Published in final edited form as:
Ophthalmology. 2009 February ; 116(2): 281–285. doi:10.1016/j.ophtha.2008.09.012.

Defining real change in measures of stereoacuity

Wendy E. Adams, FRCOphth, **David A. Leske, MS**, **Sarah R. Hatt, DBO**, and **Jonathan M. Holmes, BM, BCh**
Department of Ophthalmology, Mayo Clinic, Rochester, MN

Abstract

Purpose—To establish the thresholds for "real change" in stereoacuity by defining long term test-retest variability as 95% limits of agreement for four stereoacuity tests.

Design—Retrospective cohort study.

Participants and/or Controls—We identified 36 patients (median age 17 years, range 7 to 76 years) with any type of stable strabismus who had stereoacuity measured on two consecutive visits. Stable strabismus was defined as angle of deviation within 5 prism diopters (pd) by simultaneous prism and cover test (SPCT) and prism and alternating cover test (PACT).

Methods—Stereoacuity was measured at near using the Preschool Randot and the near Frisby stereotests and at distance using the Frisby Davis Distance (FD2) and the Distance Randot stereotests. Stereoacuity was transformed to log units for analysis. 95% limits of agreement were calculated based on a 1.96 multiple of the standard deviation of differences between test and retest.

Main Outcome Measures—95% limits of agreement for change in stereoacuity thresholds at two consecutive visits.

Results—95% limits of agreement were 0.59 log arcsec for the Preschool Randot, 0.24 for the near Frisby, 0.68 for the FD2, and 0.46 for the Distance Randot. These values correspond to the following octave steps (doublings of threshold; for example, 200 to 400 arcsec): Preschool Randot 1.95, near Frisby 0.78, FD2 2.27, and Distance Randot 1.52.

Conclusions—A change of approximately two octaves of stereoacuity threshold are needed to exceed test-retest variability for most stereoacuity tests. Changes less than two octaves cannot be distinguished from test-retest variability. When used to guide patient management, caution should be taken in interpreting changes in stereoacuity of less than two octaves.

Introduction

Change in stereoacuity has been reported to be a sign of improvement or deterioration in ocular alignment,[1,2] and deterioration of stereoacuity has been suggested, by some authors, as an indicator for surgery in conditions, such as intermittent exotropia.[3] Nevertheless, isolated measures of stereoacuity are influenced by test-retest variability, and such variability has not been extensively studied, particularly with newer tests of stereoacuity. Previous studies have focused on single modalities of stereoacuity testing and have reported test-retest variability during a single day, not over weeks or months.[4,5] Knowing the degree of test-retest variability in stereoacuity measures over weeks or months would be particularly useful in interpreting the results of testing in clinical practice.

Correspondence and reprint requests to: Dr. Jonathan M. Holmes, Ophthalmology W7, Mayo Clinic, Rochester, MN 55905. Phone: (507) 284-3760. Fax: (507) 284-8566. holmes.jonathan@mayo.edu.

None of the authors have any proprietary or financial interests to disclose.

To define real change in stereoacuity, we compared performance on four tests of stereoacuity (Preschool Randot, near Frisby, Distance Randot, and the Frisby-Davis Distance stereoacuity tests) in patients with stable strabismus across two time points, weeks or months apart.

Subjects and Methods

Institutional Review Board approval was obtained for this study. All experiments and data collection were conducted in a manner compliant with the Health Insurance Portability and Accountability Act.

Patients

In order to identify a cohort of patients in which to study test-retest variability, we searched an ocular motility database at our institution for patients with stable strabismus who had not undergone any intervening change in treatment and who had measures of stereoacuity using each of four different stereotests on two consecutive occasions at least a week, but no more than one year apart. Stable strabismus measurements were defined for the purposes of the present study as simultaneous prism and cover test (SPCT) and prism and alternating cover test (PACT) within 5 prism diopters (pd) between visits. This 5 pd limit is well within previous reports of test-retest variability for strabismus measurements.[6] Patients with intermittent strabismus were excluded because the degree of control may have differed at each visit (in contrast to our previous report on variability of stereoacuity in intermittent exotropia, which focused solely on the condition of intermittent exotropia).[7]

Patients were excluded if there was a change in torsion of greater than 5 degrees between visits measured using the Maddox double rod test. Patients who were known to have unstable neurological conditions, such as myasthenia gravis, or evolving neurological conditions were also excluded, as was any patient who had undergone surgery between visits or in the year preceding the first examination. Any patient undergoing non-surgical treatments such as convergence exercises, amblyopia treatment, or visual therapy was excluded, as were those undergoing a change in prism, i.e. from Fresnel prism to ground in prism or change in prism magnitude. Patients were also required to have stable visual acuity, defined as less than or equal to 0.1 logarithm of the minimal angle of resolution (LogMAR) change in either eye between visits. Patients were excluded if there was a change in refractive correction of greater than 1.00 diopter spherical equivalent or a change from single vision to bifocal lenses or vice versa, to reduce the effect that change in refractive correction may have on either distance or near stereoacuity. The above parameters defined patients with stable strabismus and stable visual acuity in whom we could study long term test-retest reliability of stereoacuity.

Assessment of stereoacuity

The following tests of stereoacuity were administered, using presentation protocols that have been described previously: the Preschool Randot was administered at 40 centimeters (testing 40, 60, 100, 200, 400, 800 arcsec),[8] the near Frisby at 37 to 60 centimeters (testing 40, 60, 100, 200, 400 arcsec),[9] the Frisby-Davis Distance (FD2) at 3 meters (testing 20, 40, 80, 160, 200 arcsec) [10,11,12] and Distance Randot at 3 meters (testing 60, 100, 200, 400 arcsec).[13] At each level, two out of two correct responses were required to “pass,” with the exception of the Preschool Randot test, which is designed as two of three. Stereoacuity was recorded as “nil” if the largest disparity could not be passed. Stereoacuity testing was performed as part of routine clinical evaluation. The testing may not have been performed by the same examiner at every visit, but each test was administered following a standardized testing protocol. This paradigm of possible different testers closely reflects the common clinical

situation of evaluating the results of stereoacuity testing from one clinic visit to the next, and so the results of the present study are most generalizable to the clinical setting.

Thirty-six patients (age range 7 to 76 years) met the inclusion criteria. Nineteen (53%) were children under 19 years of age. Visual acuity ranged from −0.2 to 0.5 logMAR. Distance SPCT measurements ranged from 8 pd exotropia to 35 pd esotropia, and from 0 to 12 pd of vertical deviation. Near SPCT measurements ranged from 8 pd exotropia to 6 pd esotropia and from 0 to 25 pd of vertical deviation. PACT measurements showed almost identical ranges.

The time between visits ranged from 10 to 364 days (median 161 days).

Analysis

Stereoacuity values were transformed to log arcsec for the purpose of analysis, ranging from 1.30 (20 arcsec) to 2.90 (800 arcsec). If the patient had no measurable stereoacuity, the next log level (0.3 log arcsec progression) above the largest disparity for that test was assigned as "nil," i.e. 3.2 log arcsec for the Preschool Randot, 2.90 log arcsec for the Near Frisby, 2.60 log arcsec for the FD2, and 2.90 for the Distance Randot. The assignment of the next log level to nil is commonly used in analysis of stereoacuity data,[4] and allows calculations of differences between tests and changes between visits. The reason we assigned different log levels to represent nil for each test was to avoid the bias of assigning a value of 3.2 log arcsec uniformly across all tests. If we had assigned 3.2 log arcsec uniformly, the difference between the largest measurable disparity and nil would have differed between tests, creating bias in the test-retest analysis. Differences between test and retest were calculated for each individual for each test. The 95% limits of agreement and the 95% confidence intervals around the 95% limits of agreement were calculated.[14] These values were then converted back to octave steps, which might also be described as doublings. Each doubling of the stereoacuity threshold, e.g. 100 to 200 arcsec, corresponds to a 0.3 change of the log transformed value, so we therefore divided the 95% limits of agreement value by 0.3 to give a number of octaves. Agreement between scores was also represented as Bland-Altman plots.[14]

To test for a maturational or learning effect in children and to test for any influence of presbyopia, we conducted the following analyses: 1) compared variability in 7–18 year olds (n=19) to >18 year olds (n=17); 2) compared variability in 19 to <40-year-olds (n=3) to ≥ 40-year-olds (n=14).

Results

The 95% limits of agreement are summarized in Table 1 and represented on the Bland-Altman plots in Figure 1 A, B, C and D.

The Bland-Altman plots suggest that, for each stereoacuity test, magnitude of the test-retest differences did not appear to be dependant on the level of stereoacuity (Figure 1 A–D). When converting the log values of the 95% limits of agreement back to octave steps (doublings) of stereoacuity, the half width of the 95% limit of agreement for the Preschool Randot was 1.95 octaves, for the near Frisby was 0.78 octaves, for the FD2 was 2.27 octaves, and for the Distance Randot was 1.52 octaves.

Since most steps in stereoacuity testing are in octaves (e.g., 100 to 200 arcsec), this means that, for most tests, an approximately two step difference is required to indicate a real change, apart from the near Frisby, which requires only a one step difference.

There was no overall tendency for the retest values to be higher or lower than the initial test values (mean differences: Preschool Randot 0.09, near Frisby 0.005, FD2 −0.05, and Distance Randot 0.05, P>0.05 for all comparisons).

To assess the potential impact of including both children and adults in our study, we analyzed children and adults separately. The median change in scores for 7- to 18-year-olds and >18-year-olds were similar (P>0.2 for all 4 stereotests) and not significantly different from zero (P>0.05) in all cases except for the Preschool Randot in 7- to 18-year-olds (mean difference 0.15 log arcsec better on second test, P=0.03).

These data indicate that there was no substantial learning effect or fatigue effect in the children. The magnitude of the 95% limits of agreement was similar in children and adults (Table 1), with the exception of the near Frisby, which was lower in adults (0.29 octaves). There were only 3 patients 19 to 40 years, and therefore it was not possible to evaluate potential differences in variability between pre-presbyopic and presbyopic adults.

Discussion

In our study of test-retest variability over time, using four current stereoacuity tests, patients with stable strabismus and stable visual acuity showed marked variability of stereoacuity thresholds. We found that for most tests a change of at least two octaves (doublings) is needed for change to exceed test-retest variability. Changes that exceed test-retest variability are likely to represent real changes.

Stereoacuity thresholds are often used in clinical practice as a guide to management. If stereoacuity appears to be reduced from one visit to the next, the physician is likely to conclude that the condition is worsening, and might institute treatment, even recommending surgery. Nevertheless, there are few data on what magnitude of change might be expected to be within test-retest variability, and what magnitude of change might reasonably be expected to represent real change. Our study data addresses these issues.

There were differences in the magnitude of test-retest variability among the four tests we studied. The near Frisby test had a 95% limit of agreement less than 0.3 log arcsec (one octave or doubling) overall, which leads to the conclusion that a single-octave change in threshold on the near Frisby is likely to represent real change. Nevertheless, in children, the 95% limit of agreement exceeded 0.3 log arcsec, and therefore a two-octave change in threshold on all stereotests is likely to represent real change in children. The Preschool Randot test and Distance Randot test had a 95% limit of agreement less than 0.6 log arcsec, which leads to the conclusion that a two-octave change in threshold on these tests is likely to represent real change.

Only the FD2 had a 95% limit of agreement over 0.6 log arcsec (0.68 arcsec, Table 1). This would suggest that a three-octave change is needed to be reasonably certain that a real change has occurred. Nevertheless, the FD2 was the only test we performed that had a measurable threshold better than 40 arcsec. It is possible that the inclusion of the 20 arcsec level in FD2 testing increased variability. To test this hypothesis, we conducted an additional analysis, collapsing 20 arcsec into 40 arcsec. Recalculating the 95% limits of agreement for the FD2, we found a level of 0.55, corresponding to just less than two octaves. For ease of interpretation, we therefore suggest that a change of two octaves might be used as a threshold for determining real change in all but the near Frisby test. Applying a change of two octaves as a threshold for real change is convenient because most tests of stereoacuity are designed such that many of the steps are log steps, i.e. doublings of the disparity (for example, 100 to 200 arcsec). In this way, two octaves would be from 100 to 400 arcsec or from 40 to 160 arcsec. Pragmatically, two octaves could be considered a real

change, and a change of less than two octaves could be considered indistinguishable from test-retest variability.

We are unaware of test-retest variability studies using the near Frisby, FD2, or Distance Randot test. Nevertheless, others have studied the Preschool Randot. Fawcett and Birch[4] reported the 95% limits of agreement to be 0.3 log units, slightly less than our value of 0.46. Pragmatically, this difference is inconsequential, since 0.3 corresponds to one level of the test, and our conclusion of needing to find a two-level change to be reasonably certain that the change exceeds test-retest variability corresponds to the same recommendation made by Fawcett and Birch. The small difference may be due to the different time frames for the two studies; patients were tested on the same day in Fawcett's study and weeks or months apart on our study.

One potential weakness of our study is that there may have been different examiners assessing stereoacuity at each visit. Nevertheless, this situation is common in clinical practice, where often the same technician or orthoptist or physician does not measure stereoacuity on a subsequent visit, yet these data are used for clinical management decisions. We therefore believe that the results of our study are particularly applicable to clinical practice.

An additional weakness is a potential learning or maturation effect, particularly in the children. Overall, test and retest scores were similar. We performed a separate secondary analysis for the 19 children between 7 and 18 years, and found that the change in score was not significantly different from zero for three of the four tests. If there had been a significant learning or maturation effect, we would have expected all thresholds of the retest score to have been better than the initial test score. Similarly, we found no worsening of stereoacuity thresholds on the retest in subjects over 40 years old. Therefore, we did not find learning, maturation, or presbyopic effects.

Another possible weakness of this study is the different levels assessed by each stereotest. For example, the FD2 used at 3 meters presents stereoacuity thresholds of 20 to 200 arcsec, whereas the Preschool Randot presents thresholds from 40 to 800 arcsec. Nevertheless, these differences are a function of the manufacture of the tests, and therefore reflect common clinical practice.

The lack of standardization of testable levels in each stereotest leads to a somewhat problematic application of our results. The finest levels chosen for the FD2 were those in a log progression of 20, 40, 80, and 160 seconds of arc; however, the largest disparity that can be tested is 200 arcsec (corresponding to how far the shape can be displaced within the apparatus). Using this testing protocol, caution is needed when interpreting a change from 80 to 200 arcsec, because this is less than 2 logarithmic steps. Similarly, when using the Preschool Randot test, the change from 40 to 60 arcsec is not a logarithmic step. These problems are analogous to the differences between using a classic Snellen chart for visual acuity versus using a chart with a logMAR progression. Future tests of stereoacuity should be designed with only logarithmic steps, e.g. 25, 50, 100, 200, 400, and 800 arcsec, or 20, 40, 80, 160, 320, and 640 arcsec.

The patients who had measurable stereoacuity on one visit and no measurable stereoacuity on the next visit (or vice versa) deserve further comment. This occurred once with Preschool Randot, three times with the Distance Randot, and once with the FD2; there were no occurrences with the Frisby test. It is possible that for these patients their true stereoacuity threshold was close to the coarsest measurable level and therefore would be expected to test as present on some administrations and absent on others. Including patients with "nil" stereoacuity also may have biased our results toward somewhat better agreement than we

would have found if we included only patients with measurable stereoacuity. Nevertheless, the finding of no stereoacuity on one examination and measurable stereoacuity on the next would have been missed in such a study design. We therefore felt that including patients with nil stereoacuity on the first exam was needed to represent the full spectrum of stereoacuity thresholds.

We did not study the Titmus test because it is no longer used in our routine clinical practice. We, and others, have previously reported the problems of monocular cues when administering and interpreting the results of the Titmus fly, animals, and circles.[15–18] Only circles 5 to 9 (140 arcsec to 40 arcsec) appear to be free from monocular clues.[15] We prefer to use the Preschool Randot, Distance Randot, and Near Frisby, which are free from monocular clues, and the FD2, which has a monocular test phase[12] to account for potential monocular clues. The Near Frisby is only free from monocular cues when administered correctly, perpendicular to the line of sight and not allowing head movement.[15]

Our data can be reasonably extrapolated to patients with constant strabismus. It is entirely possible that in conditions with intermittency, such as intermittent exotropia, the “normal variability” of the condition results in greater variability of stereoacuity than we are currently reporting for constant strabismus. Our previous study of stereoacuity in intermittent exotropia supports this assertion.[7] Re-analyzing those data[7] from the standpoint of reliability yields a 95% limit of agreement of more than 2 octaves for intermittent exotropia (Preschool Randot, Distance Randot, and FD2; unpublished analysis). Such variability of an underlying condition increases the challenge of using stereoacuity to monitor that condition, in which even larger thresholds for change (such as 3 octaves) would need to be observed, and perhaps observations repeated, to have any degree of certainty that the condition had changed.

Despite test-retest variability, measures of stereoacuity such as the Preschool Randot, Distance Randot, near Frisby, and FD2 will still be very useful outcome measures in clinical studies, such as studies of interventions for strabismus. In large studies, the now known test-retest variability could be accounted for by having a sufficient sample size. Nevertheless, care should be taken when interpreting measurements of stereoacuity in an individual patient, and using those values for clinical decisions. In general, a change of two octaves in stereoacuity threshold should be taken to indicate a probable real change, whereas a lesser change may well be within test-retest variability.

Acknowledgments

Supported by National Institutes of Health Grants EY015799 (JMH); Research to Prevent Blindness, Inc., New York, NY (JMH as Olga Keith Weiss Scholar and an unrestricted grant to the Department of Ophthalmology, Mayo Clinic), and Mayo Foundation, Rochester, MN.

References

1. Fu VL, Stager DR, Birch EE. Progression of intermittent, small-angle, and variable esotropia in infancy. Invest Ophthalmol Vis Sci. 2007; 48:661–4. [PubMed: 17251463]
2. Stathacopoulos RA, Rosenbaum AL, Zanoni D, et al. Distance stereoacuity. Assessing control in intermittent exotropia. Ophthalmology. 1993; 100:495–500. [PubMed: 8479706]
3. O’Neal TD, Rosenbaum AL, Stathacopoulos RA. Distance stereo acuity improvement in intermittent exotropic patients following strabismus surgery. J Pediatr Ophthalmol Strabismus. 1995; 32:353–7. [PubMed: 8587017]
4. Fawcett SL, Birch EE. Interobserver test-retest reliability of the Randot preschool stereoacuity test. J AAPOS. 2000; 4:354–8. [PubMed: 11124670]

5. Schmidt P, Maguire M, et al. Vision in Preschoolers Study Group. Random Dot E stereotest: testability and reliability in 3- to 5-year-old children. J AAPOS. 2006; 10:507–14. [PubMed: 17189143]
6. Holmes JM, Leske DA, Hohberger GG. Defining real change in prism-cover test measurements. Am J Ophthalmol. 2008; 145:381–5. [PubMed: 18045567]
7. Hatt SR, Mohney BG, Leske DA, Holmes JM. Variability of stereoacuity in intermittent exotropia. Am J Ophthalmol. 2008; 145:556–61. [PubMed: 18201680]
8. Birch E, Williams C, Hunter J, et al. Random dot stereoacuity of preschool children. J Pediatr Ophthalmol Strabismus. 1997; 34:217–22. [PubMed: 9253735]
9. Frisby JP, Davis H, McMorrow K. An improved training procedure as a precursor to testing young children with the Frisby Stereotest. Eye. 1996; 10:286–90. [PubMed: 8776462]
10. Frisby, JP.; Davis, H., editors. Clinical tests of distance stereopsis: State of the art. Lisse: Swets & Zeitlinger; 2003. p. 187-90.
11. Adams WE, Hrisos S, Richardson S, et al. Frisby Davis distance stereoacuity values in visually normal children. Br J Ophthalmol. 2005; 89:1438–41. [PubMed: 16234448]
12. Holmes JM, Fawcett SL. Testing distance stereoacuity with the Frisby-Davis 2 (FD2) test. Am J Ophthalmol. 2005; 139:193–5. [PubMed: 15652852]
13. Fu VL, Birch EE, Holmes JM. Assessment of a new distance Randot stereoacuity test. J AAPOS. 2006; 10:419–23. [PubMed: 17070476]
14. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; 1:307–10. [PubMed: 2868172]
15. Holmes, JM.; Leske, DA. In: Pritchard, C., editor. Monocular clues in tests of stereoacuity; Transactions IX International Orthoptic Congress; Nurnberg. Germany: Berufsverband der Orthoptistinnen Deutchlands e. V; 1999. p. 103-6.
16. Levy NS, Glick EB. Stereoscopic perception and Snellen visual acuity. Am J Ophthalmol. 1974; 78:722–4. [PubMed: 4415892]
17. Clarke WN, Noel LP. Stereoacuity testing in the monofixation syndrome. J Pediatr Ophthalmol Strabismus. 1990; 27:161–3. [PubMed: 2366128]
18. Fawcett SL, Birch EE. Validity of the Titmus and Randot circles tasks in children with known binocular vision disorders. J AAPOS. 2003; 7:333–8. [PubMed: 14566315]

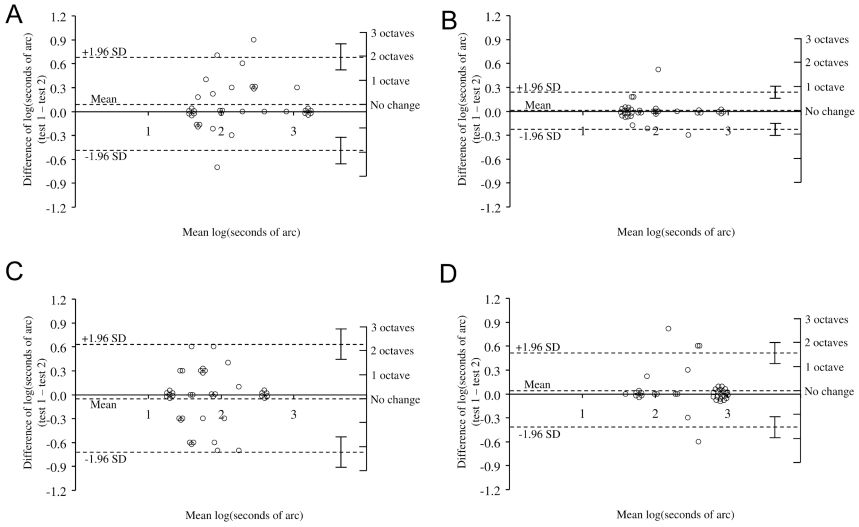

Figure 1.

Test-retest variability represented as Bland-Altman plots for the Preschool Randot (A), Near Frisby (B), Frisby-Davis Distance (C), and Distance Randot (D) stereoacuity tests. Upper and lower dotted lines show 95% limits of agreement with 95% confidence intervals.

Table 1

95% limits of agreement for each stereotest and conversion to number of octaves needed to indicate “real change”

	Overall (N=36)		Adults (ages 19 +, N=17)		Children (ages 7 to 18, N=19)	
Test	**95% Limits of agreement (95% confidence interval) (log arcsec)**	**Number of octaves***	**95% Limits of agreement (95% confidence interval) (log arcsec)**	**Number of octaves**	**95% Limits of agreement (95% confidence interval) (log arcsec)**	**Number of octaves**
Preschool Randot	0.59 (0.41, 0.76)	1.95	0.57 (0.31, 0.82)	1.89	0.60 (0.35, 0.85))	1.99
Near Frisby	0.24 (0.17, 0.31)	0.78	0.09 (0.05, 0.12)	0.29	0.32 (0.18, 0.45)	1.06
Frisby-Davis Distance (FD2)	0.68 (0.48, 0.88)	2.27	0.60 (0.33, 0.87)	1.99	0.74 (0.43, 1.05)	2.47
Distance Randot	0.46 (0.32, 0.59)	1.52	0.45 (0.25, 0.65)	1.49	0.48 (0.28, 0.68)	1.59

* The number of octaves is calculated by dividing the 95% limit of agreement in log seconds by 0.3 (which corresponds to a doubling of stereoacuity disparity).