Published in final edited form as:
Am J Ophthalmol. 2010 April ; 149(4): 672–676. doi:10.1016/j.ajo.2009.11.004.

Test-Retest Reliability of Health-Related Quality of Life Questionnaires in Adults with Strabismus

David A. Leske, **Sarah R. Hatt**, and **Jonathan M. Holmes**
Department of Ophthalmology, Mayo Clinic, Rochester, MN

Abstract

Purpose—To report the test-retest variability of the new Adult Strabismus 20 (AS-20) and the NEI VFQ-25 health-related quality of life (HRQOL) instruments in adults with strabismus.

Design—Prospective case series

Methods—55 adult patients in a clinical practice with stable strabismus completed the AS-20 and the NEI VFQ-25 at two visits, without intervening treatment. Questionnaires were completed the second time either at a subsequent office visit, immediately pre-operatively, or by mail. Intraclass correlation coefficients were calculated. 95% limits of agreement and 95% confidence intervals (CI) around the 95% limits of agreement were also calculated.

Results—There was excellent agreement of overall questionnaire scores for the AS-20 (ICC = 0.92) and NEI VFQ-25 (ICC = 0.94). 95% limits of agreement for overall scores were 14.3 points (CI 10.9 to 17.7) for the AS-20 and 11.1 points (CI 8.5 to 13.8) for the NEI VFQ-25. The lower test-retest variability of the VFQ-25 appeared to be partly due to ceiling effects with many scores at the normal end of the range.

Conclusions—The new AS-20 and the NEI VFQ-25 show excellent test-retest reliability in adults with strabismus. Change exceeding 95% limits of agreement (14 points on the AS-20 and 11 points on the VFQ-25) is indicative of real change in an individual patient. The AS-20 may be more useful than the VFQ-25 because it is less prone to ceiling effects in adults with strabismus.

Introduction

Formal assessment of health-related quality of life (HRQOL) has been recommended in the management of adult strabismus.[1] Several vision-specific HRQOL instruments have been

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.

Financial Disclosures: Grant support by National Institutes of Health Grants EY015799 and EY018810 (JMH), and Research to Prevent Blindness, New York, NY (JMH as Olga Keith Weiss Scholar and an unrestricted grant to the Department of Ophthalmology, Mayo Clinic)

Contributions to Authors: Design of the study (DAL, SRH, JMH); conduct of the study (DAL, SRH, JMH), collection of the data (SRH, JMH), management of the data (DAL, SRH, JMH), analysis of the data (DAL, JMH) interpretation of the data (DAL, SRH, JMH), Preparation of the manuscript (DAL, SRH, JMH), review of the manuscript (DAL, SRH, JMH), and approval of the manuscript (DAL, SRH, JMH).

Statement about Conformity with Author Information: Institutional Review Board approval was obtained at the author's institution and each patient gave informed consent before participating. All procedures and data collection were conducted in a manner compliant with the Health Insurance Portability and Accountability Act.

used in the evaluation of adults with strabismus,[2–9] but reports on the test-retest reliability of these instruments are sparse. Reliability data provide an estimate of variance due to random error or *measurement error*.[10–12] Test-retest reliability data describe the extent to which repetition of the test yields the same results when no underlying change in health has occurred. Limits of agreement calculated from test-retest data are particularly helpful for interpreting changes in scores over time for an individual patient, and have been used in other fields, such as prism and cover test measurements in strabismus.[13, 14] In previous reports we describe the development and initial validation of the AS-20, a strabismus-specific HRQOL questionnaire for adults.[8, 15] In the present study we report the test-retest reliability of the AS-20 questionnaire in a cohort of adults with strabismus. For comparison, we also assessed the test-retest reliability of the National Eye Institute Visual Function Questionnaire (VFQ-25)[16] in the same cohort of adult strabismus patients.

Methods

Fifty-five adult strabismus patients (median age 44, range 18 to 80 years) were prospectively recruited from outpatient clinics and completed both the AS-20 and the VFQ-25 at two time points within one year. Questionnaires were completed in the office for the first administration. The second administration was either 1) in the office at a subsequent exam (N=29, 25 to 144 days later, median 66 days), 2) immediately preoperatively (1 day later in 17, 6 days later in 1), or 3) by mail within 2 days (N=8). Patients completing the questionnaires for the second time immediately prior to surgery or by mail were instructed to complete the questionnaires as if they had not completed them before. Patients with inherently variable strabismic conditions (e.g. ocular myasthenia gravis) were excluded to limit the study cohort to patients that were stable between questionnaire administrations. We also excluded patients who had undergone strabismus surgery within 1 year prior to the first exam because patients' symptoms and perceptions might change during the post-operative period. For the office retest administration, patients were required to have stable strabismus (no change in angle of deviation > 10 PD in primary position) and no intervening treatment or change in treatment. Thirty-eight (69%) were female and 49 (89%) of patients self-reported their race as 'White.' Strabismus diagnoses were idiopathic in 32 (58%), neurological in 19 (35%) and mechanical in 4 (7%). Of our patients, 35 (64%) had diplopia, 9 (16%) had rare diplopia, and 11 (20%) did not have diplopia. Visual acuity ranged from 20/15 to 20/30 in better eye (median 20/20) and 20/15 to 20/4000 in the worse eye (median 20/20).

Responses for each item on the AS-20 were recorded using a 5-point Likert type scale ('never,' 'rarely,' 'sometimes,' 'often,' and 'always'), and converted, for each patient, to a mean score ranging from 0 (worst HRQOL) to 100 (best HRQOL).[8, 15] The VFQ-25 contains Likert type scales and also yields a mean individual patient score from 0 to 100.16 An administrable version of the AS-20 is available online at http://public.pedig.jaeb.org (Accessed: July 31, 2009) and of the VFQ-25 at http://www.nei.nih.gov/resources/visionfunction/vfq_ia.pdf (Accessed: July 31, 2009).

Statistical Analysis

For both the AS-20 and the VFQ-25, differences in scores at first and second administrations were compared using signed rank tests. Bland-Altman plots[17] were used to analyze the variability of the differences. Half widths of the 95% limits of agreement were calculated using 1.96 SD to define the 'limits' within which 95% of the differences should lie. 95% confidence intervals (CI) around the 95% limits of agreement were also calculated. Intraclass correlation coefficients were calculated between first and second administrations. Analyses were repeated to compare variability in patients with and without diplopia.

Results

Differences between first and second administrations

As expected in a test-retest study of a reliable instrument, there were no significant differences between overall and subscale scores on the AS-20 (P>0.5 for all comparisons, Table). Nevertheless, for the VFQ-25, scores were very slightly higher on the second administration (better HRQOL) for the overall score (75.8 versus 77.5, P=0.02) and for the Difficulties with Near Activities subscale (71.5 versus 75.6, P=0.01). There were no other significant differences found in VFQ-25 subscale scores between first and second questionnaire administrations.

Differences between modes of administration

Analyzed separately by mode of administration, the intraclass correlation coefficient for the AS-20 was slightly lower (indicating more variability between measures) for the office and pre-surgery modes than for the mail mode (0.91: 95% CI 0.82 to 0.96 versus 0.90: 95% CI 0.76 to 0.96 versus 0.93: 95% CI 0.71 to 0.98). For the VFQ-25, the intraclass correlation coefficient was numerically lower, but not significantly lower, for the office than for the pre-surgery or mail modes (0.92: 95% CI 0.84 to 0.96 versus 0.95: 95% CI 0.87 to 0.98 versus 0.94: 95% CI 0.75 to 0.99).

For our estimates of the 95% limits of agreement, we found a similar pattern for the AS-20, where the estimates from retests obtained by office and pre-surgery were slightly higher (indicating more variability between measures) than by mail (15.2: 95% CI 10.2 to 20.3 versus 14.5: 95% CI 8.2 to 20.8 versus 10.4: 95% CI 2.8 to 18.0). For the VFQ-25, the 95% limits of agreement were also slightly higher for the office and pre-surgery modes than by mail (12.9: 95% CI 8.6 to 17.1 vs. 10.3: 95% CI 5.8 to 14.8 vs. 5.5: 95% CI 1.5 to 9.5).

Because the estimates of different modes of administration were similar and the 95% confidence intervals of our estimates included the point estimates of the other methods of administration, we combined the data for subsequent analyses.

Overall intraclass correlations

Using a published scale,[18] agreement between exams, as measured by the intraclass correlation coefficient, was “almost perfect” (>0.80) for both the AS-20 (0.92, CI 0.87 to 0.95, Table) and the VFQ-25 (0.94, CI 0.89 to 0.96). Agreement was also “almost perfect” between questionnaire administrations for both AS-20 subscales. For the VFQ-25 subscales, agreement was “almost perfect” on 7 of the 12 subscales, “substantial” (>0.6 to 0.80) in 3 of 12 subscales, and “moderate” (>0.4 to 0.6) in 2 of 12 subscales (Table).

Overall distribution of test-retest differences

Test-retest differences are plotted against mean scores, as described by Bland and Altman, [17] in the Figure. Across AS-20 scores and across VFQ-25 scores within each instrument, variability did not appear to depend on severity (Figure). Nevertheless, the VFQ-25 scores were clustered toward the normal end of the range in these adults with strabismus, suggesting a possible ceiling effect. Comparing the first to the second administration, neither the AS-20 nor the VFQ-25 demonstrated any significant regression to the mean (data not shown).

Overall 95% limits of agreement

Half-widths of the 95% limits of agreement for the AS-20 overall score were 14.3 points (CI 10.9 to 17.7 points) and for the VFQ-25 overall score were 11.1 points (CI 8.5 to 13.8

points, Figure, Table). Limits of agreement ranged from 17.7 on the AS-20 psychosocial subscale to 19.5 on the AS-20 function subscale and from 11.5 to 36.1 on the VFQ-25 subscales (Table).

Differences between patients with and without diplopia

Thirty-five of the 55 adult patients had diplopia and 11 did not (9 patients had rare diplopia and were not analyzed in this assessment of diplopia versus not). For the AS-20, 95% limits of agreement were similar overall between patients with and without diplopia (15.2: 95% CI 10.7 to 19.8 versus 15.6: 95% CI 6.4 to 24.7). Function scale limits of agreement were numerically lower in patients with diplopia (17.8: 95% CI 12.4 to 23.1 versus 26.9: 95% CI 11.1 to 42.6) and psychosocial scale limits of agreement were numerically higher in patients with diplopia (21.0: 95% CI 14.7 to 27.3 versus 9.9: 95% CI 4.1 to 15.7). VFQ-25 overall limits of agreement were similar between patients with and without diplopia (11.6: 95% CI 8.1 to 15.0 versus 8.3: 95% CI 3.4 to 13.2).

Discussion

Both AS-20 and VFQ-25 questionnaires show excellent test-retest reliability in adults with strabismus. Test-retest reliability is particularly important for instruments that are going to be used to assess change over time. Our data define thresholds consistent with a real change in score over time. The 95% limits of agreement were 14.3 points (95% CI 10.9 to 17.7 points) for the AS-20 overall and 11.1 points (95% CI 8.5 to 13.8 points) for the NEI VFQ-25 in adults with strabismus.

Despite our findings of better test-retest reliability of the VFQ-25 in adults with strabismus, the VFQ-25 may not be the optimal instrument. We have previously reported[19] that fewer adult patients with strabismus (with and without diplopia) have sub-normal scores on the VFQ-25 than on the AS-20. The propensity for adults with strabismus to have high (normal) scores on the VFQ-25 leads to a ceiling effect, where there is no room for improvement in HRQOL scores, that might occur with clinical treatment or variability. Bradley et al[20] also described a ceiling effect using the VFQ-25 in patients with Graves' ophthalmopathy. The clustering of VFQ-25 scores toward the normal end of the spectrum is illustrated in the Figure (bottom). In this way, the apparent better test-retest variability of the VFQ-25 may purely be a corollary of the ceiling effect. Due to this ceiling effect, the VFQ-25 may be insensitive to change in adults with strabismus. In contrast to the VFQ-25, the AS-20 shows the desired distribution of scores that allows room for change, illustrated in the Figure (top).

In a previous study, patients with and without diplopia scored differently on the AS-20 subscales.[19] Nondiplopic patients more often were below normal on the AS-20 psychosocial subscale than on the function subscale (95% versus 42%; P = .002), whereas diplopic patients were more often below normal on the function subscale (85% versus 68%; P = .01). On the psychosocial subscale, more nondiplopic than diplopic patients scored below normal (95% versus 68%; P = .01); on the function subscale, more diplopic than nondiplopic patients scored below normal (85% versus 42%; P = .0005).[19] In the present study, test-retest variability in diplopic patients was numerically higher on the psychosocial subscale and lower on the function subscale compared to patients without diplopia based on 95% limits of agreement. Nevertheless, this subgroup analysis is limited by low sample size of non-diplopic patients, and hence wide 95% confidence intervals.

Data on test-retest reliability of strabismus-specific HRQOL measures are sparse, although several vision-specific questionnaires for other ophthalmic conditions have undergone such testing.[21] Test-retest reliability of the NEI VFQ has been reported for the larger, 51-item version of the questionnaire[22, 23] but not specifically for the VFQ-25. Test-retest intraclass

correlation coefficients for the VFQ-51 have been reported as being between 0.91 and 0.68 for all subscales[23] and the properties of the VFQ-25 have been described as being comparable.[16] Indeed our findings using the VFQ-25 are comparable to those reported for the VFQ-51[16] across most subscales, with the exception of the general vision and color vision subscales, which had lower reliability in the present study, and the driving and general health subscales, which had higher reliability in the present study.

A variety of approaches have been used for collecting retest data. One of the more common approaches, used for the VFQ-51,23 is to send the retest to the patient by mail although re-administering the retest at an office visit is also reported.[24] Although we administered a minority of retest questionnaires by mail, the majority of our data was obtained at a consecutive office exam. We believe that our data from a subsequent office visit (25 to 144 days later) is parallel to the clinical scenario of evaluating a patient post-treatment and therefore these test-retest thresholds would be most useful. It is noteworthy that our thresholds calculated purely from data obtained at office visits are almost identical to our reported overall thresholds.

Other aspects of reliability testing are internal consistency reliability and inter-observer reliability. Internal consistency reliability provides a measure of correlation among the items which comprise an instrument and such values for the AS-20 are reported as part of our previous study (Cronbach's alpha for all 20-items was 0.94, for the psychosocial sub-scale was 0.95, and for the function sub-scale was 0.94).[8] Inter-observer reliability is important when questionnaires are interviewer-administered and is not as relevant when questionnaires are self-administered. Future studies will address the validity of these questionnaires in adults with strabismus, such as responsiveness to changes with surgery and correlation to measures of severity such as angle of deviation and diplopia score.

There are a number of potential weaknesses to this study. First, it is possible that responses were memorized when questionnaires were completed the next day, either at home and returned by mail or immediately prior to surgery. Any memorization could potentially lead to less variability between questionnaire administrations and smaller thresholds. Nevertheless, it is noteworthy that our thresholds calculated purely from data obtained at office visits (25 to 144 days later) are almost identical to our reported overall thresholds. Secondly, reliability data may vary based on the population sample and it is possible that results would have been different had we been able to use a more racially heterogeneous population.

The AS-20 shows excellent test-retest reliability, indicating that it is likely to produce consistent results in adults with stable strabismus. These data are important for interpreting changes in AS-20 scores over time and may be useful in clinical practice and for clinical trials.

Acknowledgments

Funding/Support: Supported by National Institutes of Health Grants EY015799 and EY018810 (JMH), Research to Prevent Blindness, 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.

Other Acknowledgments: None

References

1. Mills MD, Coats DK, Donahue SP, Wheeler DT. American Academy of Ophthalmology. Strabismus surgery for adults: a report by the American Academy of Ophthalmology. Ophthalmology 2004;111:1255–1262. [PubMed: 15177982]

2. Satterfield D, Keltner JL, Morrison TL. Psychosocial aspects of strabismus study. Arch Ophthalmol 1993;111:1100–1105. [PubMed: 8166786]
3. Burke JP, Leach CM, Davis H. Psychosocial implications of strabismus surgery in adults. J Pediatr Ophthalmol Strabismus 1997;34:159–164. [PubMed: 9168420]
4. Menon V, Saha J, Tandon R, Mehta M, Khokhar S. Study of the psychosocial aspects of strabismus. J Pediatr Ophthalmol Strabismus 2002;39:203–208. [PubMed: 12148552]
5. Beauchamp GR, Black BC, Coats DK, et al. The management of strabismus in adults--III. The effects on disability. J AAPOS 2005;9:455–459. [PubMed: 16213395]
6. van de Graaf ES, van der Sterre GW, Polling JR, van Kempen H, Simonsz B, Simonsz HJ. Amblyopia & Strabismus Questionnaire: design and initial validation. Strabismus 2004;12:181–193. [PubMed: 15370526]
7. van de Graaf ES, van der Sterre GW, van Kempen-du Saar H, Simonsz B, Looman CWN, Simonsz HJ. Amblyopia and Strabismus Questionnaire (A&SQ): clinical validation in a historic cohort. Graefes Arch Clin Exp Ophthalmol 2007;245:1589–1595. [PubMed: 17549509]
8. Hatt SR, Leske DA, Bradley EA, Cole SR, Holmes JM. Development of a Quality-of-Life Questionnaire for Adults with Strabismus. Ophthalmology 2009;116:139–144. [PubMed: 19019449]
9. van de Graaf ES, Felius J, van Kempen-du Saar H, Looman CWN, Passchier J, Kelderman HS, Simonsz HJ. Construct validation of the amblyopia and strabismus questionnaire (A&SQ) by factor analysis. Graefes Arch Clin Exp Ophthalmol 2009;247:1263–8. Online: June 3. [PubMed: 19495785]
10. Bergner M, Rothman ML. Health status measures: an overview and guide for selection. Annu Rev Public Health 1987;8:191–210. [PubMed: 3555521]
11. Guyatt G, Walter S, Norman G. Measuring change over time: assessing the usefulness of evaluative instruments. Journal of Chronic Diseases 1987;40:171–178. [PubMed: 3818871]
12. Guyatt GH, Deyo RA, Charlson M, Levine MN, Mitchell A. Responsiveness and validity in health status measurement: a clarification. J Clin Epidemiol 1989;42:403–408. [PubMed: 2659745]
13. Holmes JM, Leske DA, Hohberger GG. Defining real change in prism-cover test measurements. Am J Ophthalmol 2008;145:381–385. [PubMed: 18045567]
14. Pediatric Eye Disease Investigator Group. Interobserver reliability of the prism and alternate cover test in children with esotropia. Arch Ophthalmol 2009;127:59–65. [PubMed: 19139339]
15. Hatt SR, Leske DA, Kirgis PA, Bradley EA, Holmes JM. The effects of strabismus on quality of life in adults. Am J Ophthalmol 2007;144:643–647. [PubMed: 17707329]
16. Mangione CM, Lee PP, Gutierrez PR, et al. Development of the 25-item National Eye Institute Visual Function Questionnaire. Arch Ophthalmol 2001;119:1050–1058. [PubMed: 11448327]
17. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307–310. [PubMed: 2868172]
18. Landis JR, Koch GG. The measurement of observer agreement for categorical data. Biometrics 1977;33:159–174. [PubMed: 843571]
19. Hatt SR, Leske DA, Bradley EA, Cole SR, Holmes JM. Comparison of quality of life instruments in adults with strabismus. Am J Ophthalmol 2009;148:558–562. [PubMed: 19570519]
20. Bradley EA, Sloan JA, Novotny PJ, Garrity JA, Woog JJ, West SK. Evaluation of the National Eye Institute visual function questionnaire in Graves' ophthalmopathy. Ophthalmology 2006;113:1450–1454. [PubMed: 16769116]
21. Margolis MK, Coyne K, Kennedy-Martin T, Baker T, Schein O, Revicki DA. Vision-specific instruments for the assessment of health-related quality of life and visual functioning: a literature review. Pharmacoeconomics 2002;20:791–812. [PubMed: 12236802]
22. Mangione CM, Berry S, Spritzer K, et al. Identifying the content area for the 51-item National Eye Institute Visual Function Questionnaire. Arch Ophthalmol 1998;116:227–233. [PubMed: 9488276]
23. Mangione CM, Lee PP, Pitts J, Gutierrez P, Berry S, Hays RD. Psychometric properties of the National Eye Institute Visual Function Questionnaire (NEI-VFQ). Arch Ophthalmol 1998;116:1496–1504. [PubMed: 9823352]

24. Cassard SD, Patrick DL, Damiano AM, et al. Reproducibility and responsiveness of the VF-14. An index of functional impairment in patients with cataracts. Arch Ophthalmol 1995;113:1508–1513. [PubMed: 7487617]

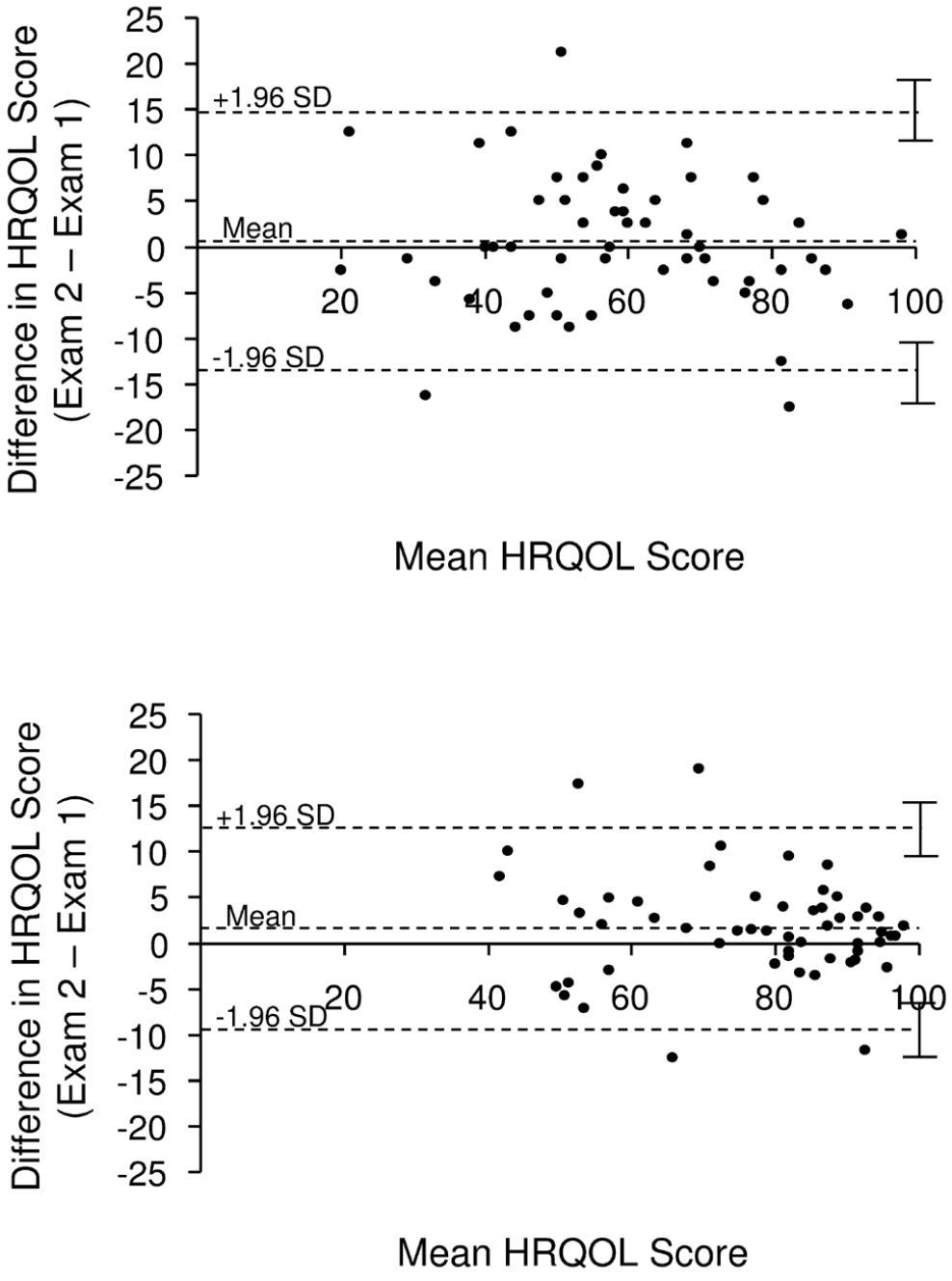

Figure.
Bland-Altman plot of Adult Strabismus Questionnaire (AS-20) (Top) and National Eye Institute Visual Function Questionnaire-25 (VFQ-25) (Bottom)

Am J Ophthalmol. Author manuscript; available in PMC 2011 April 1.

Table

Adult Strabismus Questionnaire (AS-20) and National Eye Institute Visual Function Questionnaire-25 (VFQ-25) test-retest mean scores ± standard deviation scores.

AS-20	**N**	**Test**	**Retest**	**Difference**	**P value[a]**	**95% LOA (95% CI)**	**ICC (95% CI)**
Overall	55	58.9 ± 18.5	59.5 ± 17.8	0.6 ± 7.3	0.5	14.3 (10.9 to 17.7)	0.92 (0.87 to 0.95)
Functional Scale	55	52.2 ± 22.5	52.5 ± 22.2	0.3 ± 9.9	0.9	19.5 (14.9 to 24.1)	0.90 (0.84 to 0.94)
Psychosocial Scale	55	65.6 ± 24.9	66.4 ± 25.5	0.8 ± 9.0	0.7	17.7 (13.5 to 21.9)	0.94 (0.89 to 0.96)
VFQ-25	**N**	**Test**	**Retest**	**Difference**	**P value[a]**	**95% LOA (95% CI)**	**ICC (95% CI)**
Overall	55	75.8 ± 16.8	77.5 ± 16.0	1.7 ± 5.7	0.02	11.1 (8.5 to 13.8)	0.94 (0.89 to 0.96)
General Health	55	68.2 ± 26.1	69.5 ± 23.4	1.4 ± 11.2	0.5	22.0 (16.8 to 27.1)	0.89 (0.83 to 0.94)
General Vision	55	70.2 ± 15.8	69.1 ± 17.6	−1.1 ± 16.1	0.7	31.5 (24.1 to 38.9)	0.54 (0.33 to 0.70)
Ocular Pain	55	74.5 ± 23.2	78.2 ± 21.5	3.6 ± 18.4	0.2	36.1 (27.6 to 44.6)	0.66 (0.48 to 0.78)
Near Activities	54[b]	71.5 ± 23.6	75.6 ± 21.4	4.2 ± 11.5	0.01	22.6 (17.2 to 28.0)	0.85 (0.76 to 0.91)
Distance Activities	55	76.1 ± 22.3	78.3 ± 21.3	2.3 ± 10.2	0.1	20.0 (15.3 to 24.7)	0.89 (0.81 to 0.93)
Vision Specific:							
Social Functioning	55	86.4 ± 16.9	89.3 ± 15.9	3.0 ± 11.0	0.05	21.6 (16.5 to 26.7)	0.76 (0.63 to 0.85)
Mental Health	55	59.8 ± 29.8	64.0 ± 28.4	4.2 ± 13.6	0.08	26.7 (20.4 to 33.0)	0.88 (0.81 to 0.93)
Role Difficulties	54[b]	65.7 ± 31.1	66.9 ± 28.1	1.2 ± 14.6	0.4	28.7 (21.8 to 35.5)	0.88 (0.80 to 0.93)
Dependency	52[c]	84.1 ± 23.2	84.5 ± 24.6	0.3 ± 9.0	0.6	17.7 (13.4 to 22.0)	0.93 (0.88 to 0.96)
Driving	52[c]	77.1 ± 22.7	76.4 ± 20.5	−0.6 ± 10.7	0.9	20.9 (15.8 to 26.0)	0.89 (0.82 to 0.94)
Color Vision	55	98.2 ± 6.6	98.6 ± 5.7	0.5 ± 5.9	1.0	11.5 (8.8 to 14.2)	0.55 (0.34 to 0.71)
Peripheral Vision	55	71.4 ± 26.1	73.6 ± 25.2	2.3 ± 16.9	0.3	33.0 (25.2 to 40.8)	0.78 (0.66 to 0.87)

95% LOA = half-width of the 95% limits of agreement; 95% CI = 95% confidence interval around the 95% LOA half width; ICC = Intraclass Correlation Coefficient

[a] P value based on non-parametric paired comparison (signed rank);

[b] Data missing in 1 case;

[c] Data missing in 3 cases