Published in final edited form as:

J AAPOS. 2015 February ; 19(1): 33–37. doi:10.1016/j.jaapos.2014.10.017.

Quantifying variability in the measurement of control in intermittent exotropia

Sarah R. Hatt, DBO, David A. Leske, MS, Laura Liebermann, CO, and Jonathan M. Holmes, BM, BCh

Department of Ophthalmology, Mayo Clinic, Rochester, Minnesota

Abstract

Purpose—To evaluate the performance of a series of summary measures of control and to assess reliability in quantifying exodeviation control in intermittent exotropia.

Methods—A large, simulated dataset of control scores, for 10,000 hypothetical patients with intermittent exotropia was created using Monte Carlo simulations. These data were based on children with intermittent exotropia in whom control was assessed twice during one clinical examination, using the office control score (0–5). Each simulated patient had a baseline and 11 subsequent control scores. The repeatability of a series of summary measures of control (the mean of two vs the mean of three up to the mean of six), was calculated using 95% limits of agreement (LOA).

Results—A total of 322 examinations in 152 patients were used to provide representative distributions of control scores. From the resultant Monte Carlo simulations, the 95% LOAs were 2.60 for one distance control score measure, 1.76 for the average of three, and 1.28 for the average of six. Therefore using the average of three scores, a change of <1.76 would be consistent with short-term variability, whereas a change of >1.76 would suggest a real change in control.

Conclusions—The large dataset of simulated control scores allowed us to assess the variability of specific summary measures of control. We recommend the average of 3 scores (a triple control score) as a new standard for assessing control, providing improved reliability over a single measure, while remaining implementable in clinical practice.

In intermittent exotropia, the ability to control the exodeviation is considered one of the most important measures of severity.^{1–3} Nevertheless, categorizing or quantifying a patient's exodeviation control and determining real change in control remain ongoing challenges. Although control scales^{4–8} have improved our ability to quantify exodeviation control at a given point in time, inherent moment-to-moment variability⁹ makes it difficult to assign a value that truly represents a child's control over an entire day or longer. This short-term

© 2014 by the American Association for Pediatric Ophthalmology and Strabismus. All rights reserved.

Correspondence: Dr. Jonathan M. Holmes, Ophthalmology W7, Mayo Clinic, Rochester, MN 55905 (holmes.jonathan@mayo.edu).

No authors have any financial/conflicting interests to disclose.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

variability of exodeviation control is a combination of inherent variability of the condition and any testing variability. Because short-term variability of control is characteristic of intermittent exotropia even in a stable state, any “real” change in control (ie, one consistent with a change in the underlying severity of intermittent exotropia) must exceed that which can be measured when assessing short-term variability.

Using multiple measures of control and taking an average reduces the variability of the composite measure, and we have previously reported that an average of three measures (a triple control score) over one clinical examination better represents overall control (over a day) than do single measures.¹⁰ Nevertheless, in our previous study¹⁰ our ability to explore the effect of averaging multiple measures was limited to a small cohort of 12 patients. The present study aimed to evaluate the performance of each of a series of specific averaged measures of control and to define thresholds for determining change in control over time by calculating limits of agreement for each measure and considering the practical implications of each.

Subjects and Methods

This study was approved by the Mayo Clinic Institutional Review Board. All procedures and data collection were conducted in a manner compliant with the US Health Insurance Portability and Accountability Act of 1996.

A simulated dataset of control scores for 10,000 hypothetical patients, based on actual clinical data in children with intermittent exotropia, was created to determine the short-term variability (a combination of inherent variability of the condition, and any testing variability) of control measures. The following overall process was followed: First, clinical control score data were identified on intermittent exotropia patients with 2 control scores in a single clinic examination. Second, for the clinical cohort, the frequency of subsequent control scores for given baseline scores was tabulated. For this step all examinations with 2 control scores were included (ie, more than one examination per patient where data were available). Third, clinical data were then used to create 10,000 hypothetical patients with successive control scores using Monte Carlo simulations, with the first score (from clinical data) as the “seed” for subsequent control scores. Fourth, the large, simulated dataset of 10,000 hypothetical patients was used to analyze the performance of a series of specific, summary measures of control and to calculate repeatability. Each of these steps is described in detail below.

Control Score Data—Clinical Cohort

We retrospectively identified office visits in which children with intermittent exotropia had undergone two separate control assessments during a single clinical examination, where no change in severity of intermittent exotropia would be expected and any change in control would be attributable to short-term variability. In order to obtain data on as many pairs of scores as possible, if a child had more than one examination with two measures of control, all eligible examinations were included. Inclusion criteria were as follows: age <18 years; no previous surgical intervention; basic, true divergence excess and pseudo divergence excess types of intermittent exotropia; and exodeviation 10° at distance fixation by prism and

alternate cover test. Patients with convergence insufficiency–type intermittent exotropia (near angle >10° more than distance) were excluded.

Control was assessed using the office control scale,⁴ which provides a standardized assessment of control by which an observer can quantify control on a 6-point scale. Levels 3–5 on the scale rate the duration of any spontaneous tropia during a 30-second observation period (with 5 indicating constant tropia; 3, tropia for <50% of observed time); if no spontaneous tropia is observed, the speed of recovery following dissociation is rated 0–2 (with 2 indicating recovery >5 seconds; 0, <1 second).⁴ Only distance control scores were included in the present study to examine change at the distance at which childhood intermittent exodeviations typically becomes manifest. The two assessments of control were separated in time by other clinical testing, with control typically being assessed at the beginning and at the end of the clinical examination; however, the duration between assessments was not standardized. In some cases the same examiner repeated the control assessment, but for others the orthoptist performed the first control assessment and the ophthalmologist performed the second. The examiner performing the second assessment of control was not masked to the initial control assessment.

Determining Subsequent Control Scores

The first of 2 control scores during the first eligible examination for each subject was designated as the baseline control score. The frequency of subsequent control scores (ie, the second control score) for a given baseline control score was tabulated for all examinations (ie, more than one examination per patient was allowed if data were available).

Monte Carlo Simulations

The distribution of control scores from the first eligible clinical examination of each patient in our clinical cohort was used to model the initial control score for the Monte Carlo simulation analyses of 10,000 hypothetical patients so that the distribution of baseline scores in the simulated cohort mirrored the distribution of baseline scores in the clinical cohort (eg, 22% of the clinical cohort had a baseline control score of 1, therefore 22% of the simulated cohort of 10,000 patients were also assigned a baseline score of 1). Subsequent modeled control scores were simulated using the frequency of the second control scores from all 322 examinations in our clinical cohort, given the modeled baseline score. Simulations of subsequent modeled control scores were repeated 11 times, each predicted from the likelihood of obtaining a specific control score on a second measure given a known baseline control score. Thus each simulated patient had a baseline control score and 11 subsequent control scores for a total of 12 successive control scores per patient.

Summary Measures of Control

Specific summary measures of control were created by averaging consecutive control scores, each composed of a specified number of repeated control measures. The mean of the first 2 control scores was averaged for a double control score measure, the mean of the first 3 control scores for a triple control score measure and so on, until the mean of the first 6 control scores was calculated for a sextuple measure of control. To create measures for repeatability analysis, a second summary control measure was calculated for each summary

measure by calculating the mean of scores immediately following in the dataset (eg, for the double control score measure, the repeat double was the mean of the third and fourth scores in the dataset, and the repeat measure for the sextuple measure was the mean of the 7th through 12th measures).

Repeatability of Measures

Repeatability of control score measures was assessed by calculating 95% limits of agreement (LOA), which provides a threshold within which 95% of the differences between two measurements are expected to lie. Therefore by applying these methods to summary control score and repeat summary score measurements we derived thresholds within which a change in score would be consistent with short-term variability and outside of which real change in control would be considered likely.

The simulated dataset of 10,000 patients was used to calculate 95% LOAs, with 12 successive control scores. To calculate the 95% LOAs on one measure of control, the first measure was compared with the second measure. Similarly, to calculate the 95% LOAs on a summary control score, the first summary control score was compared with the second summary control score.

Although the 95% LOAs for the average of a multiple measure can be estimated by dividing the standard deviation of test–retest differences from a single measure by the square root of the number of repeats, doing so assumes normal distributions. In our method of simulating multiple control scores, we aimed to model the clinical behavior of control, using Monte Carlo simulations to predict subsequent control scores based on the true distribution of measured control scores in a population of children with intermittent exotropia. The estimates of 95% LOAs calculated from our Monte Carlo simulations are slightly different from direct calculations using standard deviations and the square root of the number of repeated measures, but, for the reasons given, we believe they are more clinically meaningful.

Results

Control Score Data—Clinical Cohort

A total of 152 patients were identified with two measures of distance control in a single clinic examination. The distribution of baseline distance control scores (the first of 2 control scores in the first eligible examination for each patient) was 0 in 1 of 152 patients (1%), 1 in 36 (24%), 2 in 51 (34%), 3 in 8 (5%), 4 in 25 (16%), and 5 in 31 (20%).

Determining Subsequent Control Scores

Of the 152 patients, 84 had additional examinations with two measures of control (range, 2–13 examinations per subject), yielding a total of 322 examinations. Median age was 6 years (range, <1 to 17 years), and median angle of exodeviation was 25° at distance (range, 12° – 55°) and 16° at near (range, 10° esotropia to 45° exotropia). For 172 of 322 (54%) the same examiner performed both control assessments, and for 150 (47%) a different examiner

performed the second assessment. The frequencies of the first control scores for all 322 examinations and subsequent control scores are given in Table 1.

Monte Carlo Simulations

The model for Monte Carlo simulations was created based on clinical data. The distribution of baseline distance control scores for the Monte Carlo simulations mirrored the distribution of scores found for the clinical cohort of 152 clinical patients with two measures of control in a single examination (step 1 of the results), and the frequency of subsequent control scores was modeled on the frequencies found for the 322 examinations identified in the clinical cohort (step 2 of the results). With these parameters in place, Monte Carlo simulations were run to create 10,000 hypothetical patients with a baseline control score and 11 subsequent control scores, for a total of 12 successive control scores per patient. Examples of control scores on 10 simulated patients are provided in Table 2. The distribution of baseline scores in the simulated cohort was based directly on the proportions found in the clinical cohort of 152 patients: 124 (1%) had a baseline score had a baseline score of 0; 2,236 (22%), a baseline score of 1; 3,416 (34%), a baseline score of 2; 559 (6%), a baseline score of 3; 1,646 (16%), a baseline score of 4, and 2,019 (20%), a baseline score of 5.

Summary Measures of Control

The 95% LOAs on the series of specific summary control score measures are provided in Table 3. For a single control score at distance the 95% LOA was 2.60, whereas the 95% LOA for the average of three scores was 1.76 and for the average of six scores was 1.28 (Table 3).

We performed additional analyses to evaluate whether variability differs depending on whether the baseline control was less severe (scores 0, 1, and 2) or more severe (scores 3, 4, and 5). We used the triple control score and calculated the 95% LOA, when grouping scores of 0, 1 and 2 (triple of 2.0) and when grouping scores of 3, 4, and 5 (triple of 2.3). For the less severe control scores of 0, 1, and 2, the 95% LOAs were 1.33; for the more severe control scores of 3, 4, and 5, the 95% LOAs were 1.77.

Discussion

Using a large simulated dataset, we calculated the repeatability of a series of specific summary measures of control in intermittent exotropia. Single measures of control show a large degree of short-term variability (95% LOA 2.60), whereas the average of multiple measures shows much less variability. Using the average of multiple control scores reduces the variability of the composite measure, providing a more repeatable measure of control and a more robust means of assessing change in control over time.

Assessment of control is the main parameter by which severity of intermittent exotropia is judged, and worsening control has been used by some investigators as an indicator of the need for surgical intervention.^{1,2,13} Therefore quantifying change in exodeviation control is critically important for the management of intermittent exotropia. Nevertheless, despite the high importance placed on control when making management decisions, the assessment of

control in clinical practice remains inadequate. If exodeviation control is formally quantified at all (using a control scale^{4-8,14}), it tends to be done so only once during a clinic examination. In this present study, as in previous studies,^{9,10} we found a high level of variability in control score when using a single measure, with a 95% LOA of 2.60. This means that a change of >2.60 would be required to be sure of a real change in the underlying condition as opposed to short-term variability. For example, using a single measure of control, a change from a score of 1 (exodeviation following dissociation only, with spontaneous recovery in 1–5 seconds) on one examination, to a score of 3 (spontaneously manifest exotropia for <50% of observed time) on the next examination, is a change of two levels and therefore less than the 95% LOA of 2.60. Such magnitude of change is consistent with short-term variability of control and therefore cannot reliably be interpreted as a real deterioration in control. These data illustrate how unreliable and potentially misleading single measures of control are when evaluating change over time in an individual patient.

By averaging multiple measures of control, repeatability improves dramatically compared with single measures. Nevertheless, there is a need to balance practical implementation alongside optimal repeatability. Despite superior repeatability assessing control six times in an examination, in most clinical settings such testing would be impractical. The average of 3 control scores, however, may provide an optimal balance between practicality and repeatability. In the present study the 95% LOA on the average of three measures was 1.76 at distance, requiring a change in control score of two or more to be certain of a real change. In a previous study we found three measures of control to be achievable over an examination and the average of three scores (a triple control score) to be more representative of the patient's overall control than single measures.¹⁰ Based on these previous data and the data from the current study, we recommend using the average of 3 control scores as a compromise between the practical demands of testing and achieving better repeatability.

We reanalyzed data from our previous study of 12 patients who underwent multiple measures of control¹⁰ to derive 95% repeatability coefficients (the same as 95% LOAs but used when there are more than two measures) and to compare values to those found in the present study. The 95% repeatability coefficients on a triple control score (mean of three measures) were 2.15 at distance, somewhat larger than the 95% LOAs for triple control scores in the present study. The previous study included only 12 patients, and assessments of control occurred over a period of 2 days, both of which may have increased the variability of control score measures. Since recruiting large numbers of patients to perform multiple repeat measures of control is impractical, we believe the simulated data reported in this current study provide a best estimate of variability of control score measures for application in clinical practice.

When analyzing less severe and more severe control using the triple control score, we found somewhat less variability on the lower end of the control scale. It is to be expected that control scores on either end of the scale (0 and 5) will be subject to floor and ceiling effects (respectively), resulting in somewhat reduced variability when control is measured at those extremes. Nevertheless, we would expect a greater floor effect in mild cases than ceiling effect in severe cases because mild cases rarely have periods of poor control, whereas severe cases often have periods of better control. Therefore, it is not surprising that variability is

somewhat less on the lower (less severe) end of the control scale. In terms of practically applying the 95% LOAs of a triple control score, even though we found that the 95% LOAs of less severe scores were 1.33 and more severe scores were 1.77, the starting value for an individual patient may differ dramatically due to short-term variability. For example, a patient may have a score of 2 at one moment and a score of 4 a few minutes later. Therefore we believe using the overall 95% LOA of 1.76 is both reasonable and appropriately conservative.

Few previous studies report the reliability of control score measurements in intermittent exotropia or how to determine whether or not control has changed over time. Previous studies of children with intermittent exotropia have reported change in angle of deviation measurements¹¹ and stereoacuity,¹² but quantification of control and change in control have proven difficult to study. The main challenge to rigorous measurement of exodeviation control is the inherent moment-to-moment variability of the condition itself. Such variability means that even within one clinic examination, different control scores would be assigned depending on the moment at which the control assessment is performed. In the present study we demonstrated that by averaging multiple control scores, a more robust, averaged measure of control can be obtained, enabling a more rigorous assessment of change in control by accounting for some of the moment to moment variability.

There are notable limitations to this study. Due to the desire to study a large cohort of stable patients with many successive measures of control, we used a simulation method to create 10,000 patients, each with 12 successive measures of control. A limitation of Monte Carlo simulation technique is that the results depend on the model parameters, and although we used real patient data to model the simulations, it is possible that repeated measures in 10,000 real patients might differ in a way that would affect our results. Ideally one would measure control over 12 successive examinations in a large cohort of intermittent exotropia patients in whom the underlying condition had not changed, but this is practically impossible. For these reasons we consider simulations to provide the best available data on repeatability of control score measures. Although we had 322 examinations with two measures of control, some of these were repeat examinations in the same patient, therefore theoretically reducing the amount of variability in the scores. However, we only used repeat examinations to maximize our ability to represent the range of possible subsequent control score values and based the proportion of baseline scores in the modeled sample on only the first examination for each patient. There may be some bias toward finding less variability in control when the initial control score is known by the examiner performing the second control score assessment. Despite this potential bias, we found relatively high variability in control scores. Finally, we used the office control scale^{4,9,10} to quantify control and the derived repeatability data are not generalizable to other methods of quantifying control.

By quantifying the repeatability of a series of specific summary control score measures, we were able to quantify more accurately a patient's ability to control and to better determine whether or not there has been a significant change in control over time. A triple control score (the average of three measures over an examination) provides a balance between improved repeatability and the practical demands of testing, and is recommended as an

improved standard for clinical assessment, knowing that variability is greater for more severe than for less severe levels of control.

Acknowledgments

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. von Noorden, GK.; Campos, EC. Exodeviations. In: Lampert, R.; Cox, K.; Burke, D., editors. Binocular Vision and Ocular Motility: Theory and Management of Strabismus. 6. St. Louis: Mosby; 2002. p. 356-76.
2. Santiago, AP.; Ing, MR.; Kushner, BJ.; Rosenbaum, AL. Intermittent exotropia. In: Rosenbaum, AL.; Santiago, AP., editors. Clinical Strabismus Management Principles and Surgical Techniques. Philadelphia: W. B. Saunders Company; 1999. p. 163-75.
3. Hatt SR, Gnanaraj L. Interventions for intermittent exotropia. Cochrane Database Syst Rev. 2013; 5:CD003737. [PubMed: 23728647]
4. Mohny BG, Holmes JM. An office-based scale for assessing control in intermittent exotropia. Strabismus. 2006; 14:147-50. [PubMed: 16950743]
5. Stathacopoulos RA, Rosenbaum AL, Zanon D, et al. Distance stereoacuity. Assessing control in intermittent exotropia. Ophthalmology. 1993; 100:495-500. [PubMed: 8479706]
6. Petrunak JL, Rao R. The evaluation of office control in intermittent exotropia: a systematic approach. Am Orthopt J. 2003; 53:98-104. [PubMed: 21149078]
7. Haggerty H, Richardson S, Hrisos S, Strong NP, Clarke MP. The Newcastle Control Score: a new method of grading the severity of intermittent distance exotropia. Br J Ophthalmol. 2004; 88:233-5. [PubMed: 14736781]
8. Chia A, Seenyen L, Long QB. A retrospective review of 287 consecutive children in Singapore presenting with intermittent exotropia. J AAPOS. 2005; 9:257-63. [PubMed: 15956946]
9. Hatt SR, Mohny BG, Leske DA, Holmes JM. Variability of control in intermittent exotropia. Ophthalmology. 2008; 115:371-6. [PubMed: 17629562]
10. Hatt SR, Liebermann L, Leske DA, Mohny BG, Holmes JM. Improved assessment of control in intermittent exotropia using multiple measures. Am J Ophthalmol. 2011; 152:872-6. [PubMed: 21843874]
11. Hatt SR, Leske DA, Liebermann L, Mohny BG, Holmes JM. Variability of angle of deviation measurements in children with intermittent exotropia. J AAPOS. 2012; 16:120-24. [PubMed: 22525165]
12. Adams WE, Leske DA, Hatt SR, Holmes JM. Defining real change in measures of stereoacuity. Ophthalmology. 2009; 116:281-5. [PubMed: 19091410]
13. Wright, KW. Exotropia. In: Wright, KW., editor. Pediatric ophthalmology and strabismus. St. Louis: Mosby-Year Book Inc; 1995. p. 195-202.
14. Petrunak, JL.; Rao, RC.; Baker, JD. The evaluation of office control in X(T): A systemic approach. In: de Faber, JT., editor. Transactions of the 28th European Strabismological Association Meeting. London: Taylor and Francis; 2004. p. 109-12.

Table 1

Frequency of first and subsequent (second) control scores measured in 322 single clinic examinations (152 children with intermittent exotropia) at distance^a

Initial distance control score		Number (%)	Subsequent control score (%)					
			0	1	2	3	4	5
0		4 (1%)		1 (25%)	3 (75%)	-	-	-
1		72 (22%)		1 (1%)	45 (63%)	15 (21%)	6 (8%)	2 (3%)
2		110 (34%)		-	13 (12%)	56 (51%)	13 (12%)	21 (19%)
3		18 (6%)		-	1 (6%)	7 (39%)	3 (17%)	2 (11%)
4		53 (16%)		-	2 (4%)	22 (42%)	4 (8%)	10 (19%)
5		65 (20%)		-	1 (1%)	13 (20%)	3 (5%)	11 (17%)
								37 (57%)

^aThe proportion showing the same score for first and second assessments in a single clinic examination are in boldface type.

Table 2

Example data showing the baseline control score and 11 subsequent control scores on 10 patients in the simulated dataset of 10,000 patients

Patient	Baseline control score	11 subsequent control scores										
		1st	2nd	3rd	4th	5th	6th	7th	8th	9th	10th	11th
1	5	4	4	3	5	5	5	2	5	4	4	3
2	2	2	2	2	4	4	4	4	2	2	2	2
3	5	2	2	4	5	5	5	5	2	4	2	5
4	4	5	4	4	2	4	5	2	5	4	2	2
5	1	1	1	1	2	1	2	1	1	2	1	1
6	2	2	2	2	3	2	2	2	2	2	2	3
7	2	2	1	2	1	5	2	3	2	2	1	5
8	1	1	2	2	1	2	2	1	1	2	1	1
9	0	1	1	0	0	1	0	1	0	0	1	1
10	3	5	2	2	2	5	5	2	2	3	5	2

Table 3

The 95% limits of agreement on specific, summary (averaged) control score measures, ranging from a single control score to the average of 6 control scores^a

Summary control score measure	95% limits of agreement
Distance	
Single control score	2.60
Average of 2 scores	2.07
Average of 3 scores	1.76
Average of 4 scores	1.53
Average of 5 scores	1.38
Average of 6 scores	1.28

^aData calculated from a simulated cohort of 10,000 patients each with 12 successive control score measures.