Journal of Optometry 14 (2021) 275–281

Journal of Optometry

www.journalofoptometry.org

ORIGINAL ARTICLE

The repeatability and reproducibility of four techniques for measuring horizontal heterophoria: Implications for clinical practice

Nicola S. Anstice[a,b,*], **Bianca Davidson**[b], **Bridget Field**[b], **Joyce Mathan**[b], **Andrew V. Collins**[b], **Joanna M. Black**[b]

[a] *Discipline of Optometry and Vision Science, The University of Canberra, Australia*
[b] *School of Optometry and Vision Science, The University of Auckland, New Zealand*

Received 19 January 2020; accepted 25 May 2020
Available online 12 August 2020

KEYWORDS
Binocular vision;
Heterophoria;
Reliability;
Repeatability;
Cover test

Abstract
Purpose: Convergence insufficiency, the most common binocular vision anomaly, is characterised by a receded near point of convergence and an exophoria which is at least 4 prism dioptres (Δ) larger at near than at distance. However, the repeatability of standard heterophoria measures are poorly understood. This study assessed the ability of four common heterophoria tests to detect differences of 4Δ by evaluating the inter- and intra-examiner variability of the selected techniques.
Methods: Distance and near horizontal heterophorias of 20 visually-normal adults were measured with the alternating prism cover test, von Graefe prism dissociation, Howell Card and Maddox Rod by two examiners at two separate visits using standardised instructions and techniques. We investigated inter- and intra-examiner variability using repeatability and reproducibility indices, as well as Bland-Altman analysis with acceptable limits of agreement defined as ±2Δ.
Results: The Howell card test had the lowest intra-examiner variability at both distance and near, as well as the best 95% limits of agreement (±1.6Δ for distance and ±3.7Δ for near). Inter-examiner reproducibility results were similar, although at near the alternating prism cover test had better repeatability (1.1Δ, 95% confidence intervals −1.1Δ to 4.0Δ) than the Howell card (1.4Δ, 95% confidence intervals −1.9Δ to 5.9Δ).

* Corresponding author at: Discipline of Optometry and Vision Science, University of Canberra, 11 Kirinari Street, Bruce, Australian Capital Territory, Australia.
E-mail address: nicola.anstice@canberra.edu.au (N.S. Anstice).

https://doi.org/10.1016/j.optom.2020.05.005

Conclusion: The low repeatability of many standard clinical heterophoria tests limits the ability to reliably detect a 4△ difference. The Howell Card provided the most repeatable and reproducible results indicating that this technique should be used to detect small changes in heterophoria magnitude and direction.

Introduction

Eye care professionals commonly use heterophoria measures to assess binocular vision in both adults and children. Clinically, heterophoria can be measured by many techniques which use varying stimuli, methods of dissociation, and patient instructions. Although some authors argue that heterophoria measurement is not useful in the diagnosis of decompensating binocular vision disorders,[1] these tests are routinely employed as part of the diagnostic criteria for many binocular vision conditions. Convergence insufficiency is a prevalent binocular disorder affecting 2–8% of the population,[2,3] and may negatively affect reading[4] and other academic behaviours.[5] Clinicians typically diagnose convergence insufficiency using a combination of symptoms,[6] receded near point of convergence, reduced positive fusional vergence reserves, and an exophoria which is greater at near than at distance, by at least 4 prism dioptres (△).[7] Recently, the Binocular Vision Anomalies and Normative Data (BAND) study group suggested a difference between distance and near heterophorias of >1.25△ produces the highest sensitivity and specificity for detecting children with non-strabismic binocular vision anomalies.[8] However, this raises the question of whether it is possible to reliably detect small differences in heterophoria by a single examiner over time or between two examiners assessing the same patient.

Several previous studies[9–11] have examined the precision and accuracy of dissociated heterophoria tests, but these varied in the clinical techniques employed, the statistical analysis undertaken and whether inter-examiner, intra-examiner or inter-test variance was investigated. In this study, we examined the reliability and reproducibility of four clinical techniques which measure horizontal heterophoria: the alternating prism cover test, the von Graefe method, the Howell Card and the Maddox rod. In particular, we investigated the measurement agreement over time and between examiners, and whether this agreement fell within our *a priori* limit of 4△, the critical diagnostic value used by the Convergence Insufficiency Treatment Trial (CITT).[7]

Materials and methods

We conducted a prospective, examiner- and participant-masked, randomised observational study which investigated the repeatability and reproducibility of horizontal heterophoria measurements using four common clinical techniques: alternating prism cover test, von Graefe technique, Howell Card and Maddox Rod. The study was approved by the University of Auckland Human Participants Ethics Committee, in accordance with the tenets of the Declaration of Helsinki.

Based on conservative estimates of published Intraclass Correlation Coefficients (ICCs),[12] sample size calculations indicated 15 participants was sufficient to obtain 80% power with alpha = 0.05.[13] The number of participants was calculated based on two observations per subject to allow both inter-examiner and intra-examiner reliability to be calculated independently. To allow for loss-to-follow-up, we recruited a larger sample of 20 adults (>18 years of age) from the Optometry student population over seven months in 2017. After obtaining informed written consent, participants completed a short clinical history to ensure recruits were visually normal adults with no history of eye surgery or trauma, amblyopia or strabismus, ocular pathology or neurological disorders. All participants were required to have distance visual acuity (VA), of at least 0.0 logMAR (Snellen equivalent 6/6) in each eye with habitual correction, as measured on the Medmont AT-20R chart (Medmont International PTY LTD. Victoria, Australia).

Participants attended two measurement sessions; at the initial session, the two examiners assessed each participant independently using identical techniques, performed in the same clinical examination room with predetermined scripted instructions. Each participant returned for a second session within ten days, at which time the measures were repeated in the same manner and order by the same two examiners. Both examiners (authors BF and JM) were final year Optometry students with a minimum of nine months clinical training and had passed an advanced binocular vision course. A third researcher (author BD) collated all data, allowing examiners to remain masked both to their initial findings (when conducting the second session) and to each other's results. The examiner order and test sequence were randomised for each participant at the first session, although distance heterophorias were always assessed before near heterophorias. The order of tests and examiners was repeated in the second session. In all tests, except Maddox Rod, the examiners instructed participants to keep the target clear, and to report blurred vision if present, to help control accommodation.

Variables measured

The participants wore their habitual spectacle correction in a trial frame (for alternating prism cover test and Howell card) or phoropter (for von Graefe and Maddox Rod), and fixated a distance or near target while the examiner measured heterophoria using standardised clinical techniques.[14] Measurements were performed at 6 m and 40 cm, except for the Howell Card (Bernell Corp, South Bend, IN) which was used at the calibrated testing distances of 3 m and 33 cm. Measures were undertaken in full room illumination (>750 lx), except for the Maddox Rod test which was performed in dim illumination (50—100 lx) to enhance the visibility of the red streak target. For both the near Howell Card and near cover test, target distance (33 cm and 40 cm, respectively) was measured before commencing testing at each session.

For the alternating prism cover test, participants viewed a single crowded 0.1 logMAR letter at distance and N8 letter at near. The examiner observed the participant's eye movements and neutralised the deviation using a prism bar in front of the right eye. If no movement was observed, examiners added base in and base out prism until a deviation was observed and the midpoint of these prism measurements was recorded. Von Graefe testing was performed using a vertical line of letters (0.0—0.3 logMAR at distance and N8 at near), and Risley prisms. Initially, a 12△ base in prism was inserted over the right eye and 6△ base up prism over the left eye for five seconds to allow complete dissociation prior to heterophoria measurement. The base in prism magnitude was reduced until alignment was reported. Prism was added beyond neutralisation, then reduced to the alignment point again and the mean of these two results was calculated. In cases where the deviation was greater than 12△ exophoria, the prism was increased in a base in direction instead. For the Howell Card, a 6△ base down prism was placed in front of the right eye and participants were asked to report which number the arrow pointed to on the lower chart image. Even numbers (blue side) represented an exophoria, while odd numbers (yellow side) represented an esophoria. The Maddox Rod test was performed with a pen torch and the Maddox rod lens in front of the right eye with the cylinders oriented horizontally. The participant reported where the vertical red streak was positioned relative to the white light. If the white light and vertical red streak were not overlapping, Risley prisms were increased in magnitude in the appropriate direction until alignment was achieved.

The primary outcome was the inter- and intra-examiner repeatability of the selected tests for distance and near heterophoria measurement in young adults.

Statistical analysis

A univariate General Linear Model (GLM), with heterophoria measure as the dependent variable, and independent variables of test, examiner and session, was performed to evaluate differences in the mean heterophoria measure for each test at both distance and near separately. When comparing inter-examiner reliability, we took the mean of each examiner's measures across the two sessions, while for estimates of intra-examiner reliability we used only Examiner 1's measures. Intra-examiner repeatability and inter-examiner reproducibility were assessed using the within-subject standard deviation and associated 95% confidence limits on one-way analysis of variance (ANOVA).[15] Intraclass correlation coefficients (ICCs), and their 95% confidence intervals, were used to assess the intra- and inter-rater reliability[16] using a two-way mixed absolute agreement model.[17] Bland Altman analysis[18] was performed to determine the limits of agreement between each test, separately for distance and near measures. The agreement was summarized by examining the bias, standard deviation of differences, 95% limits of agreement (LoA) and confidence limits around the LoA.[19,20] The maximum acceptable 95% limits LoA were defined *a priori* as ±2.00△ based on the suggested minimum detectable eye movement under usual conditions[21] and the CITT criteria for diagnosing convergence insufficiency. Statistical analysis was performed in SPSS Statistical Package version 25 (IBM Corp, Amorak, NY, USA) and in all analyses, negative values indicate exophoria, while positive values indicate esophoria.

Results

Participants

We recruited 20 visually-normal young adults (12 females, 8 males, age range 22–26 years) from the student population of the School of Optometry and Vision Science at the University of Auckland. All participants completed all heterophoria measures at both sessions. The minimum time between sessions was one day and all participants completed the second session within 10 days. Mean heterophoria measures across all participants were small and within Morgan's normative values (Fig. 1). One participant had a moderate exophoria (4—8△) at distance, while two participants had large near exophorias (12—20△ and 10—17△, respectively).

Comparison of tests

Univariate GLM found no difference between distance heterophoria measures taken by Examiner 1 and Examiner 2 ($p = 0.668$) and no difference between Session 1 and Session 2 ($p = 0.474$). However, there was a significant effect of heterophoria test ($p = 0.013$); post-hoc Bonferroni corrected comparisons showed a significant difference between the alternating prism cover test and Maddox Rod results (mean difference = −1.1△, $p = 0.015$). There were no significant two- or three-way interactions. Near heterophoria results were similar, with no differences between examiners ($p = 0.114$), sessions ($p = 0.959$) or tests ($p = 0.640$). There were no significant two- or three-way interactions.

Estimates of repeatability and reproducibility

As there was no difference in heterophoria measures between Examiner 1 and Examiner 2, intra-examiner repeatability was estimated by comparing Session 1 and Session 2 results for Examiner 1. The mean intra-examiner variability was 0.9△ (95% CI −1.7△ to 3.4△) for distance and 1.8△ (95% CI −1.6△ to 8.2△) for near. The Howell card technique had the lowest intra-examiner variability, while the

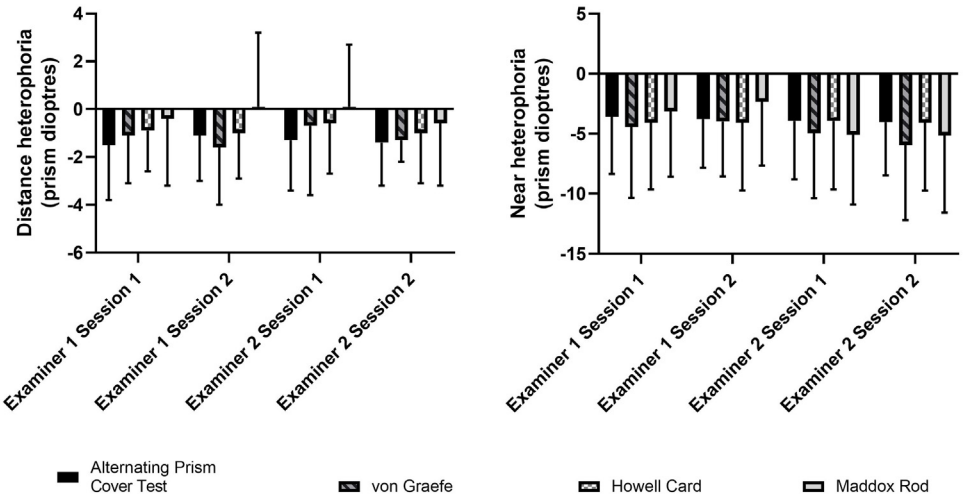

Figure 1 Mean values (± standard deviation) of horizontal heterophoria (in prism dioptres) measured using the alternating prism cover test, von Graefe, Howell Card and Maddox Rod at distance (left) and near (right).

Table 1 Intra-examiner repeatability and inter-examiner reproducibility estimated using within-subjects variability on One-way ANOVA.

Heterophoria measure	Test distance	Intra-examiner repeatability (Δ) (95% confidence limits)	Inter-examiner reproducibility (Δ) (95% confidence limits)
Alternating prism cover test	6 m	0.9 (−1.7 to 3.3)	0.6 (−1.3 to 1.9)
	40 cm	1.5 (−1.9 to 6.7)	1.1 (−1.1 to 4.0)
Von Graefe	6 m	1.1 (−1.9 to 4.4)	0.8 (−1.5 to 2.7)
	40 cm	2.3 (−1.1 to 11.7)	2.0 (−1.6 to 9.5)
Howell Card	3 m	0.6 (−1.3 to 1.9)	0.4 (−1.0 to 1.3)
	33 cm	1.3 (−1.9 to 5.2)	1.4 (−1.9 to 5.9)
Maddox Rod	6 m	1.0 (−1.1 to 4.0)	0.6 (−1.3 to 2.0)
	40 cm	1.9 (−1.6 to 9.1)	2.1 (−1.4 to 10.3)

von Graefe method had the highest (Table 1). Inter-examiner variability (estimated by comparing the mean results of Session 1 and Session 2 for each of the two examiners) showed a similar pattern to the intra-examiner results. Howell Card measures showed the lowest inter-examiner variability at distance, while alternating prism cover test had the lowest variability at near; the von Graefe and the Maddox Rod methods showed the highest inter-examiner variability at distance and near, respectively.

Reliability of intra- and inter-examiner measurements were also assessed using ICCs and their corresponding 95% CIs. A high degree of reliability was found for all heterophoria measurement techniques. Intra-examiner repeatability for Examiner 1 showed that the Howell card had the highest repeatability, while the von Graefe technique had the lowest (Table 2). The highest inter-examiner variability was seen with the Maddox Rod test at near (95% CIs 0.131–0.965).

Bland Altman analysis showed overall bias was small (<2Δ) for all heterophoria measures (Fig. 2); however, LoAs for most techniques fell outside our predefined criterion of ±2Δ except the Howell Card at distance (both intra- and inter-examiner comparisons) and Maddox Rod at distance (inter-examiner comparison). All near heterophoria measures had LoAs outside ±2Δ, with the best agreement between Examiner 1 and Examiner 2's Howell card measurements (-3.3Δ to 3.0Δ).

Discussion

Understanding reliability and repeatability allows practitioners and researchers to recognise the amount of inherent variability in any technique and how this may affect the interpretation of clinical findings.[22] One common criterion for diagnosing convergence insufficiency is a difference of ≥4Δ between distance and near heterophoria measures,[12] while more recently the BAND study has suggested a threshold of >1.25Δ[8] difference for detecting non-strabismic binocular vision disorders. But how realistic are these criteria? While ICCs for heterophoria tests were high, reliability estimates, using within-subjects repeatability and Bland Altman analyses, found 95% confidence intervals and limits of agreement outside ±2Δ which may limit the diagnostic accuracy of these techniques. Our data supports that of Holmes et al., who found high (>0.94) inter-rater ICCs

Table 2 Intraclass Correlation Coefficients (ICCs) between Session 1 and Session 2 (Examiner 1) and between Examiner 1 and Examiner 2 (average of measures of two measures).

Heterophoria measure	Test distance	Intra-examiner repeatability ICC (95% confidence limits)	Inter-examiner repeatability ICC (95% confidence limits)
Alternating prism cover test	6 m	0.877 (0.720–0.949)	0.914 (0.796–0.965)
	40 cm	0.929 (0.829–0.971)	0.948 (0.875–0.979)
Von Graefe	6 m	0.777 (0.519–0.906)	0.850 (0.655–0.937)
	40 cm	0.783 (0.529–0.908)	0.860 (0.661–0.944)
Howell Card	3 m	0.804 (0.570–0.918)	0.903 (0.775–0.960)
	33 cm	0.9612 (0.906–0.984)	0.984 (0.960–0.994)
Maddox Rod	6 m	0.907 (0.780–0.962)	0.952 (0.884–0.981)
	40 cm	0.899 (0.764–0.959)	0.867 (0.131–0.965)

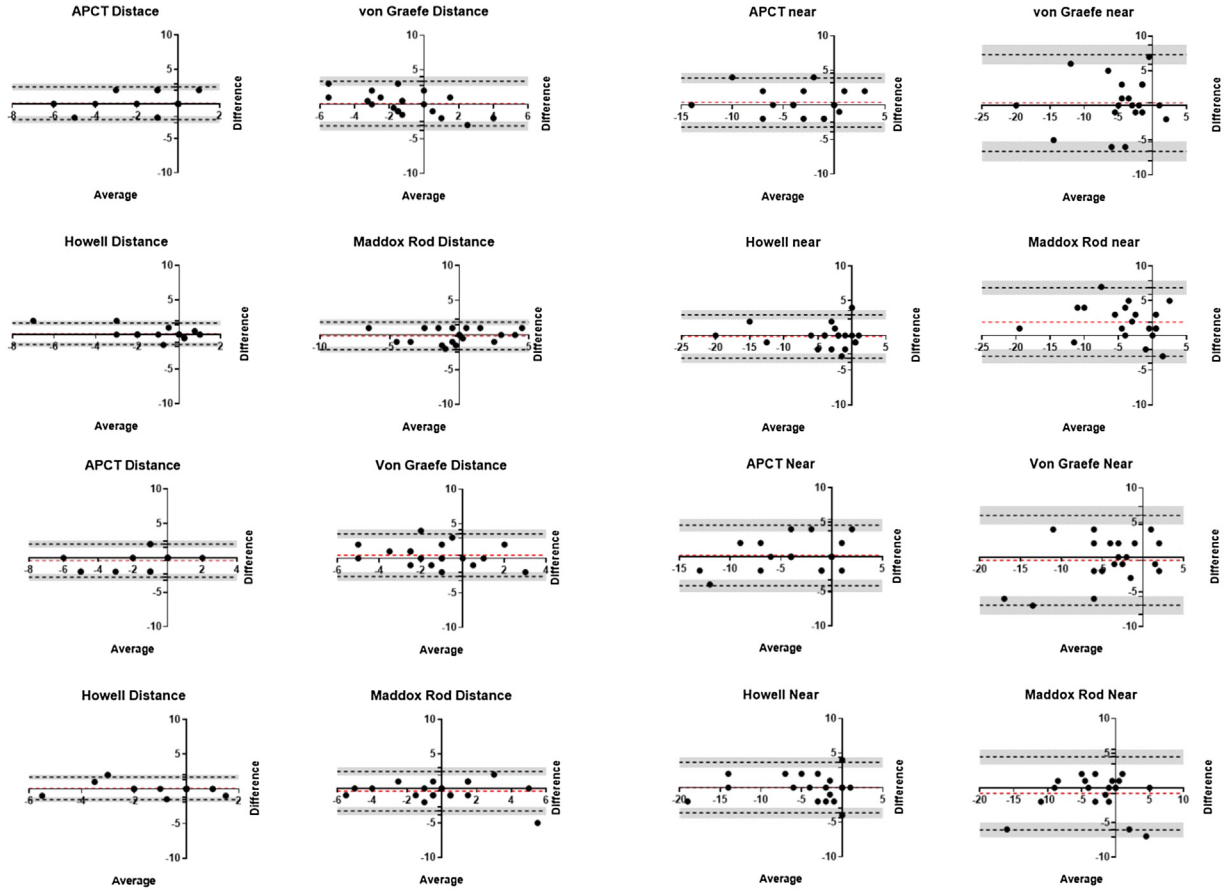

Figure 2 Bland Altman plots showing mean difference between measures (red dashed lines), 95% limits of agreement (black dashed line) and 95% confidence intervals of limits of agreement (grey shaded zone). The comparisons are between Examiner 1 and Examiner 2 at distance (top left) and near (top right) for Session 1, and between Session 1 and Session 2 for Examiner 1 at distance (bottom left) and near (bottom right).

for both simultaneous and alternating prism cover test measures[23] but wide 95% LoAs for both measures (±6.3Δ to ±10.2Δ). The authors concluded that a difference of at least 10Δ was needed to represent a real change. However, it is important to note that participants in the Holmes et al. study were 23 patients with sixth nerve palsy and three controls, whereas we recruited participants with normal binocular vision and results between the two should be compared with caution. Conversely, other studies have suggested much smaller differences can be reliably detected. Johns et al.[9] found the alternating prism cover test to have intra- and inter-examiner repeatability of <0.5Δ, while Rainey et al.[24] suggested a precision limit of 1.2–2.0Δ. While our estimates for repeatability using one-way ANOVA within-subject vari-

ance results for all heterophoria tests were 0.5–2.5△, the 95% repeatability limits were much higher.

Comparing the four techniques we investigated, the Howell Card produced the most repeatable heterophoria measures both between examiners and between visits. Other studies support this result with the Howell card providing more consistent results than von Graefe measures,[10,11,25] but more variable measures than the Muscle Imbalance Measure[25] and Thorington test,[11,26] two heterophoria tests not included in our research. There are several potential reasons for the improved repeatability compared with the other tests used in this study. As the Howell card works on the same optical principles as the Thorington technique, it is perhaps unsurprising in our test battery that this procedure was also the most repeatable. Furthermore, the Howell card was the only heterophoria test where the participants were aware of their previous results and it is possible they remembered the numbers reported and had a bias towards reporting the same number in further assessments.[11] Remembered answers would have been particularly true for measures completed on the same day, for example between Examiner 1 and Examiner 2 in each session, rather than between measures completed between Session 1 and Session 2. However, this is not supported by either the ICCs or the LoA on the Bland Altman plots as these are similar for both inter-examiner and intra-examiner variability.

One of the strengths of this study was the use of a strict testing protocol including scripted participant instructions employed by both examiners across each session to minimise the variability in administering the tests. The research was conducted in the same room for all sessions to ensure consistency in test conditions. Testing and examiner order were randomised for each participant to reduce learning and fatigue effects, while the same testing order was repeated between Session 1 and Session 2 to better compare repeatability over time. Nevertheless, several other factors can also affect the repeatability of heterophoria measures including testing distance (the Howell card was presented at 3 m and 33 cm), mode of dissociation,[27] and use of a trial frame versus phoropter.[28] In this study, the use of a phoropter for the von Graefe and Maddox Rod methods may have contributed to the greater variability, particularly at near, seen with these techniques and our results support the conclusion of Casillas and Rosenfield[28] who recommend that heterophoria measures should be conducted in free space. Free space measurements may allow for a better peripheral field of view, and thus peripheral fusion lock, increasing fusional amplitude for participants with normal binocular vision. Likewise, a larger target size, such as the number targets on the Howell Card, may recruit more of the peripheral retina which may also improve fusional vergence.

Despite the use of strict testing protocols and written examiner instructions, near heterophoria measures were more variable than distance measures, suggesting clinicians need to appropriately control accommodation in clinical heterophoria measures to minimise variability. Participants were instructed to always keep the target clear during testing, except for the Maddox Rod which utilized a non-accommodative target. Near testing with the Maddox Rod produced the most variability across all analyses, mostly likely associated with the absence of consistent accommodation control through visual feedback of blur. This result reinforces the importance of appropriate control of accommodation to reduce instability of the accommodative response and therefore the variability in near heterophoria measures.[26]

Study limitations

The examiners in this study were final year optometry students who may be considered novice practitioners and thus may have more variable results that experienced clinicians. However, previous research suggests there are no clinically significant differences in the variability of heterophoria measures between expert and less experienced clinicians.[29,30] While repeatability of heterophoria measures may be unaffected by practitioner experience, Hrynchak et al.[30] reported novice clinicians found larger deviations, possibly because they performed tests more slowly allowing full dissociation before measuring the heterophoria magnitude.[29] Likewise, participants in this study were final year optometry students and thus trained observers. Therefore, it is more likely that our results can be attributed to the variability of the techniques themselves rather than inexperienced participants altering their criteria but does mean our findings cannot be directly applied to the general population. In clinical settings, it is likely that the intra- and inter-examiner variability is larger than seen in these study results as critical elements, such as maintaining target clarity, may be more difficult to control in untrained observers.

As this primary goal of this study was to investigate intra- and inter-examiner reliability, the sample size was calculated based on estimates of ICCs from previous heterophoria studies, and using two measures from each participant. Sample size estimates would have differed had the method of agreement analysis been the primary outcome measure as this primarily investigates whether there is the same bias throughout the range of measurement values encountered in a patient population. Based on our results, the sample size required for Bland Altman analysis, using $p = 0.05$, power = 80% and a maximum allowable difference between methods of 4△, found 8–22 participants would be required for Bland Altman analysis, depending on the heterophoria technique examined.

Conclusions

In our study, the Howell Card produced the least inter-examiner and intra-examiner variability although at near even this test did not meet our pre-set repeatability criteria of ±2△. Therefore, the diagnosis of binocular vision disorders requiring the detection of small differences in heterophoria should be made with caution as this level of repeatability does not appear to be present with most clinical measurement techniques.

Conflicts of interest

The authors have no conflicts of interest to declare.

Acknowledgements

Funding: This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

References

1. Jenkins TC, Pickwell LD, Yekta AA. Criteria for decompensation in binocular vision. *Ophthalmic Physiol Opt*. 1989;9:121–125.
2. Letourneau JE, Ducic S. Prevalence of convergence insufficiency among elementary school children. *Can J Optom*. 1988;50:194–197.
3. Rouse MW, Borsting E, Hyman L, et al. Frequency of convergence insufficiency among fifth and sixth graders. The Convergence Insufficiency and reading Study (CIRS) Group. *Optom Vis Sci*. 1999;76:643–649.
4. Scheiman M, Chase C, Borsting E, et al. Effect of treatment of symptomatic convergence insufficiency on reading in children: a pilot study. *Clin Exp Optom*. 2018;101:585–593.
5. Rouse M, Borsting E, Mitchell GL, et al. Academic behaviors in children with convergence insufficiency with and without parent-reported ADHD. *Optom Vis Sci*. 2009;86:1169–1177.
6. Borsting EJ, Rouse MW, Mitchell GL, et al. Validity and reliability of the revised convergence insufficiency symptom survey in children aged 9 to 18 years. *Optom Vis Sci*. 2003;80:832.
7. Insufficiency Treatment Trial C. The convergence insufficiency treatment trial: design, methods, and baseline data. *Ophthalmic Epidemiol*. 2008;15:24–36.
8. Hussaindeen JR, Rakshit A, Singh NK, et al. The minimum test battery to screen for binocular vision anomalies: report 3 of the BAND study. *Clin Exp Optom*. 2018;101:281–287.
9. 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;81:939–946.
10. 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.
11. Goss DA, Reynolds JL, Todd RE. Comparison of four dissociated phoria tests: reliability & correlation with symptom survey scores. *J Behav Optom*. 2010;21:99–104.
12. Rouse MW, Borsting E, Deland PN. Convergence Insufficiency and reading Study (CIRS) Group. Reliability of binocular vision measurements used in the classification of convergence insufficiency. *Optom Vis Sci*. 2002;79(4):254–264.
13. Bujang MA, Baharum N. A simplified guide to determination of sample size requirements for estimating the value of intraclass correlation coefficient: a review. *Arch Orofacial Sci*. 2017;12.
14. Elliott DB. *Clinical procedures in primary eye care E-Book*. Elsevier Health Sciences; 2013.
15. McAlinden C, Khadka J, Pesudovs K. Precision (repeatability and reproducibility) studies and sample-size calculation. *J Cataract Refract Surg*. 2015;41:2598–2604.
16. Koo TK, Li MY. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *J Chiropr Med*. 2016;15:155–163.
17. Liljequist D, Elfving B, Skavberg Roaldsen K. Intraclass correlation — a discussion and demonstration of basic features. *PLoS One*. 2019;14:e0219854.
18. Altman DG, Bland JM. Measurement in medicine: the analysis of method comparison studies. *J R Stat Soc Ser D*. 1983;32:307–317.
19. Carkeet A. Exact parametric confidence intervals for Bland-Altman limits of agreement. *Optom Vis Sci*. 2015;92:e71–e80.
20. Bland JM, Altman DG. Measuring agreement in method comparison studies. *Stat Methods Med Res*. 1999;8:135–160.
21. Fogt N, Baughman BJ, Good G. The effect of experience on the detection of small eye movements. *Optom Vis Sci*. 2000;77:670–674.
22. Kottner J, Audige L, Brorson S, et al. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Int J Nurs Stud*. 2011;48:661–671.
23. Holmes JM, Leske DA, Hohberger GG. Defining real change in prism-cover test measurements. *Am J Ophthalmol*. 2008;145:381–385.
24. Rainey BB, Schroeder TL, Goss DA, Grosvenor TP. Reliability of and comparisons among three variations of the alternating cover test. *Ophthalmic Physiol Opt*. 1998;18:430–437.
25. Wong EPF, Fricke TR, Dinardo C. Interexaminer repeatability of a new, modified prentice card compared with established phoria tests. *Optom Vis Sci*. 2002;79:370–375.
26. 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;91:786–792.
27. Sanker N, Prabhu A, Ray A. A comparison of near-dissociated heterophoria tests in free space. *Clin Exp Optom*. 2012;95:638–642.
28. Casillas EC, Rosenfield M. Comparison of subjective heterophoria testing with a phoropter and trial frame. *Optom Vis Sci*. 2006;83:237.
29. Anderson HA, Manny RE, Cotter SA, Mitchell GL, Irani JA. Effect of examiner experience and technique on the alternate cover test. *Optom Vis Sci*. 2010;87:168–175.
30. Hrynchak PK, Herriot C, Irving EL. Comparison of alternate cover test reliability at near in non-strabismus between experienced and novice examiners. *Ophthalmic Physiol Opt*. 2010;30:304–309.