Test-retest reliability of the revised diplopia questionnaire

Sasha A. Mansukhani, MBBS, Sarah R. Hatt, DBO, David A. Leske, MS, and Jonathan M. Holmes, BM, BCh

PURPOSE To evaluate misclassification of diplopia "success" when using a standardized diplopia questionnaire (DQ), and to report test-retest reliability of the DQ.

METHODS We retrospectively identified a cohort of 100 patients with stable strabismus (<5$^{\Delta}$ change in prism and alternate cover test measurements at distance and near), stable visual acuity, no change in treatment, and no clinical evidence of change, with completed DQ at two consecutive office visits (median, 71 days apart; range, 5-350 days). To evaluate the rate of misclassification of "success" and "not success," we compared the second to the first administration of the DQ using two established definitions of success: (1) "rarely" or "never" for straight ahead distance and (2) "rarely" or "never" for straight-ahead distance and reading. For DQ test-retest variability, 95% limits of agreement (LOA) and intraclass correlation coefficients (ICC) were calculated on DQ scores (0-100 scale).

RESULTS When defining success as rarely or never diplopic for distance, misclassification occurred in 12 (12%) of 100 (95% CI, 6%-20%). When defining success as rarely or never diplopic for distance and reading, misclassification occurred in 14 (14%) of 100 (95% CI, 8%-22%). The 95% LOA for the DQ score were 35.2 points, and ICC was 0.85 (95% CI, 0.79-0.90).

CONCLUSIONS We have quantified misclassification and test-retest variability when using the DQ dichotomously or as a continuous measure, equipping the clinician to better interpret DQ outcome data in practice and research. (J AAPOS 2019;23:319.e1-5)

The diplopia questionnaire (DQ) was developed to standardize documentation of double vision on a frequency scale and also to quantify numerically the severity of a patient's diplopia in both the clinical and research setting.[1] The DQ has been particularly useful for studying surgical outcomes[2] and for defining thresholds of diplopia for clinical studies.[3,4] In a recent, large multicenter observational study, the Study of Adult Strabismus (SAS1) conducted by the Pediatric Eye Disease Investigator Group,[5] the DQ was used as the primary outcome measure, defining success as "rarely" or "never" in straight-ahead distance for divergence-insufficiency and as "rarely" or "never" in straight-ahead distance and reading position for small angle hypertropia. Nevertheless, the DQ, like any clinical measure, is prone to test-retest variability, which may result in misclassification of outcomes when any continuous measure (such as the DQ) is dichotomized. Currently no data are available on the extent to which there is a risk of misclassification when a threshold has been defined using the DQ.

The underlying premise for the present study was that if there was no test-retest variability, a stable patient would not be misclassified, that is, no stable patient initially classified as "success" would become "not success" or vice versa on follow-up. We therefore studied a cohort of patients with stable strabismus and repeated assessments of diplopia to estimate the risk of misclassification when dichotomizing DQ responses as "success" versus "not success" and to quantify the test-retest variability of the DQ when used as a continuous measure.

Subjects and Methods

Approval for this study was obtained from the Mayo Clinic Institutional Review Board. All experiments and data collection were conducted in compliance with the US Health Insurance Portability and Accountability Act of 1996. From a clinical database we identified patients over 15 years of age with stable strabismus (defined below) in a single strabismus practice (JMH), where the DQ has been used routinely over many years. We excluded patients with oscillopsia, monocular diplopia, central-peripheral rivalry type diplopia due to retinal misregistration,[3,6] or visual confusion. We also excluded patients with conditions known to be somewhat variable, including myasthenia gravis; intermittent

Author affiliations: Department of Ophthalmology, Mayo Clinic, Rochester, Minnesota
Supported by National Institutes of Health Grant EY024333 (JMH) and EY011751 (JMH) and Mayo Foundation, Rochester, Minnesota. None of the sponsors or funding organizations had a role in the design or conduct of this research.

Presented as a poster at the 45th Annual Meeting of the American Association for Pediatric Ophthalmology and Strabismus Annual, San Diego CA, March 27-31, 2019.

Submitted April 25, 2019.

Revision accepted August 4, 2019.

Published online October 23, 2019.
Correspondence: Dr. Jonathan M. Holmes, BM, BCh, Ophthalmology E4, Mayo Clinic, Rochester, MN 55905 (email: holmes.jonathan@mayo.edu).

1091-8531/$36.00

https://doi.org/10.1016/j.jaapos.2019.08.277

exotropia; superior oblique myokymia; acute-onset oculomotor, trochlear, or abducens nerve palsies (onset within 6 months); and active Graves' ophthalmopathy. We also excluded patients undergoing strabismus surgery within 1 year of the first examination.

Criteria for Stability

As in previous studies,[7,8] we defined strabismus stability as no change, horizontally and vertically, $>5^{\Delta}$ between examinations as measured by prism and alternate cover test at distance and near, with no change in direction of the vertical deviation and no change in torsion $>5°$. We excluded patients with any change in treatment between examinations, including starting or discontinuing prism; change in prism magnitude; switch between Fresnel or ground-in prism; change in spectacle prescription more than 1.0 D spherical equivalent; switch between lined bifocal, progressive bifocal, or single-vision spectacles; switch from spectacles to contact lenses or vice versa; and surgery (strabismus, cataract, orbital decompression, eyelid, intracranial procedures). All levels of visual acuity were allowed, but stability required no change in visual acuity in either eye by more than 0.2 logarithm of minimal angle of resolution (logMAR) to account for expected test-retest variability of visual acuity testing.[9] We also required no other clinical evidence of improvement or worsening, such as new incomitance or notable change in fusional amplitudes, and subjective report of improvement or worsening by history.

Diplopia Questionnaire Administration and Scoring

The DQ[1] was self-administered at each visit. The full questionnaire and scoring is freely available at www.pedig.net. The first question asks whether the patient has experienced double vision in any position of gaze during the previous week while wearing habitual refractive correction (including prism). Patients answering "no" do not complete the remaining questions. Those answering "yes" rate the frequency of their double vision in reading gaze, straight-ahead distance, up, down, right, left and any other position. The DQ used for the present study displayed specific frequencies for each written frequency descriptor on the printed form (always, 100%; often, 75%; sometimes, 50%; rarely, 5%; never, 0%). The original DQ[1] did not have these specific weights printed with the response options.

We required completion of the DQ at two consecutive office visits, at least 5 days but no >1 year apart (range, 5-350 days; median, 71 days). We chose a minimum interval that was long enough to minimize direct recall of responses and a maximum interval of 1 year to reduce the chance that long-term change had occurred that was not accounted for by our other stability criteria. For a given patient, we used data from the first pair of visits meeting stability criteria.

Definition of Diplopia Success versus Not Success

For the purpose of estimating the magnitude of misclassification of a dichotomous outcome ("success" vs "not success") using the DQ, "success" was defined in two ways: (1) "rarely" or "never" diplopic for straight-ahead distance (used for divergence insufficiency in the PEDIG Study of Adult Strabismus),[5] and (2) "rarely" or "never" for straight-ahead distance and reading (used for small-angle hypertropia in the PEDIG Study of Adult Strabismus).[5] Classification of "success" at the first examination was assumed to be true, and a change in classification at the subsequent visit (in the absence of any clinical change) was deemed misclassification.

Definition of Diplopia Improvement versus Not Improvement

As an alternative analysis, we considered the effect of test-retest variability on classification of "improvement" versus "not improvement." "Improvement" was defined as a decrease in frequency of at least 2 levels (eg, from "always" to "sometimes" or better, or "often" to "rarely" or better, or "sometimes" to "never"). In order to be able to improve, subjects were required to have diplopia frequency of "always," "often," or "sometimes" at the first examination. Classification of "improvement" at the first examination was assumed to be true, with a change of at least 2 levels at the subsequent visit (in the absence of any clinical change) deemed misclassification.

Definition of Diplopia Worsening versus not Worsening

We also considered the effect of test-retest variability on classification of diplopia outcomes as "worsening." "Worsening" was defined as an increase in frequency of at least 2 levels (eg, from "never" to "sometimes" or worse, or "rarely" to "often" or worse, or "sometimes" to "always"). In order to worsen, subjects were required to have diplopia frequency "never," "rarely," or "sometimes" at the first examination. Classification of "worsening" at the first examination was assumed to be true, with a change of 2 or more levels at the subsequent visit (in the absence of any clinical change) deemed misclassification.

Analysis

For analysis of dichotomous classifications (success, improvement, worsening), for each definition (above), the proportion of misclassified patients was calculated along with 95% confidence intervals. We repeated these analyses comparing younger (<55 years of age) versus older (≥55 years) patients, dividing the cohort at the median age (55 years). Cohen's κ test was performed to evaluate the agreement between the test-retest classifications. We also hypothesized that the risk of misclassification would be greater for patients further away from the success/not success threshold compared with those closer. Therefore, we also compared the rate of misclassification for straight-ahead distance in a subset of patients with "always" at the initial evaluation to a subset of patients with "sometimes" at the initial evaluation using the Fisher exact test.

The DQ was scored as a continuous measure (0-100 scale) for each examination for each patient, as described previously.[1] The scoring algorithm weights the gaze positions as 40% for straight-ahead in the distance, 40% for reading, 1% for upgaze, 8% for downgaze, 4% for right gaze, 4% for left gaze, and 3% for other gaze. Using this scoring algorithm, we calculated an

Table 1. Frequency of test and retest responses for straight-ahead distance gaze using the diplopia questionnaire in 100 patients with stable strabismus

	Never	Rarely	Sometimes	Often	Always	Total
Never	36	2	1	1	0	40
Rarely	3	7	3	1	1	15
Sometimes	3	1	10	4	4	22
Often	0	0	3	8	0	11
Always	1	0	0	1	10	12
Total	43	10	17	15	15	100

Table 2. Frequency of test and retest responses for reading gaze using the diplopia questionnaire in 100 patients with stable strabismus

	Never	Rarely	Sometimes	Often	Always	Total
Never	40	4	0	1	0	45
Rarely	4	8	7	2	0	21
Sometimes	1	5	5	5	1	17
Often	0	0	1	2	1	4
Always	0	0	0	1	12	13
Total	45	17	13	11	14	100

intraclass correlation coefficient and 95% limits of agreement to create Bland-Altman plots.[10,11]

Results

A total of 100 patients (64 females) with stable strabismus were identified. The median patient age was 55 years (range, 15-92); 89 (89%) of patients self-reported their race as white. Strabismus diagnoses were childhood-onset or presumed childhood-onset in 30 (30%), neurologic in 35 (35%), mechanical in 15 (15%), adult onset idiopathic in 18 (18%), and sensory in 2 (2%). The median angle of horizontal or vertical strabismus (largest) measured by simultaneous cover and prism test for distance and near was 4^{Δ} and 1.5^{Δ}, respectively (horizontal range:,distance $>50^{\Delta}$ esotropia to 40^{Δ} exotropia; near $>50^{\Delta}$ esotropia to 50^{Δ} exotropia; , vertical range, 0^{Δ}-35^{Δ}, distance and near). Visual acuity ranged from 20/15 to 20/200 in the better eye (median, 20/20) and from 20/15 to 20/400 in the worse eye (median, 20/25).

Diplopia Frequency and Misclassifications

The frequency of individual responses for both the test and the retest evaluations are summarized for straight-ahead distance gaze in Table 1 and for reading gaze in Table 2.

For the first definition ("rarely" or "never" for straight-ahead distance, used in the PEDIG SAS1 Divergence insufficiency study),[5] 55 patients were classified "success" at the initial evaluation. Using this definition, the misclassification was 12 (12%) of 100 (95% CI, 6%-20%). Cohen's κ coefficient of agreement was 0.76 (95% CI, 0.63-0.89). There was no difference in misclassification rate by age group: 5 (10%) in younger patients (<55 years; n = 50) and 7 (14%) in older patients (difference, −4; 95% CI, −17% to 9%).

Fifty patients were classified as success using the definition of "rarely" or "never" diplopic for distance and reading (used in the PEDIG SAS1 small-angle hypertropia study).[5] Misclassification occurred in 14 (14%) of 100 (95% CI, 8%-22%). The κ coefficient of agreement was 0.72 (95% CI, 0.58-0.86). There was no difference in misclassification rate by age group: 5 (10%) in younger patients and 9 (18%) in older patients (difference, −8%; 95% CI, −22% to 6%).

The rate of misclassification for straight-ahead distance was higher among patients with "sometimes" at initial evaluation (18%) when compared to patients with "always" (8%) on initial evaluation, however this was not statistically significant (difference, 10%; 95% CI, −13% to 32% [$P = 0.6$]).

Criteria for improvement were met by 45 of 100 for straight-ahead distance and 34 of 100 for reading. Using this definition of improvement in straight-ahead distance gaze, misclassification occurred in 4 (9%) of 45 (95% CI, 2%-21 %), whereas using the definition of improvement in reading gaze, misclassification occurred in 1 (3%) of 34 (95% CI, 0%-15%).

Criteria for worsening were met by 77 of 100 for straight ahead distance and 83 of 100 for reading. Using this definition of worsening in straight-ahead distance gaze, misclassification occurred in 8 (10%) of 77 (95% CI, 5%-19%), whereas using the definition of worsening in reading gaze, misclassification occurred in 4 (5%) of 83 (95% CI, 1%-12%).

Test-retest Reliability of the DQ Continuous Score

The mean difference in score between the test and retest DQ evaluations was 3.8 points ± 18.0 points, and the intraclass correlation coefficient was 0.85 (95% CI, 0.79-0.90). The Bland-Altman plot (Figure 1) shows test-retest differences plotted against mean score. The 95% limits of agreement between test and retest were 35.2 points (95% CI, 29.1-41.3), and visual inspection of the Bland-Altman plot revealed no marked relationship between degree of test-retest variability and severity of diplopia.

Discussion

The rate of misclassification of diplopia success in patients with strabismus using the DQ appears relatively low, ranging from 12% to 14%, although the upper ends of the 95% confidence intervals range from 20% to 22%. The rate of misclassification of diplopia improvement and worsening in patients with strabismus using the DQ also appears to be low, ranging from 3% to 10%, with the upper ends of the 95% confidence intervals ranging from 12% to 21%. Clinicians and researchers need to be aware of the potential for misclassification when interpreting clinical and research data, particularly in single-armed case-series type studies.

Issues of misclassification are more or less important depending on study design. Ideally, research studies using dichotomized DQ data would use a comparative-group

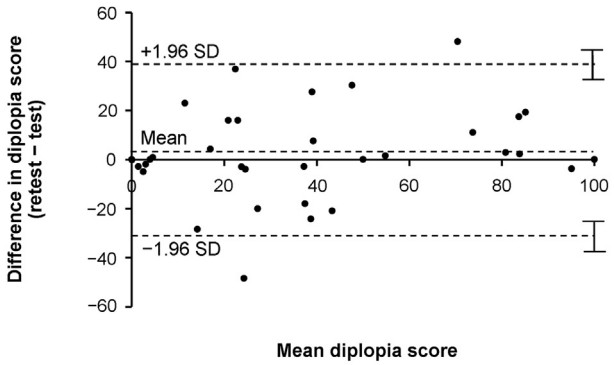

FIG 1. Test-retest variability represented as a Bland-Altman plot showing the difference in Diplopia Questionnaire (DQ) test and retest scores plotted against the mean DQ score. Limits of agreement were 35.2 points (95% CI, 29.1-41.3), with a mean difference in test-retest score of 3.8 ± 18.0 points. There was no apparent relationship between severity of diplopia and variability.

study design where misclassification would be expected to be equal between groups. Such equal risk of misclassification between groups is most likely when there is randomization of treatment assignment, and both known and unknown confounders are balanced between groups. Researchers and readers should be aware of potential biases toward more misclassification in one treatment group versus another, for example, where one group has baseline values closer to a particular threshold than another. Indeed, in our planned secondary analysis, we found a greater rate of misclassification (albeit not statistically significant) in patients whose baseline diplopia frequency was "sometimes" compared with "always." This phenomenon of increased risk of misclassification if a measurement is closer to any predefined threshold was also pointed out by Holmes regarding visual acuity thresholds in studies of amblyopia.[12] Any difference between treatment groups in baseline parameters (such as frequency of diplopia) could be mitigated by a randomized study design and further mitigated by statistical analyses that adjust for baseline factors.

To our knowledge, no previous study has addressed the potential problem of misclassification using the DQ. During the development of the DQ,[1] test-retest reliability was evaluated, but that earlier version did not incorporate numerical quantification of the written descriptors. In that previous study,[1] the intraclass correlation coefficient of DQ scores (on the same 0-100 scale) was 0.89, with 95% limits of agreement of 30.9 points, both very similar to the current study, where patient instructions indicated 75% for "often," 50% for "sometimes," and 5% for "rarely."[1]

The field of strabismus, like many others in medicine, is moving toward incorporating patient-reported outcomes for the evaluation of success of treatments. Examples of eye-specific patient-reported outcomes in strabismus are the AS-20 Questionnaire for Adults,[13] the PedEyeQ for children,[14] and the DQ. Patient-reported outcomes, either alone or in combination with motor outcomes, appear to be more representative of success or failure than motor outcomes alone,[2] and some recent studies, such as the PEDIG study SAS1, define success entirely based on the patient-reported DQ.[5] The results of these newly emerging studies can now be more reasonably interpreted with the knowledge of the risk of misclassification from the present study. In addition, we have described the rate of misclassification of diplopia improvement and worsening in patients with strabismus using the DQ, and these data will be useful in planning future studies that use improvement or worsening of frequency of diplopia as outcome measures.

Our study is not without limitations. It is possible that patients may have completed the questionnaire with recall of their previous responses, particularly when there was a shorter interval between test and retest visits; the minimum time between administrations of 5 days resulted in some overlap between recall periods in a few of our patients (n = 2), which may have reduced testing variability. Nevertheless, the median interval between tests was 71 days. In addition, it is possible that those with a longer duration between tests had adapted to their diplopia, recording a lower frequency at retest despite no clinical change. It is also possible for an apparently stable patient to experience subtle change between visits, although we excluded conditions with known fluctuations. An additional limitation is that it is possible for the patient's perceptions of "rarely," "sometimes," or "often" to change over time; the DQ used in this study incorporated a numerical descriptor of frequency to mitigate this potential effect. We had insufficient numbers to conduct analysis of variability by direction of strabismus. Also, owing to the location of our clinic, our findings may not be generalizable to more racially heterogeneous populations; a future multicenter study may help address this limitation.

The DQ shows good test-retest reliability, but dichotomizing data does introduce a low level of misclassification, which should be considered when interpreting dichotomized DQ data. Quantification of the risk of misclassification and the level of test-retest variability allows the clinician to better interpret DQ outcome data and change over time.

References

1. Holmes JM, Liebermann L, Hatt SR, Smith SJ, Leske DA. Quantifying diplopia with a questionnaire. Ophthalmology 2013;120:1492-6.
2. Hatt SR, Leske DA, Liebermann L, Holmes JM. Comparing outcome criteria performance in adult strabismus surgery. Ophthalmology 2012;119:1930-36.
3. Veverka KK, Hatt SR, Leske DA, et al. Prevalence and associations of central-peripheral rivalry-type diplopia in patients with epiretinal membrane. JAMA Ophthalmol 2017;135:1303-9.
4. Hatt SR, Leske DA, Iezzi R Jr, Holmes JM. New onset vs resolution of central-peripheral rivalry-type diplopia in patients undergoing epiretinal membrane peeling. JAMA Ophthalmol 2019;137:293-7.

5. Pediatric Eye Disease Investigator Group. Study of adult strabismus (SAS1): a prospective observational study of adult strabismus. Version 2.0. Available at: https://public.jaeb.org/pedig/stdy/337; 2016.
6. Veverka KK, Hatt SR, Leske DA, Brown WL, Iezzi R Jr, Holmes JM. Causes of diplopia in patients with epiretinal membranes. Am J Ophthalmol 2017;179:39-45.
7. Adams WE, Leske DA, Hatt SR, Holmes JM. Defining real change in measures of stereoacuity. Ophthalmology 2009;116:281-5.
8. Liebermann L, Leske DA, Hatt SR, Holmes JM. Test-retest variability of cyclodeviations measured using the double Maddox rod test. J AAPOS 2018;22:146-148.e1.
9. Siderov J, Tiu AL. Variability of measurements of visual acuity in a large eye clinic. Acta Ophthalmol Scand 1999;77:673-6.
10. Terwee CB, Bot SD, de Boer MR, et al. Quality criteria were proposed for measurement properties of health status questionnaires. J Clin Epidemiol 2007;60:34-42.
11. Bland JM, Altman DG. Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1986;1:307-10.
12. Holmes JM. Screening, confirming, and treating amblyopia based on binocularity. JAMA Ophthalmol 2014;132:820-22.
13. 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.e5.
14. Hatt SR, Leske DA, Castañeda YS, et al. Development of pediatric eye questionnaires for children with eye conditions. Am J Ophthalmol 2019;200:201-17.