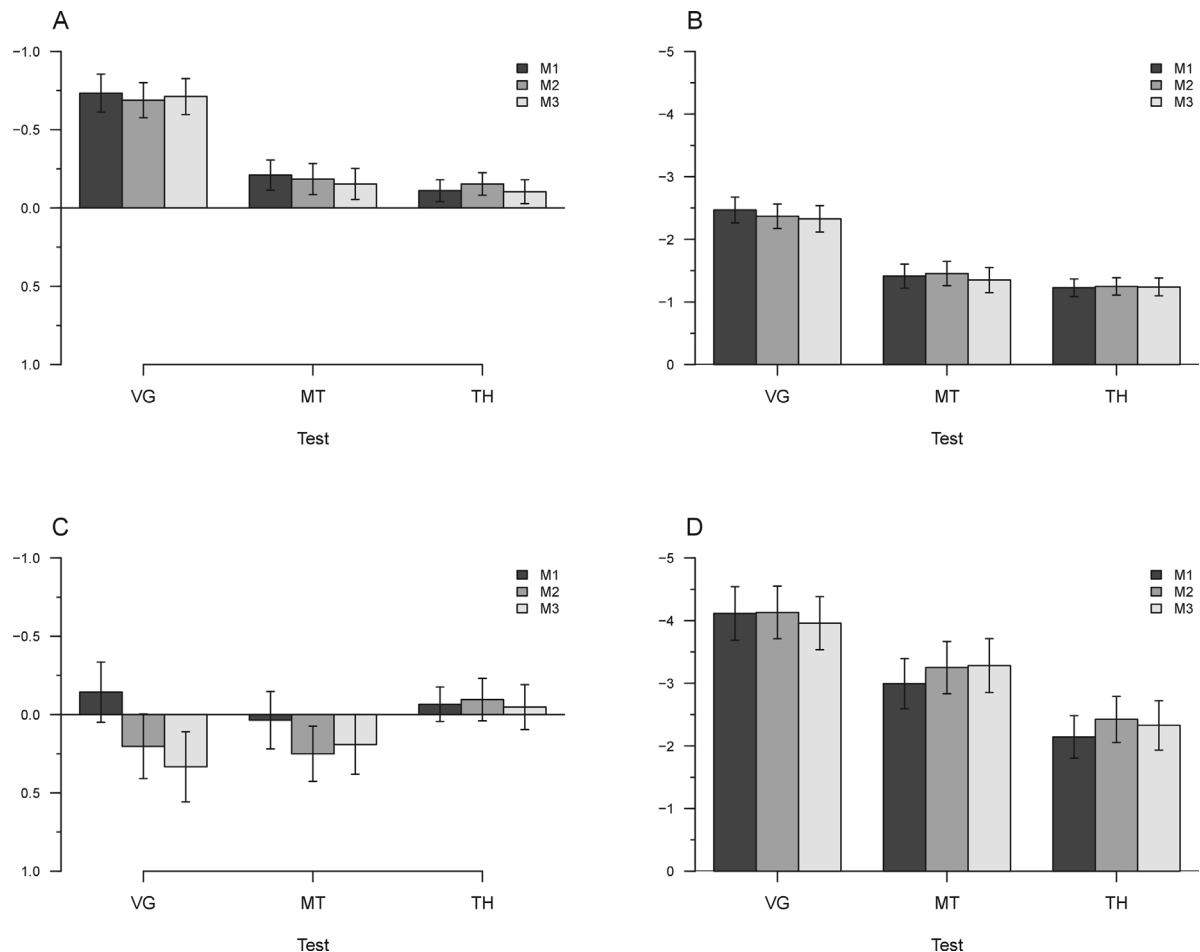


Fig. 1 Mean results of phoria evaluated at far and near for non-presbyopic (A, B) and for presbyopic (C, D). Data were expressed in prismatic diopter and separated according to test (VG = Von Graefe, MT = Modified Thorington and TH = Thorington) and measurements (M1, M2 and M3). Positive values represent esodeviations and negative exo. Bars represent ± 1 SEM.

BF10 = 0.09), MT ($p = 0.06$, BF10 = 0.63) and TH ($P = 0.07$, BF10 = 0.50). The results are shown in Fig. 1.

Agreement between measurements

The LoA coefficients for the different tests and measurements are given in Table 3 and shown in Figs. 2 and 3, separated for presbyopic and non-presbyopic participants. Following the criteria of the smallest LoA range, the best agreements were obtained for TH followed by MT and VG

for distance and near phoria. In general the agreement for presbyopic participants was lower (high LoAs).

Agreement between tests

The LoA coefficients between tests at first measurement are listed in Table 4 and shown in Fig. 4. The best agreement as smallest LoA range was between MT and TH at far and near. No consistent difference between NP and P participants was found.

Table 3 Limits of Agreement and mean bias for the three phoria tests between each measurement separated for non-presbyopic and presbyopic. The units of these data are prismatic diopters.

Distance	Test	Measurement comparison	Non presbyopic			Presbyopic		
			Lower limit	Mean diff.	Upper limit	Lower limit	Mean diff.	Upper limit
Far	Von Graefe	1 – 2	–1.28	0.05	1.37	–1.31	0.35	2
		2 – 3	–1.23	–0.02	1.18	–0.98	0.13	1.24
		1 – 3	–1.25	0.02	1.29	–1.42	0.48	2.38
	Modified Thorington	1 – 2	–0.97	0.03	1.03	–1.34	0.21	1.77
		2 – 3	–1.36	0.03	1.42	–1.2	–0.06	1.08
		1 – 3	–1.14	0.06	1.26	–1.66	0.15	1.97
	Thorington	1 – 2	–0.97	–0.04	0.88	–1.31	–0.03	1.25
		2 – 3	–1.23	0.05	1.33	–0.85	0.05	0.94
		1 – 3	–1.03	0.01	1.04	–1.42	0.02	1.46
Near	Von Graefe	1 – 2	–1.9	0.1	2.1	–2.41	–0.02	2.37
		2 – 3	–1.9	0.04	1.98	–1.75	0.17	2.09
		1 – 3	–1.87	0.14	2.15	–2.99	0.15	3.3
	Modified Thorington	1 – 2	–1.46	–0.04	1.37	–2.51	–0.26	2
		2 – 3	–1.68	0.1	1.89	–1.79	–0.03	1.73
		1 – 3	–1.42	0.06	1.55	–3.04	–0.29	2.47
	Thorington	1 – 2	–1.28	–0.02	1.23	–2.59	–0.28	2.03
		2 – 3	–1.48	0.01	1.5	–1.51	0.1	1.7
		1 – 3	–1.23	–0.01	1.21	–2.68	–0.18	2.31

Table 4 Limits of Agreement between tests and mean bias for the three phoria tests separated for non-presbyopic and presbyopic. The units of these data are prismatic diopters.

Distance	Test	Non presbyopic			Presbyopic		
		Lower limit	Bias	Upper limit	Lower limit	Bias	Upper limit
Far	VG – MT	–3.48	–0.52	2.43	–2.99	–0.18	2.64
	MT – TH	–2.33	–0.1	2.13	–2.47	0.1	2.67
	TH – VG	–2.48	0.62	3.73	–3.03	0.08	3.19
Near	VG – MT	–6.22	–1.05	4.11	–6.42	–1.12	4.18
	MT – TH	–4.59	–0.19	4.22	–5.28	–0.85	3.57
	TH – VG	–4.17	1.24	6.65	–3.21	1.97	7.15

Interexaminer differences

The main aim of this study was to compare tests and measurements, without considering the role of examiner by whom they were appropriately employed, in order only to collect data in different offices and expand the age range. Each examiner collected data for a different number of participants within different age ranges, depending on their availability. For these reasons a comparison cannot be defined as interexaminer differences. However, a comparison with a repeated measure ANOVA showed no significant effect of Examiner ($p = \text{n.s.}$ at far and $p < 0.05$ at near), but Bayesian showed an uncertain result ($\text{BF}_{10} = 0.67$ at far and 2.54 at near).

Normative data definition

Since we recruited more than three hundred participants, the goal was to define specific norms (also called expected

values)⁴ for each test taking into account the specificity of each test, distance and the distribution of data.

Significant differences in mean data were found in the previous analysis between non-presbyopic and presbyopic participants. For this reason, normative data cannot be defined without taking into account the influence of age. Since the two groups did not comprise balanced numbers (i.e. 231 participants for NP and 84 for P) and for a minimum calculation of percentile one hundred cases were necessary, we decided to develop normative data only for non presbyopic participants. Considering only the first administration, we have performed the Shapiro-Wilk test for normality and calculated the mean and SD, as well as one sample t -test, Wilcoxon signed ranked test and Bayesian one sample t -test against zero. These data are listed in [Table 5](#).

In order to ascertain whether different tests require different norms, using only the first administration of each test, we have performed two One-Way ANOVAs with the factor tests for far and near phoria. The results at far show a

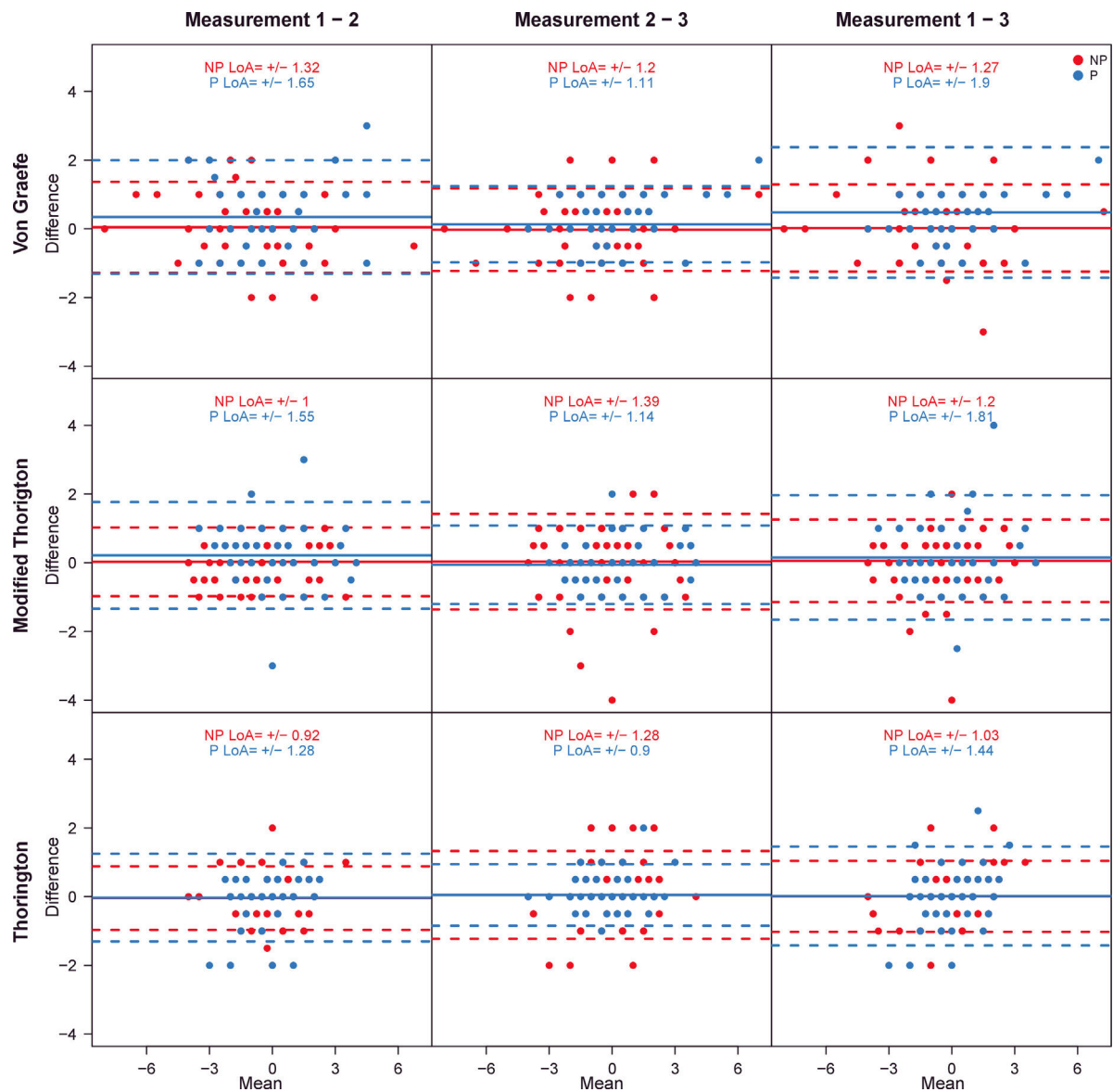


Fig. 2 Bland – Altman plots that show the agreement between measurements at far. The solid line represents the mean difference, the dashed lines represent the upper and lower limits of agreement. Values were separated for non-presbyopic and presbyopic. Many points were overlapped. The units of X and Y axes are prismatic diopters.

Table 5 Descriptive statistics of the sample of 231 non presbyopic participants together with the test for the distribution of data, one sample *t*-test, Wilcoxon signed ranked test and Bayesian one-sample *t*-test. Mean results are expressed in prismatic diopters.

Distance	Test	Mean	SD	Shapiro-Wilk p-value	One sample <i>t</i> - test p-value	Wilcoxon signed rank test p-value	Bayesian one-sample <i>t</i> - test BF10
Far	VG	−0.73	1.84	<0.0001	<0.0001	<0.0001	1.46e+06
	MT	−0.21	1.46	<0.0001	0.03	0.03	0.75
	TH	−0.11	1.06	<0.0001	0.11	0.17	0.25
Near	VG	−2.46	3.13	<0.0001	<0.0001	<0.0001	6.10e+22
	MT	−1.41	2.94	<0.0001	<0.0001	<0.0001	1.57e+09
	TH	−1.22	2.13	<0.0001	<0.0001	<0.0001	1.24e+13

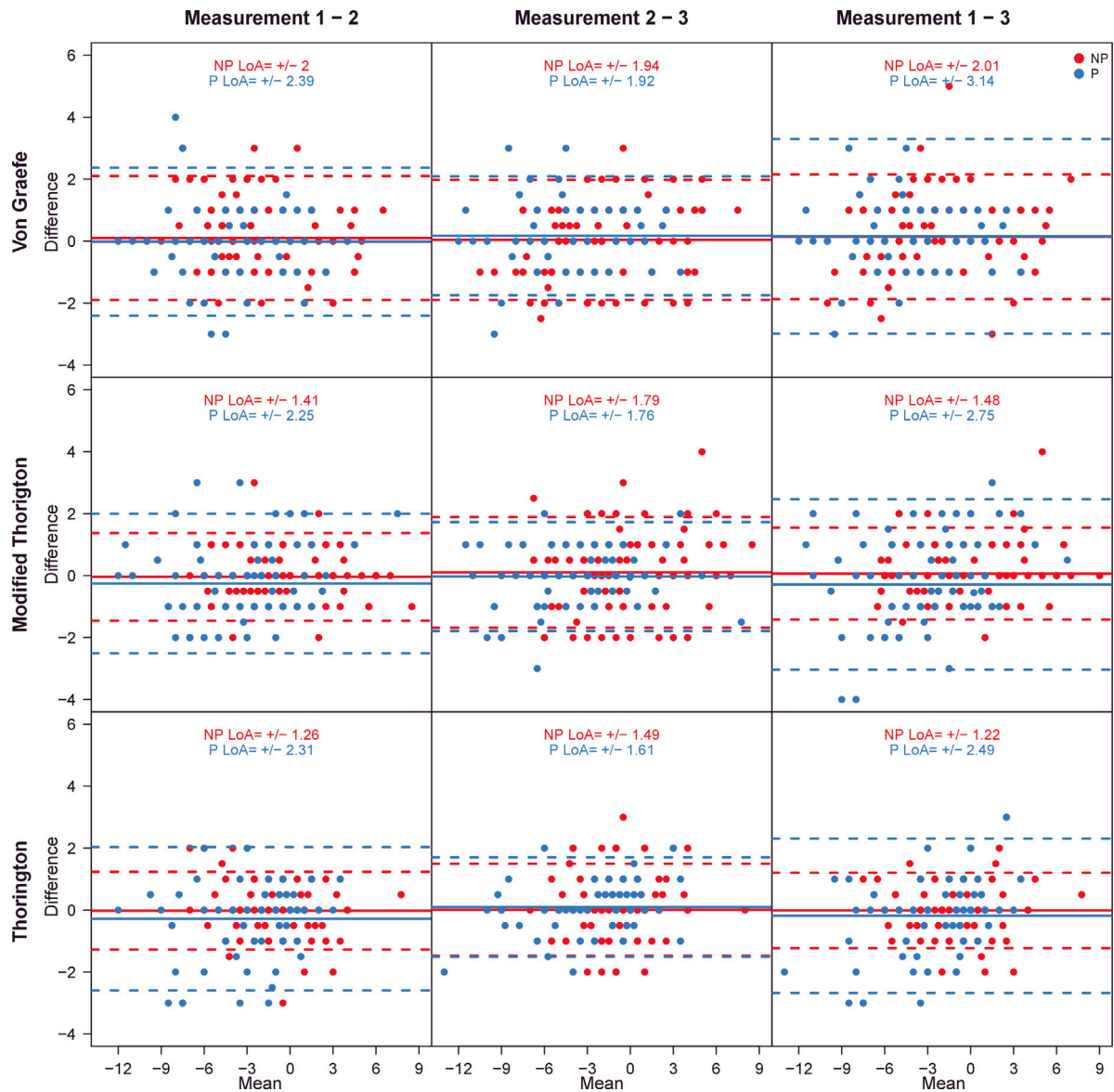


Fig. 3 Bland – Altman plots that show agreement between measurements at near. The solid line represents the mean difference, the dashed lines represent the upper and lower limits of agreement. Values were separated for non-presbyopic and presbyopic. Many points were overlapped. The units of X and Y axes are prismatic diopters.

significant result [$F_{(2,460)} = 25.58$, $p < 0.0001$, $\eta^2_p = 0.10$, $BF_{10} = 2.41E+08$). Post-hoc comparisons with Bonferroni correction show a significant difference between VG and TH ($p < 0.0001$), between VG and MT ($p < 0.0001$), but a non-significant difference between MT and TH. In addition, Bayesian Post-Hoc comparisons confirm these results: between VG and MT $BF_{10} = 32161$, between VG and TH $BF_{10} = 1.04E+06$, between MT and TH $BF_{10} = 0.18$. At near, there is a significant effect [$F_{(2,460)} = 31.58$, $p < 0.0001$, $\eta^2_p = 0.12$, $BF_{10} = 4.36E+10$]. Post-hoc comparisons with Bonferroni correction show a significant difference between VG and TH ($p < 0.0001$), between VG and MT ($p < 0.0001$), but a non-significant difference between MT and TH. Bayesian Post-Hoc comparisons confirm these results: between VG and MT $BF_{10} = 1.64E+06$, between VG and TH $BF_{10} = 9.39E+09$, between MT and TH $BF_{10} = 0.16$.

For a large sample, the Shapiro-Wilk test tends to detect even trivial shifts from the null hypothesis of normality.⁴² For this reason, we have produced a Q–Q plot and graphical representation of the distribution of score for each phoria tests (Fig. 5). Since the large part of phoria tests did present a non-normal distribution (Table 5 and Figure 6), for a practical application, we have calculated also the non-parametric percentiles that indicate the value below which a given percentage of a group of observations fall⁴³ (Table 6). Since subjects could present only the eso- or exo- values, the scale was split for the two directions, reporting only the lower bound of the scale. Using this scoring method, a percentile below 15 is comparable with 1SD below the mean and a percentile below 5 is comparable with 2SD below the mean.⁴³

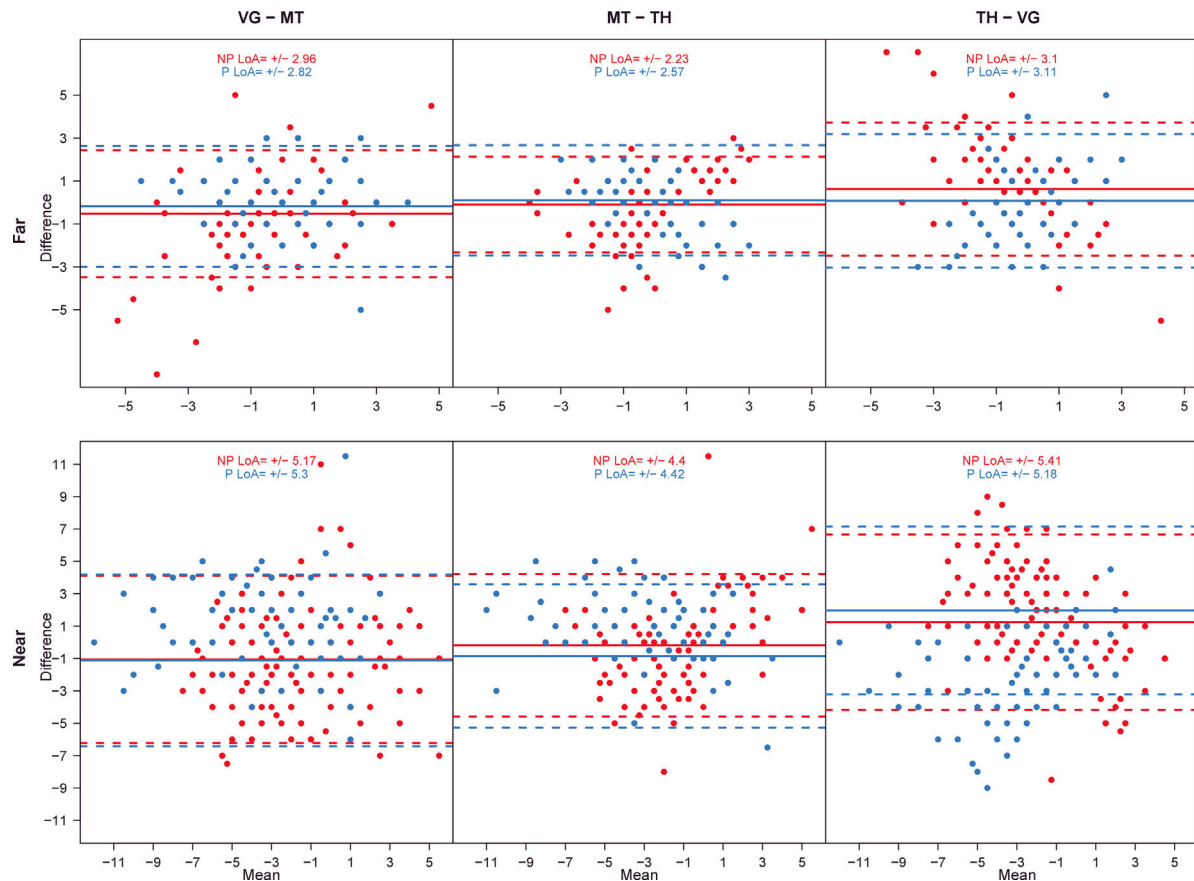


Fig. 4 Bland – Altman plots that showed agreement between test at far and near at first measurement. The solid line represents the mean difference, the dashed lines represent the upper and lower limits of agreement. Values were separated for non-presbyopic and presbyopic. Many points were overlapped. The units of X and Y axes are prismatic diopters.

Table 6 Percentile subdivision of the phoria measurements for the three test separated by distance and test. The units of these data are prismatic diopters.

		Far			Near		
		VG	MT	TH	VG	MT	TH
Percentiles exo deviation	20th percentile	−2.0	−1.0	−1.0	−5.0	−4.0	−3.0
	15th percentile	−2.0	−1.5	−1.0	−5.0	−4.0	−3.0
	10th percentile	−3.0	−2.0	−1.5	−6.0	−5.0	−4.0
	5th percentile	−3.5	−2.5	−2.0	−7.0	−6.0	−5.0
	1th percentile	−7.0	−4.0	−3.5	−9.0	−7.0	−7.0
Percentiles eso deviation	20th percentile	0.0	1.0	0.5	0.0	1.0	0.0
	15th percentile	1.0	1.0	1.0	0.0	2.0	1.0
	10th percentile	2.0	2.0	1.0	2.0	3.0	1.0
	5th percentile	2.0	2.0	1.5	4.0	4.0	2.0
	1th percentile	3.0	4.0	2.0	5.0	6.0	4.0

Discussion

In this study we have aimed to compare the results, to assess the within-session repeatability and define the norms of three phoria tests. In clinical assessment, these evaluations are necessary, because, without them, there can

be no confidence in the values produced, and therefore the clinicians cannot draw the correct conclusions from the measurements.^{11,27,44} The results are discussed separately below for repeatability, test comparison, and normative data.

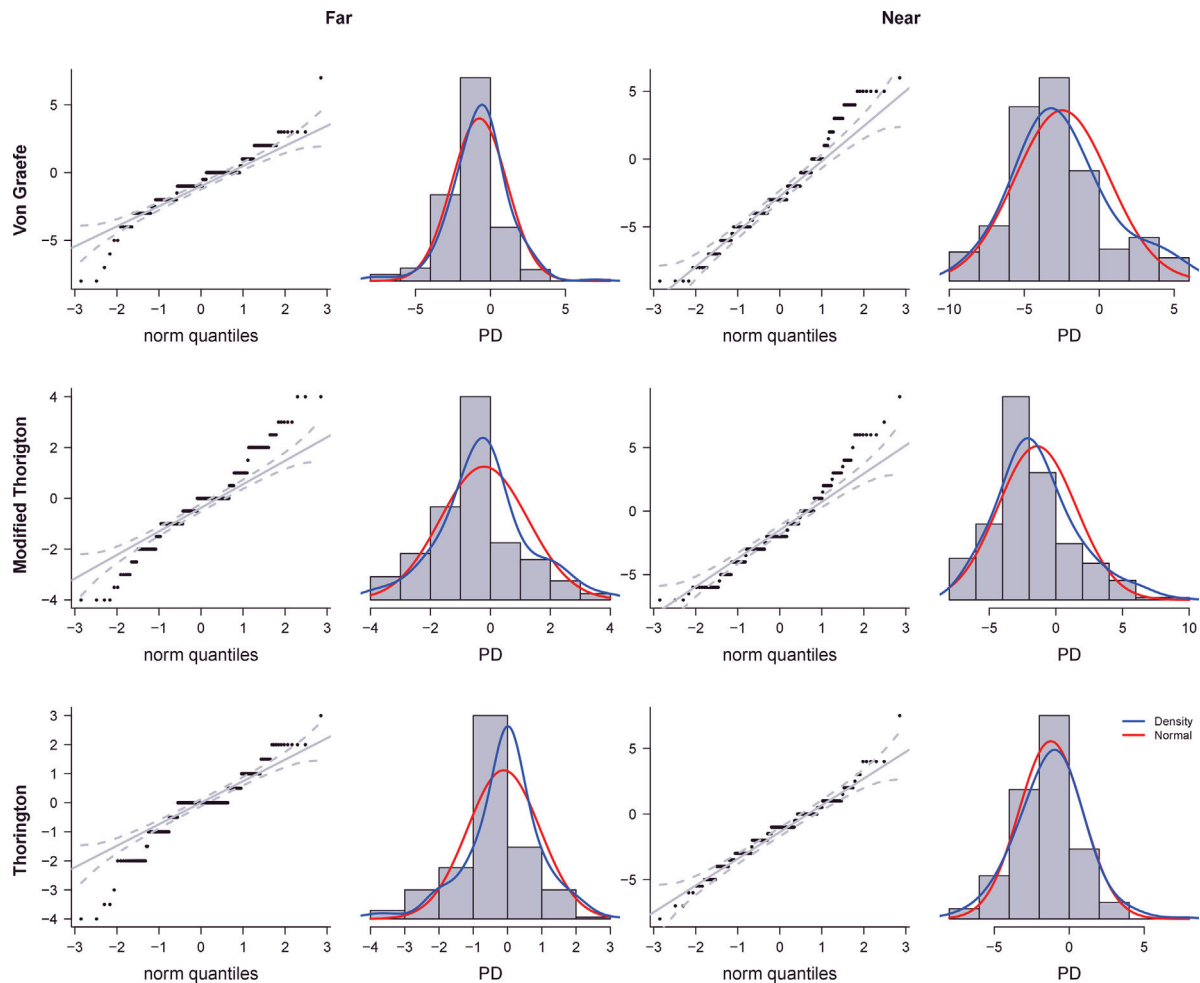


Fig. 5 Q-Q plot and density distribution function (with normal distribution) for each phoria test and distance for non presbyopic. PD = prismatic diopters.

Measurement repeatability

The terms “reliability” and “repeatability” have been applied differently in a number of studies which has resulted in some confusion, making the border between them sometimes appear unclear.^{13,14,31,45–49} In specific cases, other terms have also been used incorrectly, including reliability for test-retest repeatability,¹⁵ consistency for agreement between tests,^{7,14,50,51} repeatability for agreement¹⁴ and agreement for comparison.⁵² The results of repeatability were therefore discussed in terms of both correlation and agreement.

The aim of the study on repeatability was to verify the requirements of multiple (typically three) repetitions of measurements in clinical practice in a short period of time corresponding to the length of the clinical examination. The results are clear in the reporting of a strong relationship between measurements, with near to zero bias but with moderate LoAs at far (about 2.8 for VG, 2.6 for MT and 2.2 for TH) and at near (about 4.3 for VG, 3.6 for MT and 3.2 for TH) with larger LoAs for presbyopic participants. For both distances, the best agreement between measurements was found for TH, medium for MT and lower for VG.

These results agree with the classical test/re-test procedure after a period of days with values that appear coherently to be slight lower.^{15,16} Specifically, the modified Thorington and Thorington tests were more repeatable than the von Graefe.^{15,16} The repeatability was better using a trial frame compared with the phoropter¹⁶, and CT at near gives some differences in measurements but they are not clinically significant.¹⁸ The Thorington and cover tests measurements showed small differences.¹⁴ Different tests performed using phoropter and trial frame in the measurement of phoria gave more stable results with the latter, and more variable with the former.^{16,53} The small values obtained here could have arisen from the difference in testing times. In our case, this was within a single session, in the other cases it was between sessions performed on different days. Observing the mean bias between measurements and tests, it appears that VG and MT at far drift to eso-values with repetitions in presbyopic but not for non-presbyopic participants. At near, some differences between sessions were found. However, the application of the Bayesian analyses has shown that some results are not sufficiently compelling and consequently they are not key factors. The strong difference between tests is the main difference found.

Test comparison

The relationship between tests found to be medium. In fact, in both of the measurements (far and near) bias and the LoAs appear larger. Overall, the mean biases between tests show the well-known phenomena that VG gives more exo- values than the other two methods.⁵⁴ Also, the LoAs demonstrate the large dispersion of the data. When comparisons were made with VG, larger LoAs were found. In fact, small LoAs were found between MT and TH, both at near and far. As a result of a medium relationship between tests, it seems that not all phoria tests are fully interchangeable and there is a requirement for specific norms for each specific test denoted by the different variance. Interestingly, a very small difference between non-presbyopic and presbyopic participants were found which does not support a previously observed large difference in agreement between cover and Von Graefe tests in comparing these two age ranges.²⁴ Otherwise, some studies of test comparisons gave similar results, with a small bias and high agreement,¹⁴ but the Von Graefe test showed more exo- values,⁵⁴ and this gives more support to the diagnosis of convergence insufficiency.²³ The use of a detailed card, together with the MT method, provides better control of accommodation,²² and a different endpoint of CT was significantly different but not clinically relevant ($<0.5^\Delta$).¹⁸

Normative data

Since testing was performed with the best refractive correction, we have also defined the norms for each specific test. Since a difference between non-presbyopic and presbyopic participants was found and in the latter group there were relatively few cases, the normative data definition was limited to non-presbyopic. Data seem not to be distributed normally, and consequently we have provided both the mean and SD scores (as other norms for comparisons), together with the percentile ranking separated for eso- and exo- values. In defining normative data, when comparing different tests a difference was found and this result confirms the requirement for a specific norm for each test. Moreover, when 1 or 2 standard deviations were added to the mean value in order to consider the cut-off limits,⁴ the difference between tests grew larger, and the requirement for specific values became important. In comparing our data with those reported in the literature, we found small values and small SDs for some data,⁴ and similar results to more recent norms.⁵⁵ Different samples, age and test modalities may explain these differences.⁴⁴

Conclusions

Our primary finding was that the repetition of tests in clinical practice is not necessary since the within-session repeatability is very high. Secondly, we have reproduced the results of the Von Graefe technique, with more exo- values obtained when measuring phoria with phoropter. Based on these results, we recommend that subjective measurements of phoria can best be quantified using either the MT or TH techniques in free space and that the values be compared with the given, specific and updated norms.

Conflict of interest

The phoria card used for Thorington technique has the name of the first author. Nevertheless, he has no commercial interest in its sale.

Acknowledgments

We would thank the students and collaborators who helped the authors in collecting data (given in alphabetical order): Francesco Cantù, Tiziana Cattaneo, Vasco Lisena, Carlotta Pomarè Montin, Andrea Giuseppe Rubini and Roberta Signorelli.

References

1. Scheiman M, Rouse MW. *Optometric management of learning-related vision problems*. Elsevier Health Sciences; 2006.
2. AAO. AOA. Vision, Learning, and Dyslexia A Joint Organizational Policy Statement American Academy of Optometry American Optometric Association. *Optom Vis Sci*. 1997;74(10).
3. AOA, Available from: <https://www.aoa.org/optometrists/tools-and-resources/clinical-care-publications/clinical-practice-guidelines>, 2015.
4. Scheiman M, Wick B. *Clinical management of binocular vision: Heterophoric, accommodative, and eye movement disorders*. 5th ed. Lippincott Williams & Wilkins; 2019.
5. Saladin JJ. Phorometry and stereopsis. In: Butterworth-Heinemann, ed. *Borish' clinical refraction*. Elsevier; 2006:899–962.
6. Schroeder TL, Rainey BB, Goss DA, Grosvenor TP. Reliability of and comparisons among methods of measuring dissociated phoria. *Optom Vis Sci*. 1996;73(6).
7. Rainey BB, Schroeder TL, Goss DA, Grosvenor TP. Inter-examiner repeatability of heterophoria tests. *Optom Vis Sci*. 1998;75:719–726.
8. Wong EPF, Fricke TR, Dinardo C. Interexaminer repeatability of a new, modified prentice card compared with established phoria tests. *Optom Vis Sci*. 2002;79(6):370–375.
9. Lyon DW, Goss DA, Horner D, Downey JP, Rainey B. Normative data for modified Thorington phorias and prism bar vergences from the Benton-IU study. *Optom Am Optom Assoc*. 2005;76(10):593–599.
10. Elliott DB. Clinical procedures in primary eye care. *Elsevier Health Sciences*. 2013.
11. Elliott DB. What is the appropriate gold standard test for refractive error? *Ophthalmic Physiol Opt*. 2017;37(2):115–117.
12. Elliott DB. Evidence-based optometry and in-practice research. *Ophthalmic Physiol Opt*. 2012;32(2):81–82.
13. Schroeder TL, Rainey BB, Goss DA, Grosvenor TP. Reliability of and comparisons among methods of measuring dissociated phoria. *Optom Vis Sci*. 1996;73(6).
14. Cebrian JL, Antona B, Barrio A, Gonzalez E, Gutierrez A, Sanchez I. Repeatability of the modified Thorington card used to measure far heterophoria. *Optom Vis Sci*. 2014;91(7):786–792.
15. Goss DA, Reynolds JL, Todd RE. Comparison of four dissociated phoria tests: reliability and correlation with symptom survey scores. *J Behav Optom*. 2010;21(4):99–104.
16. Casillas E, Rosenfield M. Comparison of subjective heterophoria testing with a phoropter and trial frame. *Optom Vis Sci*. 2006;83(4):237–241.
17. Rainey BB, Schroeder TL, Goss DA, Grosvenor TP. Reliability of and comparisons among three variations of the alternating cover test. *Ophthalmic Physiol Opt*. 1998;18(5):430–437.

18. Johns HA, Manny RE, Fern K, Hu Y-S. The intraexaminer and interexaminer repeatability of the alternate cover test using different prism neutralization endpoints. *Optom Vis Sci*. 2004;81(12):939–946.
20. Rosenfield M, Chun TW, Fischer SE. Effect of prolonged dissociation on the subjective measurement of near heterophoria. *Ophthalmic Physiol Opt*. 1997;17(6):478–482.
21. Rosenfield M. Tonic vergence and vergence adaptation. *Optom Vis Sci Off Publ Am Acad Optom*. 1997;74(5):303–328.
22. Pointer JS. An enhancement to the Maddox Wing test for the reliable measurement of horizontal heterophoria. *Ophthalmic Physiol Opt*. 2005;25(5):446–451.
23. Maples WC, Savoy RS, Harville J, Golden LR, Hoenes R. Comparison of distance and near heterophoria by two clinical methods. *Optom Vis Dev*. 2009;40(2).
24. Cantó-Cerdán M, Cacho-Martínez P, García-Munoz Á. Measuring the heterophoria: Agreement between two methods in non-presbyopic and presbyopic patients. *J Optom*. 2018;11(3):153–159.
25. Facchin A, Maffioletti S, Martelli M, Daini R. Different trajectories in the development of visual acuity with different levels of crowding: The Milan Eye Chart (MEC). *Vision Res*. 2019;156:10–16.
26. Hofstetter HW. A comparison of duane's and donders' tables of the amplitude of accommodation. *Optom Vis Sci [Internet]*. 1944;21(9). Available from: https://journals.lww.com/optvissci/Fulltext/1944/09000/A_COMPARISON_OF_DUANE_S_AND_DONDERS_TABLES_OF_THE.1.aspx.
27. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15(2):155–163.
28. McGraw KO, Wong SP. Forming inferences about some intraclass correlation coefficients. *Psychol Methods*. 1996;1:30.
29. Fleiss JL, Cohen J. The equivalence of weighted kappa and the intraclass correlation coefficient as measures of reliability. *Educ Psychol Meas*. 1973;33:613–619.
30. Viera AJ, Garrett JM. Understanding interobserver agreement: the kappa statistic. *Fam Med*. 2005;37:360–363.
31. Facchin A, Maffioletti S. The reliability of the DEM test in the clinical environment. *Front Psychol*. 2018;9.
32. Wetzels R, Grasman RPPP, Wagenmakers E-J. A default bayesian hypothesis test for ANOVA designs. *Am Stat [Internet]*. 2012;66(May (2)):104–111, <http://dx.doi.org/10.1080/00031305.2012.695956>. Available from:.
33. Wagenmakers E-J, Marsman M, Jamil T, et al. Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. *Psychon Bull Rev [Internet]*. 2018;25(February (1)):35–57, <http://dx.doi.org/10.3758/s13423-017-1343-3>. Available from:.
34. Wagenmakers E-J, Love J, Marsman M, et al. Bayesian inference for psychology. Part II: Example applications with JASP. *Psychon Bull Rev [Internet]*. 2017, <http://dx.doi.org/10.3758/s13423-017-1323-7>. Available from:.
35. Jeffreys H. *Theory of probability*. Oxford: Oxford University Press; 1961.
36. Kass RE, Raftery AE. Bayes factors. *J Am Stat Assoc*. 1995;90(430):773–795.
37. Bland JM, Altman D. Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*. 1986;327:307–310.
38. Altman DG, Bland JM. Measurement in medicine: The analysis of method comparison studies. *Stat*. 1983;32:307–317.
39. Saul T, Taylor K. Normative data for the crowded logMAR Kay's pictures vision test in children. *Br Ir Orthopt J*. 2015;9:36–43.
40. R Core Team, Vienna, Austria Available from: <http://www.r-project.org/>, 2019.
41. JASP Team, Available from: <https://jasp-stats.org/>, 2017.
42. Seier E. Comparison of tests for univariate normality. *Interstat*. 2002;1:1–17.
43. Crawford JR, Garthwaite PH. Percentiles please: The case for expressing neuropsychological test scores and accompanying confidence limits as percentile ranks. *Clin Neuropsychol [Internet]*. 2009;23(February (2)):193–204. Available from: <https://www.ncbi.nlm.nih.gov/pubmed/18609335>.
44. Howarth PA, Heron G. Repeated measures of horizontal heterophoria. *Optom Vis Sci*. 2000;77(11):616–619.
45. Costa Santos C, Bernardes J, Ayres-de-Campos D, Costa A, Costa C. The limits of agreement and the intraclass correlation coefficient may be inconsistent in the interpretation of agreement. *J Clin Epidemiol [Internet]*. 2011;64(March (3)):264–269, <http://dx.doi.org/10.1016/j.jclinepi.2009.11.010>. Available from:.
46. Costa Santos C, Bernardes J, Ayres-de-Campos D. Observer reliability and agreement: Differences, difficulties, and controversies. *J Clin Epidemiol [Internet]*. 2011;64(June (6)):702, <http://dx.doi.org/10.1016/j.jclinepi.2010.12.002>. Available from:.
47. Kottner J, Streiner DL. The difference between reliability and agreement. *J Clin Epidemiol [Internet]*. 2011;64(June (6)):701–702, <http://dx.doi.org/10.1016/j.jclinepi.2010.12.001>. Available from:.
48. Zaki R, Bulgiba A, Ismail R, Ismail NA. Statistical methods used to test for agreement of medical instruments measuring continuous variables in method comparison studies: A systematic review. *PLoS One*. 2012;7(5):e37908.
49. 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 [Internet]*. 2011;31(July (4)):330–338, <http://dx.doi.org/10.1111/j.1475-1313.2011.00851.x>. Available from:.
50. Bartlett JW, Frost C. Reliability, repeatability and reproducibility: Analysis of measurement errors in continuous variables. *Ultrasound Obstet Gynecol Off J Int Soc Ultrasound Obstet Gynecol*. 2008;31(4):466–475.
51. Watson PF, Petrie A. Method agreement analysis: A review of correct methodology. *Theriogenology*. 2010;73(9):1167–1179.
52. Jones D, Westall C, Averbeck K, Abdoell M. Visual acuity assessment: A comparison of two tests for measuring children's vision. *Ophthalmic Physiol Opt*. 2003;23(6):541–546.
53. Lam AKC, Lam AKC, Charm J, Wong K. Comparison of near heterophoria tests under varying conditions on an adult sample. *Ophthalmic Physiol Opt*. 2005;25(2):162–167.
54. Goss DA, Moyer BJ, Teske MC. A comparison of dissociated phoria test findings with von Graefe phorometry and modified Thorington testing. *J Behav Optom*. 2008;19(6):145–149.
55. Wajuihian SO. Normative values for clinical measures used to classify accommodative and vergence anomalies in a sample of high school children in South Africa. *J Optom*. 2019;12(July (3)):143–160.