Published in final edited form as:
Ophthalmology. 2013 July ; 120(7): 1492–1496. doi:10.1016/j.ophtha.2012.12.032.

Quantifying diplopia with a questionnaire

Jonathan M. Holmes, B.M., B.Ch., **Laura Liebermann, C.O.**, **Sarah R. Hatt, D.B.O.**, **Stephen J. Smith, M.D.**, and **David A. Leske, M.S.**
Department of Ophthalmology, Mayo Clinic, Rochester, MN

Abstract

Purpose—To report a diplopia questionnaire (DQ) with a data-driven scoring algorithm.

Design—Cross-sectional study.

Participants—To optimize questionnaire scoring: 147 adults with diplopic strabismus completed both the DQ and the Adult Strabismus-20 (AS-20) health-related quality of life (HRQOL) questionnaire. To assess test-retest reliability: 117 adults with diplopic strabismus. To assess responsiveness to surgery: 42 adults (46 surgeries).

Methods—The 10-item AS-20 function subscale score (scored 0 to 100) was defined as the gold standard for severity. A range of weights was assigned to the responses and to the gaze positions (from equal weighting to greater weighting of primary and reading). Combining all response option weights with all gaze position weights yielded 382,848 scoring algorithms. We then calculated 382,848 Spearman rank correlation coefficients comparing each algorithm with the AS-20 function subscale score.

Main outcome measures—To optimize scoring, Spearman rank correlation coefficients (measuring agreement) between DQ scores and AS-20 function subscale scores. For test-retest reliability, 95% limits of agreement, and intraclass correlation coefficient (ICC). For responsiveness, change in DQ score.

Results—For the 382,848 possible scoring algorithms, correlations with AS-20 function subscale score ranged from −0.64 (best correlated) to −0.55. The best-correlated algorithm had response option weights of 5 for rarely, 50 for sometimes, and 75 for often, and gaze position weights of 40 for straight ahead in the distance, 40 for reading, 1 for up, 8 for down, 4 for right, 4 for left, and 3 for other, totaling 100. There was excellent test-retest reliability with an ICC of 0.89 (95% confidence interval 0.84 to 0.92) and 95% limits of agreement were 30.9 points. The DQ score was responsive to surgery with a mean change of 51 ± 34 ($p<0.001$).

Conclusions—We have developed a data-driven scoring algorithm for the diplopia questionnaire, rating diplopia symptoms from 0 to 100. Based on correlations with HRQOL, straight ahead and reading positions should be highly weighted. The DQ has excellent test-retest reliability and responsiveness, and may be useful in both clinical and research settings.

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

This study was presented in parts at the Association for Research in Vision and Ophthalmology meetings, May 2011 and May 2012, Ft. Lauderdale FL.

The authors have no financial or conflicting interests to disclose.

It would be very helpful to have a method of quantifying severity of diplopia in patients with strabismus, particularly when assessing recovery, deterioration, or response to treatment. Of the various methods available for assessing diplopia severity, a previous diplopia questionnaire[1] has been shown to be a rapid and simple test with specific advantages over a diplopia field plotted on a Goldmann perimeter.[2] The original questionnaire assessed diplopia in seven positions of gaze using a 3 response options (never, sometimes, always). Using this original questionnaire[1, 2] in clinical practice, patients sometimes requested additional response options, specifically “rarely” and “often.” The purpose of this study was to develop a revised diplopia questionnaire, to then develop an optimal scoring algorithm, and to evaluate the test-retest reliability of the new instrument.

Subjects and Methods

Institutional review board approval was obtained for this study. All procedures and data collection were conducted in a manner compliant with the Health Insurance Portability and Accountability Act (HIPAA). Verbal consent and written HIPAA authorization were obtained for each patient.

Diplopia Questionnaire

The 7 gaze positions assessed by the new questionnaire remained unchanged from the original questionnaire.[1] Specifically, patients were asked whether over the course of the last week, they experienced diplopia for reading (the patient’s typical reading position), straight ahead distance gaze, down, right, left, up, and in any other gaze position (Table 1). One additional question was added at the beginning of the questionnaire asking the patient if they ever see double in any position of gaze. This question was answered either “yes” or “no” and patients who did not have any diplopia skipped the rest of the questionnaire. The questionnaire is available at http://pedig.jaeb.org/ViewPage.aspx?PageName=DiplopiaQuest (accessed December 11, 2012).

For the version of the questionnaire reported in the present manuscript, we added “rarely” and “often” categories to the original questionnaire so that, for each question, the patient could select from one of 5 response categories: “never,” “rarely,” “sometimes,” “often,” or “always.”

We also removed the question, “can you get rid of it?” (diplopia) which had been in the original diplopia questionnaire,[1] because it was sometimes confusing to the patient and added complexity to the potential scoring algorithms.

Patients

Assessment of response options weighting: Knowing that our aim was to develop a scoring algorithm for the final questionnaire, we asked 101 patients to interpret the responses “rarely,” “sometimes,” and “often,” using a visual analogue scale, after they completed their questionnaire. We planned to weight “never” as the minimum on the final scoring algorithm (zero) and “always” as the maximum (100), so we used those values as the minimum and maximum on the visual analog scale. Patients were asked to draw a mark on a line from never to always, indicate their rating of “rarely,” “sometimes,” or “often” if they had used one or more of these options when completing the questionnaire. 52 patients provided data on “rarely,” 80 on “sometimes,” and 34 on “often.”

Evaluation of scoring algorithms: To provide an independent quantitative measure of severity of diplopia, with which we could compare diplopia questionnaire scores, we used the 10- question function subscale of the Adult Strabismus-20 (AS-20) questionnaire (Table

2) with response options "never", "rarely", "sometimes", "often" and "always."[3] The AS-20 has been found to be a reproducible,[4] representative,[3] and responsive[5, 6] instrument evaluating health-related quality of life in adult strabismus. We planned to calculate correlation coefficients between diplopia questionnaire scores and AS-20 scores, using many different possible weighting schemes for the new diplopia questionnaire, to determine which scoring algorithm best correlated with the AS-20 function subscale score.

147 consecutive adult strabismus patients with diplopic strabismus, or a history of treated diplopic strabismus, were enrolled in this phase of the study (median age 56 years, range 11 to 86 years). To derive a scoring system that was as generalizable as possible, any type of strabismus was eligible for inclusion: 32 (22%) had childhood onset / idiopathic strabismus, 66 (45%) neurogenic, 32 (22%) mechanical or restrictive, 15 (10%) convergence or divergence insufficiency, and 2 (1%) sensory. 89 (61%) were female and for 140 (95%) race was reported as 'White.' 36 (24%) of the 147 patients were also included in the study of response option weighting.

Scoring of Diplopia Questionnaire—Analogous to the weights assigned in the original version of the diplopia questionnaire,[1] we initially scored the response options by allocating a score of 0 to never, 25 to rarely 50 to sometimes, 75 to often, and 100 to always. Gaze positions were initially weighted in the same way as the original questionnaire[1, 2] with a higher proportion of the score assigned to straight ahead distance and reading (Table 1). We recognized that these weighting schemes were arbitrary, so we then developed a range of reasonable values for weights of the response options and for weights of the gaze positions.

Using data from the patients who rated "rarely," "sometimes," and "often," on the visual analog scale, we set limits to the possible range of scores for each category. The minimum score in the range was defined by the lower quartile of patient ratings on the visual analog scale and the maximum score in the range was defined by the upper quartile. Nevertheless, if the original arbitrary value (25 for rarely, 50 for sometimes, and 75 for often) was outside that range, we used that value as the minimum or maximum. We assigned a score of 0 for "never" and a score of 100 for "always." In creating possible sets of scores, we applied the rule that rarely must score less than sometimes, which must score less than often. These methods yielded the following permissible ranges: rarely 5–25%, sometimes 20–50%, and often 65–80%. The ranges differed in size for each category because they were based on the aforementioned rules. Using increments of 5%, combining these possibilities yielded 128 possible scoring algorithms for response options.

We then created a series of rules for possible scoring of gaze positions. All scores had to sum to 100. Following established scoring rules for diplopia fields[7] and cervical range of motion[1, 8] methods of assessment of diplopia, "straight ahead" and "reading" positions were always assigned higher weights than other positions, with a minimum weighting of 15 each. The weighting assigned for "upgaze" was assigned to be lower than the weighting assigned for "down," "right," and "left" gaze scores. "Right" and "left" gaze were assigned equal weighting. The "Any other" position was assigned a lower score than "down," "right," and "left" gaze. The maximum weighting for any position was limited to 50, and the maximum overall score was 100 if the patient checked "always" for all gaze positions. Increments of 5 points were used for "straight ahead" and "reading," increments of 2 points for "down," "right," and "left," and increments of 1 point for "up" and "any other." This method yielded 2,991 possible scoring algorithms for gaze positions.

Analysis

We combined all above scoring algorithms for response options with all above scoring algorithms for gaze positions, yielding 382,848 possible scoring algorithms. We then applied each of 382,848 possible scoring algorithms to the diplopia questionnaire responses of each of the 147 patients. We then assessed the relationship between total diplopia score and AS-20 function score by calculating a Spearman rank correlation coefficient. The 382,848 possible scoring algorithms for the diplopia questionnaire yielded 382,848 correlation coefficients. We then determined which of the 382,848 scoring algorithms was best correlated to the AS-20 function subscale score.

Test-retest reliability study—To assess test-retest reliability of the new diplopia questionnaire we then recruited 117 patients with diplopia to complete the diplopia questionnaire first at their office visit and again 1 to 180 days later (median 21 days). In 47 (40%) of these patients, the diplopia questionnaire was completed again one week later by mail. In 12 (10%) patients, the questionnaire was completed at their preoperative visit (on the day before surgery) and again the next morning before their surgery. In 58 (50%) patients, the retest was at the time of a follow-up clinic visit. Patients were excluded if there was any treatment (surgery, prism, or exercises) between test and re-test, or if there condition was considered potentially variable or transient (e.g. myasthenia gravis or acute sixth nerve palsy). 53 (36%) of these patients were among the 147 patients in the initial component of the study.

For assessment of test-retest reliability we calculate an intraclass correlation coefficient (ICC) and 95% limits of agreement. We also represented the data as a Bland-Altman plot.

Responsiveness—To assess responsiveness of the diplopia questionnaire, we evaluated the preoperative and 6- week postoperative scores in a consecutive cohort of 42 adults with diplopia due to paretic and restrictive strabismus, who underwent a total of 46 strabismus surgeries. Diagnoses included third, fourth, sixth, and multiple cranial nerve palsies, thyroid eye disease, post-scleral buckle strabismus, and Brown’s syndrome.

We calculated the mean change in score and the proportion who exceeded the 95% limits of agreement. We compared the preoperative and postoperative scores using a signed-rank test.

Results

AS-20 function subscale scores ranged from 8 to 100 (median 54, quartiles 35, 75). For the 382,848 possible scoring algorithms, the Spearman rank correlations with the AS-20 function subscale score were fairly narrowly distributed, ranging from −0.64 (best correlated) to −0.55. Negative values are expected because the most severe diplopia is scored as 100 and the most severe deficit in HRQOL is scored as 0.

Our initial arbitrary scoring algorithm (25 for rarely, 50 for sometimes and 75 for often) yielded a Spearman rank correlation of −0.58, indicating that the initial scoring algorithm was not unreasonable, but also was not optimal.

The best-correlated algorithm (r_s = −0.64) had response option weights of 5 for rarely, 50 for sometimes, and 75 for often, and gaze position weights of 40 for straight ahead in the distance, 40 for reading, 1 for up, 8 for down, 4 for right, 4 for left, and 3 for other, totaling 100. (Table 1)

Differences between first and second administrations

The mean difference in scores between the first and second administration was 0.06 points (range, −50 to 55). Test-retest differences are plotted against mean score, as described by Bland and Altman,[9] in Figure 1. Agreement between examinations, as measured by the ICC was excellent (ICC=0.89, 95% confidence interval 0.84 to 0.92). The 95% limits of agreement between test and retest were 30.9 points.

Responsiveness

The mean diplopia score improved from preoperatively to 6-weeks postoperatively (71 ± 31 to 20± 25, mean change 51 ± 34, p<0.0001). Regarding improvement more than the 95% limits of agreement, there were 8 cases where the preoperative score was too low to improve by 30.9 points. For the remaining 38 surgeries, 31 (82%) had pre-operative to post-operative improvement that exceeded the 95% limits of agreement.

Discussion

We have developed a diplopia questionnaire with a data-driven scoring algorithm that allows the clinician or researcher to assess the severity of a patient's diplopia based on patient report, assigning a numerical score from 0 (no diplopia) to 100 (diplopia everywhere and always). The data-driven scoring algorithm is based on correlation with health-related quality of life. A fillable Excel spreadsheet is available online at http://pedig.jaeb.org/ViewPage.aspx?PageName=DiplopiaQuest (accessed December 11, 2012). Numerical scoring of this patient reported outcome will have broad application for assessment of diplopia severity and assessments of treatment outcomes. The diplopia questionnaire shows excellent test-retest reliability in adults with strabismus, with an intraclass correlation coefficient of 0.89, (95% CI 0.84 to 0.92) and 95% limits of agreement of 30.9 points.

The diplopia questionnaire is one of several methods that can be used to quantify diplopia. The Cervical Range of Motion (CROM) method has been reported to have advantages over the standard Goldmann perimeter method,[8] due to its portability and ability to assess diplopia in a real world environment. Nevertheless, the CROM still requires special equipment and is limited to capturing a patient's diplopia at a specific point in time on a specific day. Both the CROM and the Goldmann provide a diplopia score based on standardized, objective testing whereas the diplopia questionnaire provides an entirely patient-reported outcome.

The advantage of a questionnaire is that it requires no special equipment, is very easy to administer, and is a "patient-reported outcome measure" that reflects the patient's experience. We previously found[2] that the original version of the diplopia questionnaire had advantages over a diplopia field plotted on a Goldmann perimeter, because a questionnaire can quantify diplopia during the course of the patient's everyday life. In the new questionnaire, we continue to specifically ask the patient to rate their diplopia "over the past week."

In the present study we refined the original diplopia questionnaire by first providing additional response options; "rarely" and "often," in addition to "never," "sometimes" and "always." We have now provided external validation of a scoring algorithm, by relating the scores to a validated measure of HRQOL, rather than applying an arbitrary scoring algorithm. In the future we envision an electronic version of the diplopia questionnaire, which might directly interface with the electronic medical record.

Using our method of calculating correlation coefficients with AS-20 functional subscale scores, we found that the best correlation was when we assigned a very high weight to straight ahead in the distance and to reading, with much lower weights assigned to other gaze positions. This data-driven finding is consistent with the assumptions made when a scoring algorithm was arbitrarily designed by Sullivan et al. for diplopia field performed on the Goldmann perimeter[7] and for our previous scoring of the cervical range of motion method and our original diplopia questionnaire.[1] By deriving gaze-position weighting from correlations with function-related HRQOL, we believe diplopia scores reflect the patient's everyday experience. For the majority of patients, low weighting of eccentric gazes is consistent with little functional impairment from diplopia in these positions. The very high weights assigned to straight ahead in the distance and to reading might suggest that we could ignore the other gaze positions and not sacrifice the utility of the instrument, but clinical practice suggests that there are some patients who are only troubled by diplopia in eccentric gazes, and therefore it is important to have an instrument that captures such problems.

In our preliminary assessment of responsiveness, we found an overall marked mean improvement in diplopia questionnaire score following strabismus surgery for paretic and restrictive diplopic strabismus, confirming that the instrument captures patient symptoms preoperatively and postoperatively. Additional studies are needed to evaluate predicted differences between patients who are successfully aligned versus those who are not, analogous to our work on the AS-20 HRQOL questionnaire.[6] We found that 18% of patients, with preoperative scores high enough improve more than 95% limits of agreement, did not do so. The 30.9 point threshold for the 95% limits of agreement threshold may be considered large, but likely reflects the inherent variability when using any questionnaire-based score. Similarly, failure to improve beyond 95% limits of agreement may be caused, in part, by test retest variability and warrants further analysis in terms of other measures of success or failure. Analogous to the AS-20 questionnaire, some caution should be applied when interpreting individual patient diplopia scores. The diplopia questionnaire instrument may be particularly useful in cohort studies, where average values are compared.

Our study is not without limitations. We choose the AS-20 function score as the gold-standard to evaluate the optimal scoring and weighting of the diplopia questionnaire. It is possible that the function subscale of the AS-20 does not entirely represent the severity of a patient's diplopia, but we believe that other alternatives may be more prone to misrepresentation. For example, we previously found that the diplopia field plotted on the Goldman perimeter can seriously under-represent or over-represent a patient's diplopia in the real world.[2] In addition, other health-related quality of life questionnaires such as the VFQ-25 could be used to determine optimal scoring, but the VFQ-25 has been found to be less sensitive to the concerns of adult patients with strabismus than the condition specific AS-20.[10] It may be possible to use our new Rasch-derived version of the AS-20[11] to perform a similar analysis, but since we found two separate function subscales,[11] a general function subscale and a reading function subscale, a weighting rule for HRQOL subscales would need to be developed to analyze the HRQOL data. Currently, it is not clear how one should weight reading function versus general function. To date, development and validation of the diplopia questionnaire has focused primarily on patients with binocular diplopia, and it may therefore be less sensitive to other strabismus-related symptoms such as visual confusion. Another potential limitation is that our weighting of the diplopia score was derived from a spectrum of patients with differing diagnoses. Weighting may have been different if derived from a single sub-group (for example paretic strabismus). Nevertheless, by including patients across the spectrum of strabismus our scoring algorithm is broadly generalizable. It may be possible to evaluate scoring algorithms for specific sub-groups in future analyses.

Our new diplopia questionnaire with data-driven scoring algorithm is useful patient-reported outcome measure that can be used to assess severity of diplopia both in the clinical and research setting. The diplopia questionnaire has good test-retest reliability and may be useful for evaluating changes over time and with treatment.

NIH-PA Author Manuscript
NIH-PA Author Manuscript

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Acknowledgments

Financial support: This study was supported by National Institutes of Health Grant 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. None of the sponsors or funding organizations had a role in the design or conduct of this research.

References

1. Holmes JM, Leske DA, Kupersmith MJ. New methods for quantifying diplopia. Ophthalmology. 2005; 112:2035–2039. [PubMed: 16185766]
2. Adams WE, Hatt SR, Leske DA, Holmes JM. Comparison of a diplopia questionnaire to the Goldmann diplopia field. J AAPOS. 2008; 12:247–251. [PubMed: 18258467]
3. Hatt SR, Leske DA, Bradley EA, et al. Development of a qualityof- life questionnaire for adults with strabismus. Ophthalmology. 2009; 116:139–144. [PubMed: 19019449]
4. Leske DA, Hatt SR, Holmes JM. Test-retest reliability of healthrelated quality-of-life questionnaires in adults with strabismus. Am J Ophthalmol. 2010; 149:672–676. [PubMed: 20138603]
5. Hatt SR, Leske DA, Holmes JM. Responsiveness of health-related quality-of-life questionnaires in adults undergoing strabismus surgery. Ophthalmology. 2010; 117:2322–2328. [PubMed: 20832120]
6. Hatt SR, Leske DA, Liebermann L, Holmes JM. Changes in healthrelated quality of life 1 year following strabismus surgery. Am J Ophthalmol. 2012; 153:614–619. [PubMed: 22285013]
7. Sullivan TJ, Kraft SP, Burack C, O'Reilly C. A functional scoring method for the field of binocular single vision. Ophthalmology. 1992; 99:575–581. [PubMed: 1584576]
8. Hatt SR, Leske DA, Holmes JM. Comparing methods of quantifying diplopia. Ophthalmology. 2007; 114:2316–2322. [PubMed: 17512980]
9. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet. 1986; 1:307–310. [PubMed: 2868172]
10. Hatt SR, Leske DA, Holmes JM. Comparison of quality-of-life instruments in childhood intermittent exotropia. J AAPOS. 2010; 14:221–226. [PubMed: 20417138]
11. Leske DA, Hatt SR, Liebermann L, Holmes JM. Evaluation of the Adult Strabismus-20 (AS-20) Questionnaire using Rasch analysis. Invest Ophthalmol Vis Sci. 2012; 53:2630–2639. [PubMed: 22447864]

NIH-PA Author Manuscript

NIH-PA Author Manuscript

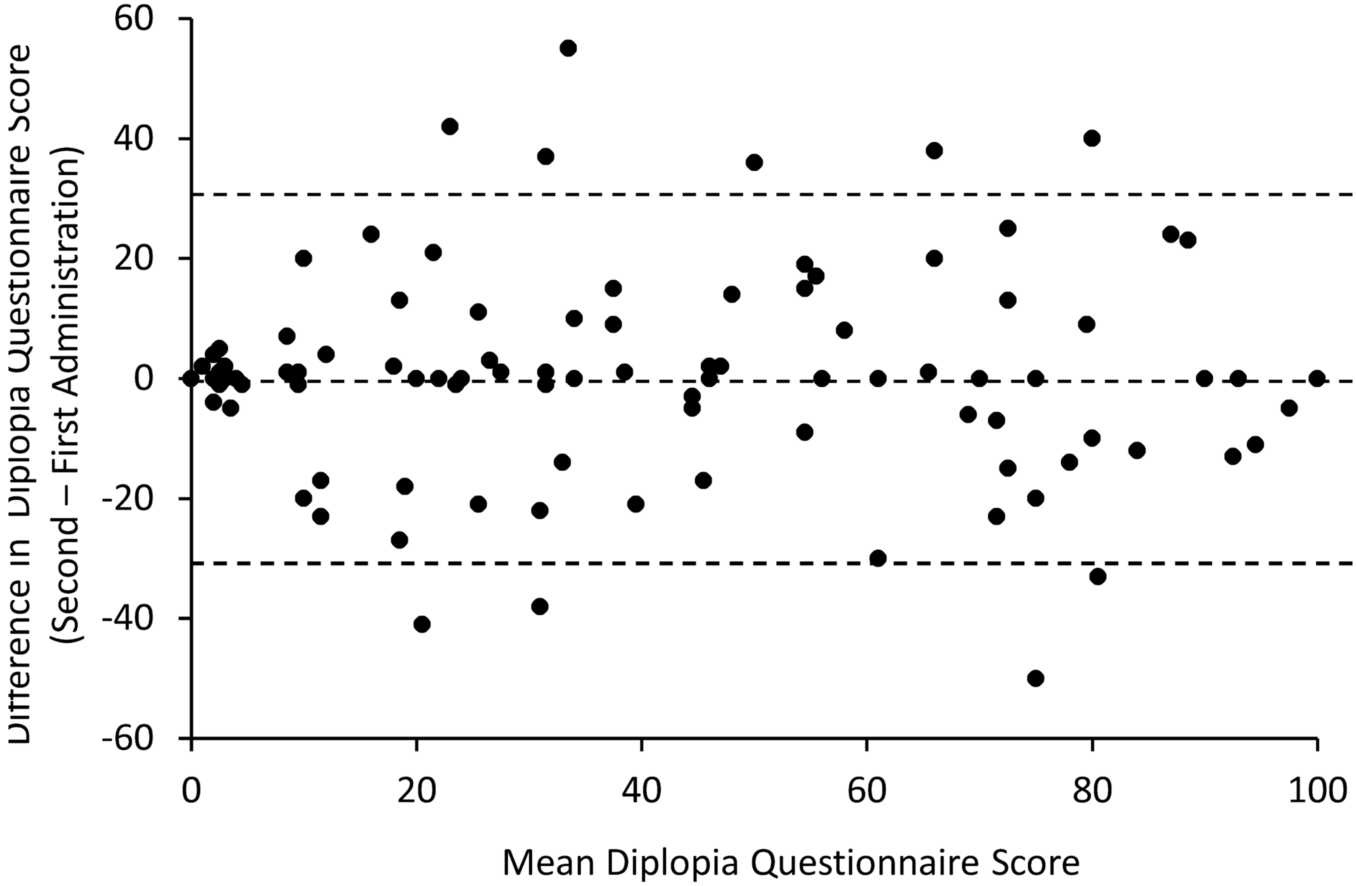

Figure 1.
Test-retest data represented as a Bland-Altman plot for Diplopia Questionnaire applying the proposed new scoring algorithm. The 95% limits of agreement were 30.9 points on a 0 to 100 scale.

NIH-PA Author Manuscript

NIH-PA Author Manuscript

Table 1

Diplopia Questionnaire. Positions of gaze and assigned scores showing initial scoring, prior to the current study and revised scoring, derived from the current study.

Gaze positions	Initial, arbitrary scoring					Revised scoring based on current study				
	Always	**Often**	**Sometimes**	**Rarely**	**Never**	**Always**	**Often**	**Sometimes**	**Rarely**	**Never**
1- reading	16	12	8	4	0	40	30	20	2	0
2- straight	24	18	12	6	0	40	30	20	2	0
3- up	8	6	4	2	0	1	0.75	0.5	0.05	0
4- down	16	12	8	4	0	8	6	4	0.4	0
5- right	16	12	8	4	0	4	3	2	0.2	0
6- left	16	12	8	4	0	4	3	2	0.2	0
7- any other	4	3	2	1	0	3	2.25	1.5	0.15	0

If any item 1 to 6 is blank, an overall score is not calculated

If item #7 is left blank, a “never” value (0) is imputed.

Table 2

Adult Strabismus-20 health-related quality of life questionnaire, showing the 10 function subscale questions

1) I cover or close one eye to see things better
2) I avoid reading because of my eyes
3) I stop doing things because my eyes make it difficult to concentrate
4) I have problems with depth perception
5) My eyes feel strained
6) I have problems reading because of my eye condition
7) I feel stressed because of my eyes
8) I worry about my eyes
9) I can’t enjoy my hobbies because of my eyes
10) I need to take frequent breaks when reading because of my eyes