Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Are Distance Stereoacuity Tests Interchangeable? Effects of Measurement Range and Disparity Resolution on Agreement

Jorge Jorge[1]

jorge@fisica.uminho.pt

University of Minho

Miguel Coelho[1]

University of Minho

João Pedro Jorge[2]

Universidade Nova de Lisboa

Research Article

Keywords: distance stereoacuity, stereopsis, method comparison, repeatability, agreement, disparity resolution

Posted Date: November 24th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11707158/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Abstract

Purpose.

To assess the repeatability and interchangeability of four distance stereoacuity methods and examine how differences in measurement range and disparity resolution affect between-method comparisons.

Methods.

Distance stereoacuity was assessed at 3 m in 65 young adults using COI, Eyenext, Distance Randot, and frontal Senaptec, in randomised order, with retesting after 10 minutes. Valid measurable thresholds were analysed on a $\log_{10}$(seconds of arc) scale. Repeatability was evaluated using paired Wilcoxon tests, intraclass correlation coefficients (ICC(A,1)), concordance correlation coefficients (CCC), and Bland–Altman limits of agreement. Between-method differences and agreement were evaluated in 61 participants with complete initial measurements, first on the native scales, then after capping thresholds to 60–200 arcsec and finally after harmonising to three ordinal levels (60, 100, and 200 arcsec).

Results.

Median initial thresholds were 60, 80, 100, and 41 arcsec for COI, Eyenext, Distance Randot, and Senaptec, respectively. The methods differed on their native scales (Friedman $\chi^2(3) = 28.007$; $p < 0.001$; Kendall's $W = 0.153$), with significant Holm-adjusted differences for Distance Randot versus COI and Senaptec. The global difference was smaller after capping to 60–200 arcsec ($\chi^2(3) = 9.348$; $p = 0.025$; $W = 0.051$), with no significant adjusted pairwise differences, and was not statistically significant after three-level harmonisation ($\chi^2(3) = 6.371$; $p = 0.095$; $W = 0.035$). Nevertheless, individual agreement remained limited. COI–Eyenext had the highest native-scale agreement (ICC = 0.665; CCC = 0.661) and three-level linear weighted kappa ($\kappa = 0.477$), whereas the remaining pairwise ICCs were 0.114–0.254 and linear weighted kappas 0.103–0.196. No method showed a statistically significant short-term test–retest shift after Holm correction; repeatability ICCs ranged from 0.527 to 0.636, with wide limits of agreement.

Conclusions.

Harmonising measurement scales attenuated group-level differences but did not establish equivalence or adequate agreement between individuals. These four methods should not be used interchangeably; the same instrument and protocol should be retained for longitudinal assessment.

Key Points

Measured stereoacuity depends partly on each test's available range and disparity steps, meaning that different instruments may report systematically different thresholds.

Making the measurement scales more comparable reduced differences between instruments, but individual participants could still receive substantially different results across tests.

For clinical follow-up and research, using the same stereotest and protocol is safer than converting thresholds obtained from different instruments.

Introduction

Stereoacuity is one of the most widely used clinical measures of binocular sensory function, but the apparently simple output in seconds of arc can obscure substantial differences in how individual tests generate that number. Clinical stereotests differ in stimulus construction, binocular separation, task demands, available disparity levels, and stopping rules. These differences matter because a stereoacuity threshold is not a direct physical measurement in the same way as length or mass; it is an estimate produced by a particular psychophysical procedure. O'Connor and Tidbury highlighted that conventional stereotests sample only a limited portion of stereoscopic function and often provide relatively few disparity levels, which restricts measurement precision.[1]

This issue is especially important at distance, where fewer validated tests are available and where the numerical ranges of commonly used methods differ substantially. Test-specific normative limits have been reported even when the same underlying construct is being assessed; Piano and colleagues found different upper limits of normality for Distance Randot and Frisby-Davis distance testing and cautioned against treating stereotests as interchangeable.[2] Similar conclusions were reached by Leske, Birch, and Holmes, who showed that real-depth and Randot tests produced systematically different thresholds both at near and at distance.[3] The commercially available Distance Randot itself has good normative discrimination and high within-test reproducibility, but those properties do not establish equivalence with other distance tests.[4]

Digital methods have widened the measurable range and made it possible to present stereoscopic stimuli with greater flexibility, but they have not removed the problem of between-test disagreement. In 64 adults, Kim and colleagues compared a 3-D monitor test with Distance Randot at 3 m; although the two tests were related, the reported correlation was modest and agreement was not exact for all participants.[5] Computerised distance stereoacuity systems have also shown good repeatability while retaining non-negligible limits of agreement with conventional tests.[6] These findings suggest that a new or digital platform cannot be assumed to be numerically equivalent to an established clinical test simply because both report thresholds in seconds of arc.

One reason for disagreement may be the architecture of the measurement scale itself. Vancleef and colleagues showed that the higher thresholds obtained with the TNO test could be partly reproduced by the test protocol and the available disparity values, rather than being explained solely by a global-versus-local stereopsis distinction.[7] Even changes between editions of the same stereotest can produce clinically important shifts in measured threshold.[8] More recent work with digital stereotests has likewise shown that testing range and stimulus parameters affect measured thresholds and between-

test relationships.[9, 10] Binocular separation method may add another source of variation: red–green anaglyph, polarised, and active-shutter presentations can yield different distance stereoacuity estimates even when the nominal task is similar.[11]

The practical problem is therefore twofold. First, tests with different lower limits, upper limits, and step sizes may show group-level differences simply because participants are forced into different response categories. Second, even if scale architecture is harmonised and the group-level difference disappears, the same individuals may still be classified differently by different methods. These are related but distinct questions, and they require both comparison of central tendency and explicit analysis of agreement.

The primary aim of the present study was to evaluate the comparability and interchangeability of four methods of distance stereoacuity assessment—COI, Eyenext, Distance Randot, and Senaptec—and to determine the extent to which between-test differences could be explained by differences in measurement range and disparity resolution. Secondary aims were to compare short-term test–retest repeatability and to determine whether individual agreement remained limited after scale harmonisation. We hypothesised that the four methods would differ on their native scales, that these differences would be attenuated by harmonising range and resolution, and that scale harmonisation would not by itself produce sufficient individual agreement to support interchangeability.

Methods

Study design and participants

A repeated-measures comparative study was conducted to evaluate four methods of distance stereoacuity assessment.

Participants were eligible if their habitual visual acuity was at least 1.0 in decimal notation, with or without optical correction, and if they had no manifest strabismus or known ocular disease. Individuals with a history of refractive surgery or strabismus were excluded. All measurements were performed using the participants' habitual optical correction, when applicable.

Testing was conducted under appropriate ambient illumination. Written informed consent was obtained from all participants. The study was approved by the Ethics Committee of the University of Minho (CEICVS 113/2020) and conducted in accordance with the Declaration of Helsinki.

Stereoacuity assessment

Four distance stereoacuity methods were evaluated: the Stereopsis Test of the Centro de Optometría Internacional (COI; Madrid, Spain), the Eyenext digital vision chart system (EyeNext Srl, Avigliana, Italy), the Distance Randot Stereotest (Stereo Optical Co., Inc., USA), and the Senaptec Sensory Station depth-

perception assessment (Senaptec Inc., Beaverton, OR, USA). Only the frontal measurement was considered for Senaptec.

All tests were performed at a viewing distance of 3 m. COI, Eyenext, and Senaptec were administered using the red–green glasses supplied by the manufacturers, whereas Distance Randot was performed using polarised glasses.

Test order was randomised to minimise order and learning effects. Participants were instructed to identify the stereoscopic targets according to each test's procedure. Testing progressed towards finer disparity levels and was discontinued after two consecutive incorrect responses. The smallest correctly identified disparity was recorded as the stereoacuity threshold.

Each test was repeated after a 10-minute interval to assess short-term repeatability.

Statistical analysis

Stereoacuity thresholds were analysed on a $\log_{10}$(seconds of arc) scale, with lower values indicating better stereoacuity. Only measurable thresholds were included in the analyses. Initial measurements were used for between-method comparisons, whereas repeated measurements were used to assess short-term repeatability.

Descriptive statistics were calculated for each method. Test–retest differences were assessed using paired Wilcoxon signed-rank tests. Repeatability and absolute agreement were evaluated using the two-way absolute-agreement, single-measure intraclass correlation coefficient (ICC(A,1)), Lin's concordance correlation coefficient (CCC), and Bland–Altman analysis.[12] Mean differences and 95% limits of agreement were calculated on the logarithmic scale and back-transformed into retest/test ratios. Differences between the four stereotests were assessed using the Friedman test, followed by pairwise Wilcoxon signed-rank tests. Kendall's W was used to quantify the overall effect size.

To examine the influence of differences in measurement range and resolution, two sensitivity analyses were performed. First, measurements were capped at 60 and 200 arcsec to establish a common numerical range. Second, values were grouped into three ordered categories (≤ 60, > 60–100, and > 100 arcsec), represented by 60, 100, and 200 arcsec, respectively.

Agreement between methods was assessed using ICC(A,1), CCC, Bland–Altman analysis, Spearman's rho, and Kendall's tau-b. Rank correlations were also examined after harmonisation of the measurement range. Agreement on the three-category scale was evaluated using linear and quadratic weighted Cohen's kappa, exact agreement, and extreme disagreement. Holm correction was applied to multiple pairwise comparisons. Statistical significance was set at $p < 0.05$.

Results

Participant characteristics and descriptive statistics

The study included 65 participants, of whom 40 (61.5%) were women and 25 (38.5%) were men. Mean age was 23.34 ± 3.68 years, with a median of 22 years (interquartile range [IQR] 21–24; range 19–38 years).

Valid initial measurements were available for 65 participants with COI, 64 with Eyenext, 62 with Distance Randot, and 65 with Senaptec. Median thresholds were 60, 80, 100 and 41 arcsec, respectively. The corresponding IQRs were 40–120, 40–160, 60–100 and 20–176 arcsec. Table 1 provides the descriptive statistics on the $\log_{10}$ scale and the corresponding values in arcseconds. The observed disparity levels varied substantially between methods (Fig. 1).

Table 1
Descriptive statistics and observed disparity levels for the initial measurement with each distance stereotest.

Method	**n**	**Median log10 [P25; P75]** **(arcsec)**	**Observed levels (arcsec)**	**Observed range (arcsec)**
COI	65	1.778 [1.602; 2.079] 60 [40; 120]	20, 40, 60, 80, 100, 120, 140, 160, 180, 200	20–200
Eyenext	64	1.903 [1.602; 2.204] 80 [40; 160]	20, 40, 80, 160, 320	20–320
Distance Randot	62	2.000 [1.778; 2.000] 100 [60; 100]	60, 100, 200, 400	60–400
Senaptec	65	1.613 [1.301; 2.246] 41 [20; 176]	10, 20, 31, 41, 52, 62, 72, 83, 93, 145, 176, 217, 240	10–240

Note. Logarithmic values are expressed as $\log_{10}$(seconds of arc); values in arcseconds are inverse-transformed medians and quartiles. The observed levels describe those represented in this sample, not necessarily every programmed disparity level available on a device.

Test–retest repeatability

Valid paired measurements were available for 65 participants with COI, 63 with Eyenext, 58 with Distance Randot and 65 with Senaptec. No systematic difference was detected between the initial and repeated measurements for any method after correction for multiple comparisons (all Holm-adjusted $p > 0.20$).

Test–retest absolute agreement was moderate. ICC(A,1) values were 0.527 for COI, 0.607 for Eyenext, 0.601 for Distance Randot and 0.636 for Senaptec; Lin's CCC showed a similar pattern (Table 2). The

geometric mean retest/test ratios were 0.850, 0.946, 0.866 and 0.808, respectively. The corresponding 95% Bland–Altman ratio limits were wide, particularly for COI (0.202–3.574) and Senaptec (0.147–4.443). Thus, the absence of a systematic retest shift did not imply close agreement for individual measurements.

Table 2
Short-term test–retest repeatability of the four distance stereoacuity methods.

Method	n	ICC(A,1) [95% CI]	Lin CCC [95% CI]	Retest/test ratio [95% LoA]	Repeatability factor
COI	65	0.527 [0.320; 0.708]	0.524 [0.317; 0.705]	0.850 [0.202; 3.574]	4.204
Eyenext	63	0.607 [0.435; 0.733]	0.604 [0.431; 0.730]	0.946 [0.271; 3.308]	3.495
Distance Randot	58	0.601 [0.295; 0.780]	0.598 [0.291; 0.777]	0.866 [0.319; 2.347]	2.711
Senaptec	65	0.636 [0.440; 0.796]	0.633 [0.437; 0.794]	0.808 [0.147; 4.443]	5.497

Note. ICC(A,1) denotes the two-way absolute-agreement single-measure intraclass correlation coefficient; CCC, Lin's concordance correlation coefficient; LoA, 95% Bland–Altman limits of agreement. Ratios and LoA are backtransformed from $\log_{10}$ differences. The Eyenext estimates are based on 63 valid pairs.

Differences between stereotests

The four methods differed significantly when compared on their original measurement scales (Friedman $\chi^2(3) = 28.007$, $p < 0.001$; Kendall's $W = 0.153$; $n = 61$). The results of the overall comparison before and after harmonisation are presented in Table 3.

Effect of harmonising measurement range and resolution

Capping all four methods to the common 60–200 arcsec range reduced the overall effect (Friedman $\chi^2(3) = 9.348$, $p = 0.025$; Kendall's $W = 0.051$), and no pairwise comparison remained significant after Holm correction. Once both measurement range and disparity resolution were harmonised to the three-level 60/100/200 arcsec scale, the overall difference was no longer statistically significant ($\chi^2(3) = 6.371$, $p = 0.095$; Kendall's $W = 0.035$). Relative to the original analysis, Kendall's W decreased by approximately 77% (Table 3; Fig. 2).

Table 3
Overall comparison of the four distance stereotests before and after harmonisation.

Analysis	$\chi^2(3)$	p	Kendall W	Interpretation
Native scales	28.007	< 0.001	0.153	Global difference
Common range 60–200 arcsec	9.348	0.025	0.051	Smaller global difference; no Holm-significant pair
Common range + 60/100/200 resolution	6.371	0.095	0.035	No significant global difference

Pairwise comparisons between stereotests

After Holm correction, Distance Randot differed significantly from COI (median difference = 0.222 $\log_{10}$; $p < 0.001$) and Senaptec (Senaptec minus Distance Randot: median difference = − 0.333 $\log_{10}$; $p = 0.003$). The other four comparisons were not statistically significant (Table 4).

Table 4
Pairwise comparisons of distance stereoacuity methods on the original measurement scales. Differences are calculated as method B minus method A.

Comparison	Median difference log10	Clinical ratio B/A	Holm p
Eyenext − COI	0.058	1.143	0.083
Distance Randot − COI	0.222	1.667	< 0.001
Senaptec − COI	−0.056	0.880	0.410
Distance Randot − Eyenext	0.097	1.250	0.126
Senaptec − Eyenext	−0.187	0.650	0.126
Senaptec − Distance Randot	−0.333	0.465	0.003

Note. Differences are expressed in $\log_{10}$(seconds of arc). The ratio is method B divided by method A on the arcsecond scale. P values were adjusted using Holm's method across six comparisons.

Agreement between stereotests

Reducing the differences between group distributions did not yield strong individual agreement. COI and Eyenext showed the highest absolute agreement on the original scales (ICC = 0.665; CCC = 0.661), while the remaining pairwise ICCs ranged from 0.114 to 0.254. Bland–Altman ratio limits were wide, indicating that one test could yield a markedly different threshold from another for the same participant (Table 5).

Table 5
Pairwise agreement between distance stereoacuity methods on original and harmonised scales (n = 61 complete cases).

Pair	ICC / CCC	Native ratio [LoA]	Spearman ρ (Holm p)	Kendall τ-b (Holm p)	Linear weighted κ [95% CI]	Exact / extreme disagreement
COI–Eyenext	0.665 / 0.661	1.173 [0.395; 3.480]	0.689 (< 0.001)	0.596 (< 0.001)	0.477 [0.290; 0.640]	57.4% / 4.9%
COI–Distance Randot	0.253 / 0.250	1.549 [0.354; 6.784]	0.332 (0.045)	0.273 (0.048)	0.196 [−0.004; 0.388]	44.3% / 14.8%
COI–Senaptec	0.213 / 0.210	0.888 [0.103; 7.674]	0.217 (0.280)	0.161 (0.298)	0.112 [−0.112; 0.339]	49.2% / 27.9%
Eyenext–Distance Randot	0.186 / 0.184	1.321 [0.286; 6.090]	0.178 (0.339)	0.157 (0.316)	0.147 [−0.057; 0.344]	44.3% / 16.4%
Eyenext–Senaptec	0.254 / 0.251	0.757 [0.098; 5.849]	0.326 (0.045)	0.258 (0.049)	0.131 [−0.082; 0.333]	42.6% / 23.0%
Distance Randot–Senaptec	0.114 / 0.112	0.573 [0.070; 4.714]	0.173 (0.339)	0.148 (0.316)	0.103 [−0.109; 0.306]	42.6% / 23.0%

Note. Native ratios express the second-named method relative to the first; LoA are 95% Bland–Altman ratio limits. Correlation p values are Holm-adjusted across six pairs. Weighted kappa was calculated for the common three-category scale. Extreme disagreement corresponds to a difference of two categories.

Discussion

The present study demonstrates that differences between distance stereoacuity methods depend partly on the way each instrument samples and reports disparity, but also that removing much of the variation in their numerical scales does not make their results interchangeable. All four methods were administered at the same viewing distance and within the same participants, yet their native-scale threshold distributions differed significantly. Only two of the six adjusted pairwise comparisons—Distance Randot versus COI and Distance Randot versus Senaptec—remained significant, indicating that the overall Friedman result should not be interpreted as a difference between every pair of tests. More importantly, absolute agreement between individual measurements was generally weak. This distinction between group-level differences and individual agreement is central to the clinical interpretation of our findings.

The descriptive results illustrate why comparisons based only on central tendency can be misleading. Median thresholds ranged from 41 arcsec with Senaptec to 100 arcsec with Distance Randot. Because

lower thresholds represent better stereoacuity, these values might appear to imply better performance with some instruments than others. However, each test offers different disparity levels, range limits, stimulus characteristics, and response demands. A difference in a test-derived threshold cannot therefore be attributed to a true difference in the participant's underlying binocular function without an external reference standard. Prior work comparing real-depth and random-dot stereotests and establishing test-specific normative values supports the need to interpret thresholds within the method used.[2, 3]

Sequential harmonisation provided a direct sensitivity analysis of the influence of the reporting scale. On the original scales, the global difference was statistically significant, although its effect size was small (Kendall's W = 0.153). Capping all thresholds below 60 arcsec at 60 and above 200 arcsec at 200 reduced W to 0.051. When values were additionally collapsed into three common categories (60, 100, and 200 arcsec), W decreased further to 0.035 and the Friedman test was no longer statistically significant (p = 0.095). Thus, the between-method contrast was sensitive to the range and granularity of the data.

This pattern is consistent with previous evidence that stereotest design and available disparity steps influence reported thresholds. Vancleef et al. showed that disparity sampling and test procedure contributed to elevated TNO thresholds,[7] while van Doorn et al. demonstrated that different editions of the same test could yield clinically relevant differences.[8] Digital stereotest studies also report effects of measurable range and stimulus properties, including dot size, density, and temporal presentation.[9, 10] These studies provide plausible mechanisms for variation between instruments; they do not identify the causes of the differences observed here. More recently, Cao et al. compared a novel glasses-free distance stereotest with the Distance Randot and reported differences in measured thresholds, despite both methods assessing distance stereopsis.[13] Their findings further support the importance of considering test-specific characteristics when interpreting stereoacuity measurements.

The harmonisation findings require a particular caution. Capping values compresses the observed distributions, and reducing a continuous or multi-step outcome to only three levels necessarily increases ties and discards information. Both operations can reduce rank-based test statistics even if the underlying measurements remain clinically different. Consequently, the absence of a significant global difference after full harmonisation is neither a demonstration of equivalence nor proof that scale architecture accounted for a specific proportion of the original difference. Establishing equivalence would require a prospectively defined clinically acceptable margin and an analysis designed to test that margin.

The agreement analyses address the clinically distinct question of whether a threshold from one method can substitute for a threshold from another in the same person. COI and Eyenext were the most closely related pair: native-scale ICC(A,1) = 0.665 and CCC = 0.661, with Spearman's rho = 0.689. This relationship persisted after capping the range (rho = 0.629), and the three-level classification yielded a linear weighted kappa of 0.477 (95% CI 0.290–0.640), exact agreement in 57.4% of participants, and extreme two-category disagreement in 4.9%. These figures indicate appreciable association but only partial category concordance; they do not justify unrestricted substitution of one method for the other.

Agreement was lower for the other five pairs, with native-scale ICCs of 0.114–0.254 and linear weighted kappas of 0.103–0.196 on the harmonised scale. Moreover, the majority of rank associations did not remain statistically significant after the common-range analysis. Correlation, absolute agreement, and category agreement measure different properties: two tests can rank participants similarly while assigning substantially different thresholds, and a low-resolution categorical scale may conceal differences that matter to monitoring. The present pattern therefore supports instrument-specific interpretation rather than a universal cross-test conversion factor. This accords with earlier reports of differences between clinical stereotests,[3, 14] although agreement estimates from different samples or devices should not be compared as if they were obtained under identical conditions. In contrast, Barnett-Itzhaki et al. reported more favourable agreement between distance stereoacuity methods in children aged 4–17 years.[15] Differences in study populations, instruments, and testing protocols may partly explain these contrasting findings, highlighting the importance of establishing interchangeability for specific instruments and populations.

Variation in binocular image separation may contribute to the observed disagreement. Distance Randot uses polarised viewing, whereas COI, Eyenext, and the frontal Senaptec assessment use red–green separation. Zhao et al. reported differences in distance stereoacuity according to whether red–green, polarised, or active-shutter separation was employed.[11] Nevertheless, separation method alone is insufficient to explain our observations: Senaptec did not show high agreement with COI or Eyenext despite sharing the broad red–green approach, whereas COI and Eyenext agreed most closely. Test-specific stimulus geometry, contrast, response format, and threshold sequence are therefore additional plausible contributors, but they were not experimentally isolated in this study.

None of the four methods showed evidence of a consistent test–retest shift after multiplicity correction (all Holm-adjusted p values > 0.20). This finding should not be confused with high repeatability at the individual level. ICC(A,1) estimates were 0.527 for COI, 0.607 for Eyenext, 0.601 for Distance Randot, and 0.636 for Senaptec. The corresponding Bland–Altman ratio limits were wide, particularly for Senaptec (0.147–4.443) and COI (0.202–3.574). Thus, repeated measurements could differ considerably within an individual even when the group showed no detectable directional change.

Test–retest variability is well recognised in stereoacuity measurement. Argilés et al. demonstrated that enhanced task instructions and visual examples improved stereoacuity outcomes in adults, suggesting that participant understanding and task familiarisation should be considered when interpreting test–retest variability.[16] Wang et al. found strong short-term reproducibility of the Distance Randot within its available disparity steps in a normative sample,[4] while Adams et al. examined how much change must occur before it can confidently be distinguished from measurement variability. [17] Antona et al., Mehta and O'Connor, and Ale Magar et al. also documented differences in repeatability across stereotests.[14, 18, 19] Direct numerical comparisons with our ICCs would be inappropriate because the study populations, threshold sequences, retest intervals, and reliability statistics differ. The consistent lesson is that a non-significant paired test does not guarantee that an observed change in one person is meaningful.

For clinical follow-up, these findings argue against treating values reported in seconds of arc as interchangeable across instruments. The importance of establishing appropriate reference values is

illustrated by Jorge et al., who reported normative distance stereoacuity values in professional football players. Although their study did not compare different stereotests, it highlights the importance of clearly defined measurement procedures when interpreting stereoacuity thresholds across populations.[20] Furthermore, Jorge and Jorge reported an association between poorer distance stereopsis and reduced dynamic visual acuity in professional football players, reinforcing the functional relevance of distance stereoacuity assessment. Together, these findings emphasise the importance of obtaining reliable and interpretable measurements, particularly when stereoacuity is monitored over time.[21] An apparent improvement or deterioration following a change of stereotest may reflect methodological differences rather than a change in binocular sensory status. Whenever feasible, clinicians should retain the same stereotest, viewing distance, test version, optical separation system, and threshold procedure across visits. If a change of instrument is unavoidable, the new measurement should be treated as a new method-specific baseline, not converted using a fixed correction factor.

For research, the manufacturer, model or software version, minimum and maximum available disparities, step sequence, and decision rule should be reported explicitly. Studies combining data from different instruments should account for method as a potential source of heterogeneity. Group-level comparisons alone are insufficient to establish that two tests can be substituted at the individual level; agreement analyses and prespecified clinical tolerance limits are required.[12]

The within-participant design, randomised test order, common 3 m viewing distance, and repeated measurements strengthened the comparison. The use of a common complete-case sample (n = 61) for the main four-method analysis avoided attributing differences to unequal participant composition. Assessing distributional differences, repeatability, absolute agreement, rank association, and ordinal agreement separately provided a broader picture than any single statistic.

Several limitations temper interpretation. The sample was relatively small and consisted of young adults with generally good vision; the findings may not generalise to children, older adults, or patients with amblyopia, strabismus, or poor stereopsis. Four participants lacked a valid first measurement for at least one method, and the main between-method analysis was confined to the 61 complete cases; repeatability sample sizes also differed across instruments. Analyses of measurable responses may not fully reflect differences in testability at the limits of each device. The 10-minute interval estimates immediate repeatability, not day-to-day or longer-term stability. No independent reference standard was available to establish which test was most accurate. Finally, the instruments varied simultaneously in several design features, preventing causal attribution of their disagreement to range, resolution, binocular separation, or any other single component. The three-level categorisation was deliberately coarse and should be viewed solely as a sensitivity analysis, not as a proposed clinical classification.

Future work should assess the same methods in larger, more diverse samples, including participants with reduced stereoacuity, and use longer retest intervals to estimate clinically relevant measurement variability. Experimental comparisons that hold the stimulus constant while altering one feature at a time —such as disparity sequence, binocular separation, or response procedure—would help disentangle the mechanisms of test-specific thresholds. Equivalence or interchangeability should be evaluated against clinically justified margins rather than inferred from non-significant difference tests. Where discrete test

levels constrain the observations, ordinal or interval-censored approaches may complement conventional analyses.

Conclusions

COI, Eyenext, Distance Randot, and frontal Senaptec yielded significantly different distance stereoacuity distributions on their native scales. Capping thresholds to a common range and then reducing them to three shared categories attenuated the global differences, but this finding does not establish equivalence. Individual agreement remained limited, including for COI and Eyenext, the closest pair. These instruments should not be considered interchangeable for individual assessment, and longitudinal evaluations should use a consistent method and protocol.

Declarations

Declarations

Ethics approval and consent to participate. The study was approved by the Ethics Subcommittee for Life and Health Sciences of the University of Minho (CEICVS 113/2020). Written informed consent was obtained from all participants.

Funding. This work was supported by the Portuguese Foundation for Science and Technology (FCT) in the framework of the Strategic Funding UID/04650/2025 (https://doi.org/10.54499/UID/04650/2025).

Competing interests. The authors declare no competing interests.

Data availability. The participant-level dataset is not publicly archived. Requests for de-identified data and the analysis syntax will be considered by the corresponding author in accordance with institutional approvals and applicable data-protection requirements.

Author contributions. Jorge Jorge: conceptualisation, supervision, statistical methodology, manuscript review and editing. Miguel Coelho: investigation, data curation and original draft preparation. João Pedro Jorge: analysis and interpretation of results, critical review and editing of the manuscript. All three authors reviewed and approved the submitted manuscript.

References

1. 1. O'Connor AR, Tidbury LP. Stereopsis: are we assessing it in enough depth? Clinical and Experimental Optometry [Internet]. Blackwell Publishing Ltd; 2018. p. 485–94. https://doi.org/10.1111/cxo.12655
2. 2. Piano MEF, Tidbury LP, O'Connor AR. Normative Values for Near and Distance Clinical Tests of Stereoacuity. Strabismus. 2016;24(4):169–72. https://doi.org/10.1080/09273972.2016.1242636

3. 3. Leske DA, Birch EE, Holmes JM. Real depth vs randot stereotests. Am J Ophthalmol. 2006;142(4):699–701. https://doi.org/10.1016/j.ajo.2006.04.065
4. 4. Wang J, Hatt SR, O'Connor AR, Drover JR, Adams R, Birch EE, et al. Final version of the Distance Randot Stereotest: normative data, reliability, and validity. J AAPOS. 2010;14(2):142–6. https://doi.org/10.1016/j.jaapos.2009.12.159
5. 5. Kim J, Yang HK, Kim Y, Lee B, Hwang JM. Distance stereotest using a 3-dimensional monitor for adult subjects. Am J Ophthalmol. 2011;151(6):1081–1086.e1. https://doi.org/10.1016/j.ajo.2010.09.034
6. 6. Ma DJ, Yang HK, Hwang JM. Reliability and validity of an automated computerized visual acuity and stereoacuity test in children using an interactive video game. Am J Ophthalmol. 2013;156(1):195–201.e1. https://doi.org/10.1016/j.ajo.2013.02.018
7. 7. Vancleef K, Read JCA, Herbert W, Goodship N, Woodhouse M, Serrano-Pedraza I. Overestimation of stereo thresholds by the TNO stereotest is not due to global stereopsis. Ophthalmic Physiol Opt. 2017;37(4):507–20. https://doi.org/10.1111/opo.12371
8. 8. van Doorn LLA, Evans BJW, Edgar DF, Fortuin MF. Manufacturer changes lead to clinically important differences between two editions of the TNO stereotest. Ophthalmic Physiol Opt J Br Coll Ophthalmic Opt. 2014;34(2):243–9. https://doi.org/10.1111/opo.12101
9. 9. McCaslin AG, Vancleef K, Hubert L, Read JCA, Port N. Stereotest Comparison: Efficacy, Reliability, and Variability of a New Glasses-Free Stereotest. Transl Vis Sci Technol. 2020;9(9):29. https://doi.org/10.1167/tvst.9.9.29
10. 10. Read JCA, Wong ZY, Yek X, Wong YX, Bachtoula O, Llamas-Cornejo I, et al. ASTEROID stereotest v1.0: lower stereo thresholds using smaller, denser and faster dots. Ophthalmic Physiol Opt. 2020;40(6):815–27. https://doi.org/10.1111/opo.12737
11. 11. Zhao L, Zhang Y, Wu H, Xiao J. The difference of distance stereoacuity measured with different separating methods. Ann Transl Med. 2020;8(7):468. https://doi.org/10.21037/atm.2020.03.73
12. 12. Bland JM, Altman DG. Measuring agreement in method comparison studies. Stat Methods Med Res. 1999;8(2):135–60. https://doi.org/10.1177/096228029900800204
13. 13. Cao LQ, Cui B, Li XL, Zhou BY, Qin LW, Li MG, et al. Novel Glasses-Free Distance Stereotest Using Eye Tracking: Evaluation of Sensitivity, Validity, Reliability, and Monocular Cue Masking. J Multidiscip Healthc. 2025;18:1791–801. https://doi.org/10.2147/JMDH.S511165
14. 14. Antona B, Barrio A, Sanchez I, Gonzalez E, Gonzalez G. Intraexaminer repeatability and agreement in stereoacuity measurements made in young adults. Int J Ophthalmol. 2015;8(2):374–81. https://doi.org/10.3980/j.issn.2222-3959.2015.02.29
15. 15. Barnett-Itzhaki G, Barnett-Itzhaki Z, Ela-Dalman N, Mezad-Koursh D. Evaluating distance stereoacuity in children 4–17 years of age with a novel digital application. J AAPOS. 2024;28(3):103904. https://doi.org/10.1016/j.jaapos.2024.103904
16. 16. Argilés M, Vinuela-Navarro V, Pérez-Maña L, Alonso-Matarin S, Asensio-Jurado L, Rovira-Gay C, et al. Effects of task explanation on stereoacuity measurements using a random-dot stereotest. J

Optom. 2026;19(4):100625. https://doi.org/10.1016/j.optom.2026.100625

17. 17. Adams WE, Leske DA, Hatt SR, Holmes JM. Defining real change in measures of stereoacuity. Ophthalmology. 2009;116(2):281–5. https://doi.org/10.1016/j.ophtha.2008.09.012
18. 18. Mehta J, O'Connor A. Test retest variability in stereoacuity measurements. Strabismus. 2023;31(3):188–96. https://doi.org/10.1080/09273972.2023.2252853
19. 19. Ale Magar JB, Shah SP, Sleep MG, Willett FA, Dai SH. Validity and repeatability of contour-based visotec distance stereoacuity test. Clin Exp Optom. 2023;106(3):283–9. https://doi.org/10.1080/08164622.2022.2033599
20. 20. Jorge J, Fuste R, Sousa R. Normative visual performance standards in male professional football players: a comprehensive assessment of visual acuity, binocular function, and visuo-cognitive abilities. Clin Exp Optom. 2026;109(5):998–1006. https://doi.org/10.1080/08164622.2025.2610973
21. 21. Jorge J, Jorge JP. Relationship between dynamic visual acuity and static visual acuity, refractive error, and binocular vision in elite soccer players. Clin Exp Optom. 2024;107(8):820–5. https://doi.org/10.1080/08164622.2024.2301981

Figures

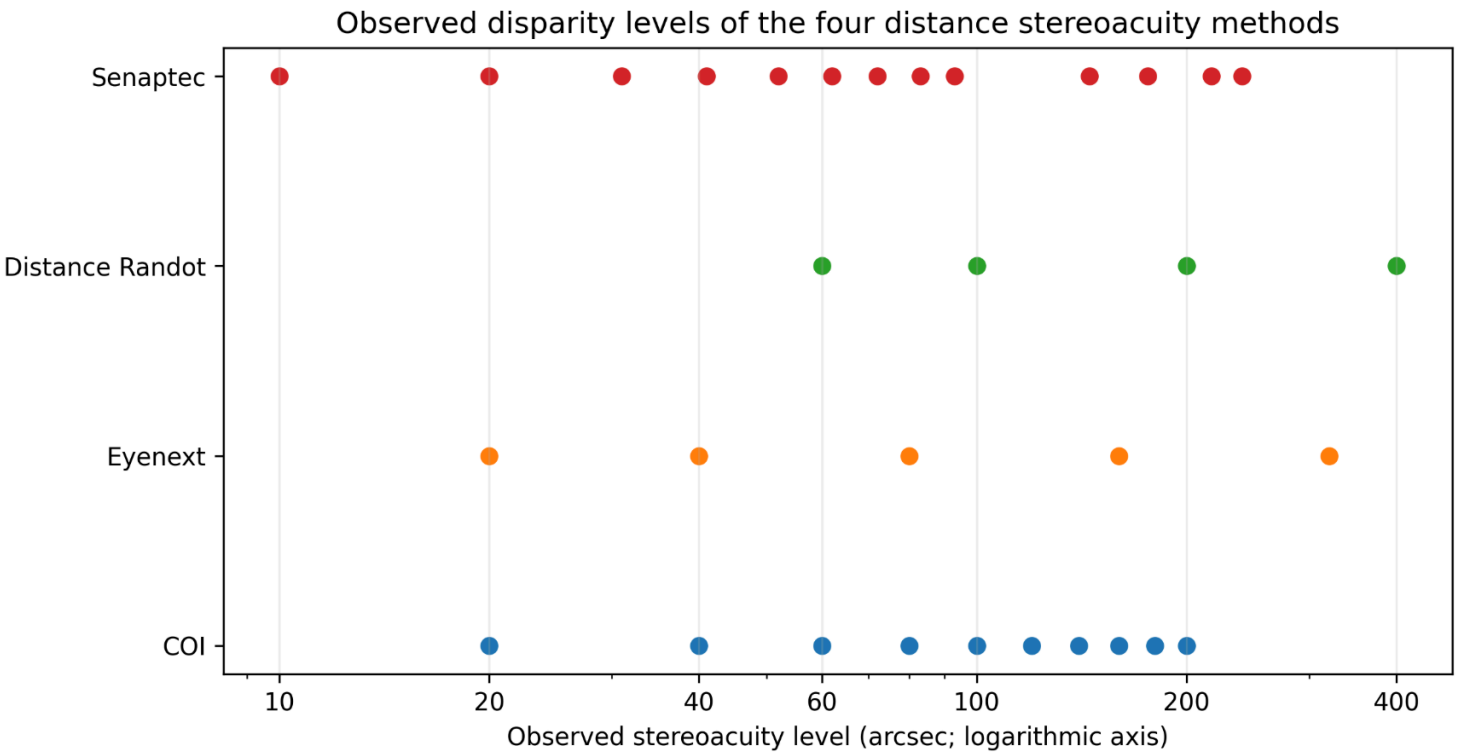


Figure 1

Observed disparity levels for the four distance stereotests during initial measurement (logarithmic axis).

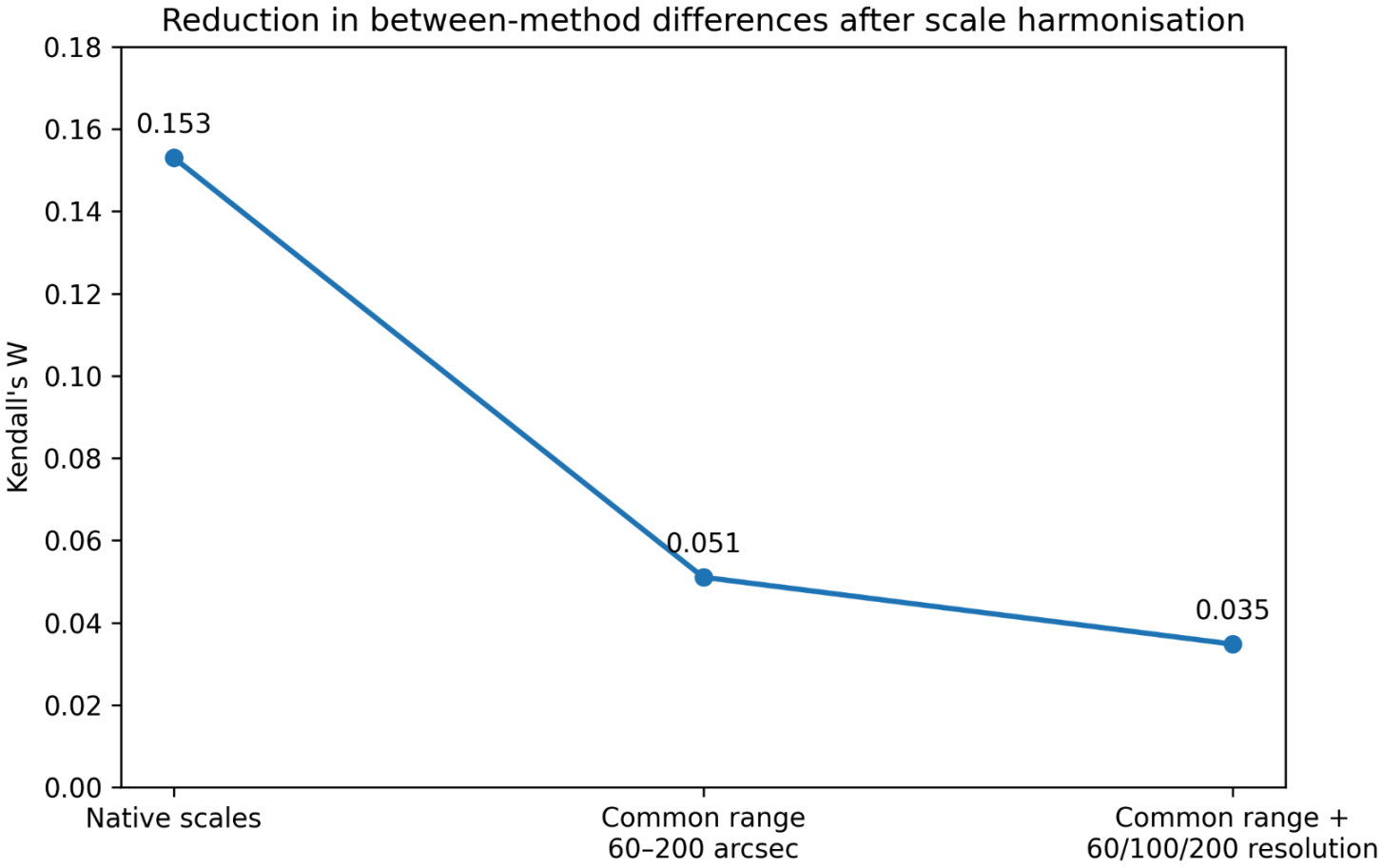


Figure 2

Kendall’s W before and after harmonisation of measurement range and disparity resolution.

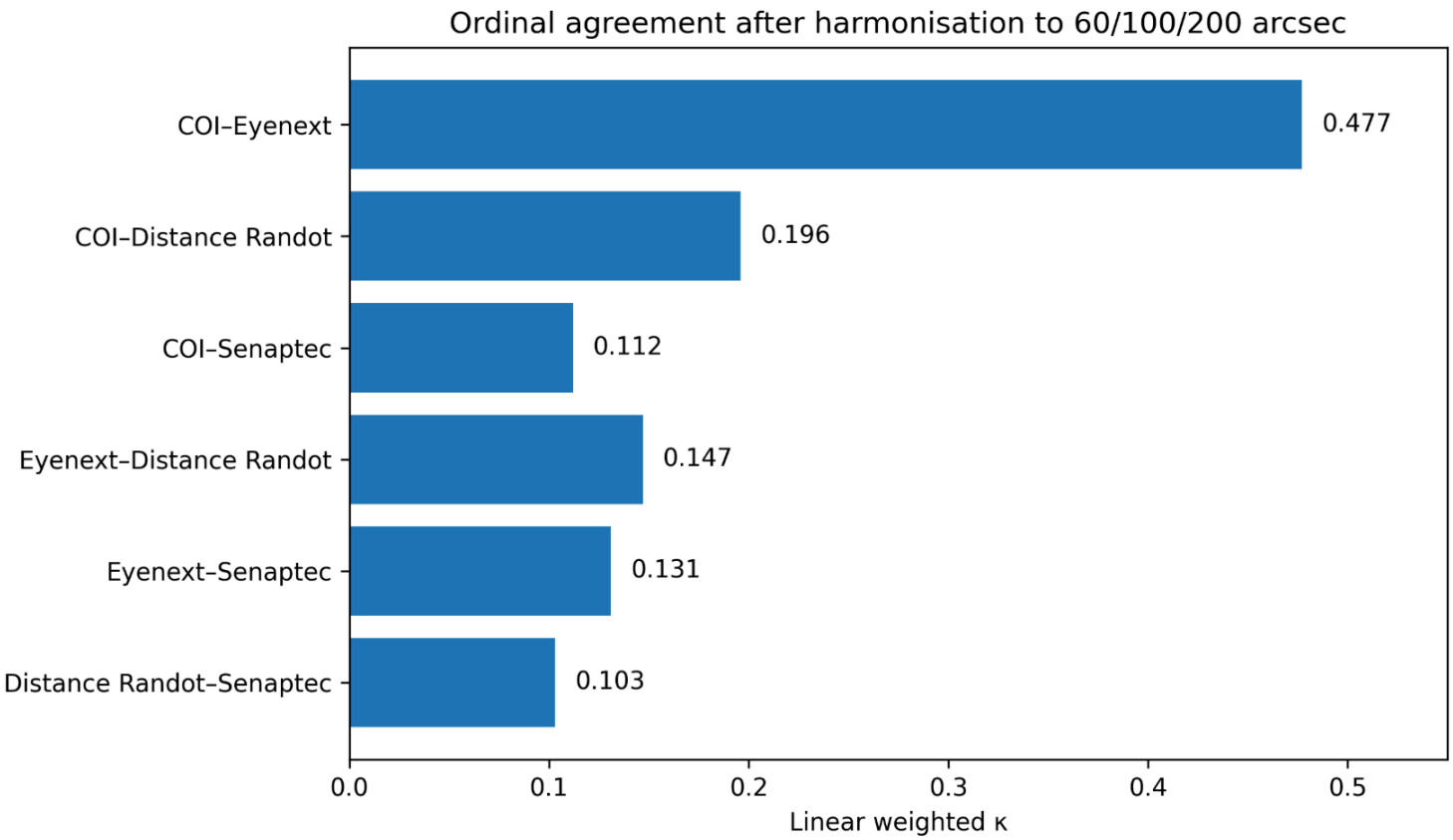


Figure 3

Linear weighted kappa for each test pair after harmonisation to the common three-category scale.