Quantitative Evaluation of Saccadic and Smooth Pursuit Eye Movements
Is It Reliable?

M. Versino,* G. Castelnovo,* R. Bergamaschi,* A. Romani,* G. Beltrami,†
D. Zambarbieri,† and V. Cosi*  

**Purpose.** The authors evaluated the reliability of the coefficients of the (1) amplitude/duration and (2) amplitude/peak velocity relationships of the mean precision values and the mean latency values (saccadic eye movements) and the coefficients of the target velocity/gain relationship (smooth pursuit eye movements). They computed test-retest maximum variability limits for these parameters.  

**Methods.** After a 1-week interval, saccadic and smooth pursuit eye movements were recorded twice from 20 healthy subjects; 12 of these subjects underwent a third recording session. The estimate of the intraclass coefficient of reliability, R, was adopted to evaluate the reliability of eye movement quantitative analysis.  

**Results.** The data demonstrated that the reliability was fairly good for the amplitude/peak velocity relationship, was good for the precision, and was excellent for the amplitude/duration, the target velocity/gain relationships, and the latency.  

**Conclusions.** Quantitative analysis of both saccadic and smooth pursuit eye movements is reliable. One statistic used to estimate reliability, ie, the within-subjects mean square value, also enables the determination of test-retest normal variability values for both the variances and the differences of measurements. Invest Ophthalmol Vis Sci. 1993; 34:1702-1709.
estimate of intraclass coefficient of reliability and (2) determine the normal range of variability for repeated measurements.

MATERIALS AND METHODS

Subjects

Saccades and smooth pursuit eye movements were recorded from 20 healthy volunteers (10 men and 10 women; mean age, 30.1 yr; range, 23–59 yr). The research followed the tenets of the Declaration of Helsinki. All subjects gave their consent to testing procedures, and they underwent two recording sessions, held with a 1-week interval, and performed at the same hour of the day (usually late in the morning or late in the afternoon). Twelve subjects underwent a third recording session after a period of time varying from 4–7 mo (mean, 5.6 mo; median, 6 mo).

Recording Procedures

The subjects sat in a dark silent room with their heads stabilized in a four-point headrest. Eye movements were recorded binocularly using electro-oculography (with recording electrodes placed at both outer canthi and the ground electrode at an ear lobe). The electro-oculography signal was filtered (low-pass cutoff frequency, 40 Hz) and stored on a floppy disk (sampling rate, 250 Hz). Eye movements were calibrated at the beginning of each session using the following target presentation sequence: 0°, +20° (right), 0°, −20° (left), and 0°. This sequence was repeated until reliable electro-oculography tracings were obtained.

Saccades were elicited by using a refixation paradigm. In each trial, the 15 light-emitting diodes, which were placed on a semicircular frame, every 5° (from −35° to +35°) at a distance of 80 cm from the subject, were lit according to a pseudorandom sequence in which the lateral diodes were activated twice and each centrifugal saccade was followed by a recentering eye movement. Each session was composed of two trials.

Smooth pursuit eye movements were elicited by means of a semicircular array on which 255 light-emitting diodes were arranged side by side and positioned 104.7 cm from the subject. The activation of these diodes was controlled by the personal computer in such a way as to generate a constant velocity (triangular ramps) movement of the visual target over an amplitude of 60° (±30°). Nine differing velocities ranging from 10–50°/sec were presented according to a pseudorandom sequence in which any couple of movements (ie, from right to left and from left to right) was made at the same velocity and the time interval between two subsequent movements ranged between 1–1.8 sec. Each session was composed of five trials (total number of ramps, 58).

Measurement and Analysis of Eye Movement Parameters

For saccadic eye movements, an interactive program was able to measure the latency (L), the amplitude (A), the duration (D), and the peak velocity (Vp) for both primary and secondary saccades (Fig. 1). We did not consider recentering saccades because of their predictability. Thus, all analyses concerning saccadic eye movements refer to centrifugal saccades.

These values were used to compute the following relationships:

1. Amplitude/duration relationship: \(D = a + b \times A\), where a and b represent, respectively, the intercept of the line with the axis of the ordinates and the slope of the line.

2. Amplitude/peak velocity relationship: \(Vp = A/(c + D \times A)\) where \(1/d\) corresponds to the value of \(Vp\) when \(A\) is infinite (maximum \(Vp\) value) and \(1/c\) corresponds to eye acceleration when \(A\) is equal to zero. By considering the inverse of \(Vp\) and \(A\), the equation can be expressed in a linearized form, \(1/Vp = d + c/A\). We chose this equation instead of the more common exponential equation based on the estimation of the maximum \(Vp\) value because this latter might be underestimated. In turn, if we consider that a relation exists between the average and peak velo-

![FIGURE 1. An example of 2 sec of saccade recording. The upper tracing corresponds to the position signal; the middle tracing, to the velocity signal; and the bottom tracing, to target activation.](image-url)
ity of saccades and that the average velocity is given by the actual amplitude divided by the duration, we obtain the equation mentioned.

3. Precision: \( P = \frac{A}{TA} \), where \( A \) is the amplitude of the primary saccade and \( TA \) is the target amplitude.

For smooth pursuit eye movements, an interactive program was able to remove the catch-up saccades superimposed on the smooth pursuit eye movement (Fig. 2). Then, it computed a gain value, which was expressed as the ratio of the amplitude of smooth tracking (ie, saccades excluded) to the amplitude of the total tracking (smooth tracking + saccade tracking), or in other words, the gain value was expressed as the ratio between the integral of eye and target velocity signals after the elimination of catch-up saccades. Many investigators calculate the gain value as the ratio between eye and target velocity computed in one or several points of the ramps. It is evident that a relationship must exist between these two methods of calculating gain and this relationship has been previously studied by several authors. In addition, our method is based on the entire ramp, rather than on a few data points. For each subject (and for each test), we computed the target velocity/gain relationship, expressed by the equation \( G = m + n \times V \), where \( G \) is the gain, \( V \) is the target velocity, and \( m \) and \( n \) represent, respectively, the intercept of the line with the axis of the ordinates and the slope of the line.

**Intraclass Correlation Coefficient of Reliability: R**

For a detailed explanation of R, the reader should refer to Fleiss. Herein, we will describe in short the theoretic basis of R. If we measure a variable in a subject, the score we obtain \( (X) \) is made of two components, ie, the “true” value of the variable \( (T) \) or the error-free score and the “random error” \( e \), which will vary around a mean of zero with a variance equal to \( \sigma_e^2 \), which should be constant if we assume that the distribution of errors is independent of the value of \( T \). The existence of the random error can be appreciated as the difference between the \( X \) scores when we repeat the same measurement in the same subject under the same conditions. Therefore, if we repeat the measurement of a variable several times in a population of subjects, the variance of \( X \) is \( \sigma_X^2 = \sigma_T^2 + \sigma_e^2 \) where \( \sigma_T^2 \) is the variance of \( T \).

The intraclass correlation coefficient of reliability is defined as

\[
R = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2}
\]

This represents the portion of variability that can be attributed to between-subject variability and will increase as the ratio \( \sigma_T^2/\sigma_e^2 \) increases, or in other words, the lesser the variability caused by random error, the greater the reliability of the measure will be. R ranges from 0 to 1, and it was ranked as follows: Values below 0.4 may be taken to represent poor reliability; those between 0.4 and 0.75, fair to good reliability; and those greater than 0.75, excellent reliability.

If we measure a variable in a sample of subjects several times (the number of measurements is not necessarily the same for all subjects), the weighted mean of the subjects’ variance, the within-subject mean square (WMS), is seen to be an unbiased estimator of \( \sigma_e^2 \). WMS is computed as

\[
\frac{\sum (k_i - 1)s_i^2}{K - N}
\]

where \( k_i \) and \( s_i^2 \) are, respectively, the number of measurements and the variance of the \( i \)th subject, \( K \) is the total number of measurements (ie, \( \Sigma k_i \)), and \( N \) is the number of the subjects.

In turns, \( \sigma_T^2 \) is estimated by the ratio \( (\text{BMS} - \text{WMS})/k_0 \), where BMS is the between-subjects mean square and is computed as

\[
\frac{\sum (k_i - 1)s_i^2}{K - N}
\]
Reliability of Eye Movement Measures

\[ \sum k_i \frac{(\bar{X}_i - \bar{X})^2}{N - 1}, \]

where \( \bar{X}_i \) is the mean of the ith subject and \( \bar{X} \) is the overall mean. In addition, \( k_0 \) is computed as

\[ k_0 = k - \frac{s^2}{N^2}. \]

where \( k \) and \( s^2 \) are the mean and the variance of the numbers of replicate measurements (in our case, \( k_0 = 2.59 \)). Thus, the estimate of \( R \), \( \hat{R} \), is equal to

\[ \hat{R} = \frac{BMS - WMS}{BMS + 1.59WMS}. \]

We computed the \( \hat{R} \) value for the coefficients \( a, b, c, d, m, \) and \( n \) and for the mean precision value and mean latency values.

However, the value of \( \hat{R} \) is underestimated if there is a fixed effect or, in other words, if the repeated testing induces either an increasing or a decreasing trend. Before computing \( \hat{R} \) for our parameters, we tested this hypothesis by means of a repeated-measures analysis of variance with sex as a two-level between-subjects factor. The within-subjects factor was the T-RT factor, which could have either three levels (12 subjects) or two levels (20 subjects). Consequently, the repeated-measures analysis of variance was done twice. The parameters belonging to the same relationship (ie, the intercept and the slope of the amplitude/duration, the amplitude/peak velocity, and the target velocity/gain relationship) were analyzed together.

Finally, we investigated whether the differences between measurements of the same variable were influenced by the subject's age. For each parameter, we computed the differences between the first and the second measurement (DIF1 in 20 subjects), the second and the third measurement (DIF2 in 12 subjects), and the first and the third measurement (DIF3 in 12 subjects). For each of these differences, we computed the absolute value, and we performed a regression analysis in which the subject's age was the independent variable.

We considered the absolute value because our main interest was the identity of the differences (ie, does age affect the degree of modification?) rather than its sign (does age involve a differing trend of modification?)

We fixed the significance level at \( P = 0.01 \) for all analyses.

**Normal Range of Variability**

We computed the normal range of variability for both the variance of three subsequent measurements and the difference between two subsequent measurements. The normal ranges were computed on the basis of the WMS value.

**Normal Range of Variances.** WMS represents the intrasubject variance of measurement computed in a sample of 20 subjects. If we consider another subject, the ratio between their variance and WMS is an F ratio with 1,32 degrees of freedom. Thus, this subject derives from the same population of our normal subject sample if their variance does not exceed the value \( F_{0.01(1,32)} \times WMS \) (significance level \( \alpha = 0.01 \)).

**Normal Range of Differences.** This method applies only in the absence of both a fixed or a subject's age effects. If these two conditions are met, any difference between subsequent measurements performed in the same subject does not statistically differ from zero. Actually, the differences vary around the expected value of zero as a result of the random error of measurement. Because the variance of two measures is a function of their difference, the same method used before enables us to define the upper normal range for the absolute value of the difference (\( d_{max} \))

\[ d_{max} = \sqrt{2} \times F_{0.01(1,32)} \times WMS \]

**RESULTS**

We will begin with the results concerning the repeated-measures analysis of variance and the regression analysis because a significant effect of both the within-subjects factor (T-RT), the interaction of T-RT and the between-subjects factor sex (T-RT * sex), and the regression independent variable “subject's age” should be considered in the computation of both \( \hat{R} \) and normal ranges of variability. The results concerning the T-RT factor and the T-RT * sex interaction factor are reported in detail in Table 1 (3 measurements and 12 subjects) and in Table 2 (2 measurements and 20 subjects). For all parameters, we did not detect a significant effect either for the T-RT factor or for the T-RT * sex interaction factor. Thus, for all parameters, the differences observed in a series of measurements accomplished in the same subjects are explainable by a random error of measurement only, and this was true for both sexes.

None of the absolute values of DIF1, DIF2, and DIF3 was linearly related with the subject’s age (regression analysis), and no other trend was identifiable from the scatterplot of the data. For each parameter, the variance of DIF1, DIF2, and DIF3 was homogeneous using both Bartlett-Box F and Cochran’s C tests.

Then, we computed \( \hat{R} \) values and the normal range for both the variance and the difference of repeated measures; these data are reported in Table 3.
The reliability of a parameter should always be ascertained and is mandatory when the parameter does not derive directly from a measurement but from the "manipulation" of raw data (computation of mean values or coefficients derived from regression analysis). Synthesis is needed because raw data are not informative for clinical purposes. We used R, the estimate of the intraclass coefficient of reliability, to check the reliability of saccade and smooth pursuit parameters. Concerning saccades, R values were above or slightly below the threshold of excellence for all parameters considered, with the exception of the amplitude/peak velocity relationship coefficients. Thus, quantitative evaluation of saccades is reliable and can be used in longitudinal studies. However, it seems preferable to make a selection among saccade parameters. For clinical purposes, both the amplitude/duration and the amplitude/peak velocity relationships are regarded as a neurophysiologic marker of the activity of burst cells lying in the paramedian pontine reticular formation,21 and attention is usually focused on a maximum velocity value corresponding to a theoretic saturation level of velocity that cannot be exceeded and that corresponds to the inverse of the coefficient d, 1/d, in our equation. However, as shown by our data, in longitudinal studies, it would be better to consider the coefficients from the amplitude/duration relationship, which are much more reliable. This idea was indirectly supported by a report, which recorded saccadic eye movements in patients with multiple sclerosis during an acute phase of the disease before and after corticosteroid treatment.10 Although the data on both peak

**DISCUSSION**

This table also reports the average subject coefficient of variation.

In short, the coefficients a and b of the saccade amplitude/duration relationship, the coefficients m and n of the smooth pursuit target velocity/gain relationship, and the saccade latency showed excellent R values, the saccade precision showed good R values, and we observed the lowest, but acceptable, R values for the amplitude/peak velocity R value coefficients. These data were partly confirmed by the average of individual subject's coefficients of variation, which followed the same ranking of R, with the exception of the lowest coefficient of variation observed for precision.

**TABLE 1. Repeated Measures MANOVA for Three Measurements**

<table>
<thead>
<tr>
<th></th>
<th>Mean 1</th>
<th>SD 1</th>
<th>Mean 2</th>
<th>SD 2</th>
<th>Mean 3</th>
<th>SD 3</th>
<th>T - RT</th>
<th>T - RT × Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>26.45</td>
<td>2.94</td>
<td>26.46</td>
<td>3.26</td>
<td>27.42</td>
<td>2.80</td>
<td>1.64</td>
<td>0.246</td>
</tr>
<tr>
<td>b</td>
<td>2.56</td>
<td>0.27</td>
<td>2.65</td>
<td>0.34</td>
<td>2.57</td>
<td>0.19</td>
<td>2.82</td>
<td>0.111</td>
</tr>
<tr>
<td>c × 10⁴</td>
<td>181.23</td>
<td>19.32</td>
<td>173.69</td>
<td>19.04</td>
<td>184.12</td>
<td>14.15</td>
<td>3.83</td>
<td>0.063</td>
</tr>
<tr>
<td>d × 10⁴</td>
<td>15.16</td>
<td>1.98</td>
<td>16.16</td>
<td>2.68</td>
<td>15.82</td>
<td>2.29</td>
<td>2.20</td>
<td>0.774</td>
</tr>
<tr>
<td>P × 10²</td>
<td>94.8</td>
<td>2.3</td>
<td>93.5</td>
<td>2.3</td>
<td>94.2</td>
<td>2.20</td>
<td>3.83</td>
<td>0.063</td>
</tr>
<tr>
<td>L</td>
<td>235.42</td>
<td>30.79</td>
<td>233.28</td>
<td>33.86</td>
<td>238.98</td>
<td>33.38</td>
<td>0.26</td>
<td>0.718</td>
</tr>
<tr>
<td>m</td>
<td>103.48</td>
<td>8.56</td>
<td>103.52</td>
<td>8.61</td>
<td>101.97</td>
<td>9.59</td>
<td>0.34</td>
<td>0.720</td>
</tr>
<tr>
<td>n</td>
<td>-1.00</td>
<td>0.15</td>
<td>-1.01</td>
<td>0.16</td>
<td>-1.01</td>
<td>0.19</td>
<td>0.29</td>
<td>0.748</td>
</tr>
</tbody>
</table>

Mean values and the standard deviations for the three measurements are reported for each parameter as are the F and the P values for T - RT factor and T - RT × sex interaction.

**TABLE 2. Repeated Measures MANOVA for Two Measurements**

<table>
<thead>
<tr>
<th></th>
<th>Mean 1</th>
<th>SD 1</th>
<th>Mean 2</th>
<th>SD 2</th>
<th>T - RT</th>
<th>T - RT × Sex</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>25.44</td>
<td>3.18</td>
<td>25.61</td>
<td>3.28</td>
<td>0.27</td>
<td>0.611</td>
</tr>
<tr>
<td>b</td>
<td>2.66</td>
<td>0.37</td>
<td>2.69</td>
<td>0.38</td>
<td>4.83</td>
<td>0.045</td>
</tr>
<tr>
<td>c × 10⁴</td>
<td>176.78</td>
<td>20.81</td>
<td>168.49</td>
<td>20.07</td>
<td>3.07</td>
<td>0.063</td>
</tr>
<tr>
<td>d × 10⁴</td>
<td>15.56</td>
<td>2.22</td>
<td>16.96</td>
<td>3.07</td>
<td>4.14</td>
<td>0.055</td>
</tr>
<tr>
<td>P × 10²</td>
<td>94.3</td>
<td>2.6</td>
<td>93.7</td>
<td>2.4</td>
<td>0.23</td>
<td>0.639</td>
</tr>
<tr>
<td>L</td>
<td>237.24</td>
<td>28.27</td>
<td>235.48</td>
<td>34.59</td>
<td>0.01</td>
<td>0.923</td>
</tr>
<tr>
<td>m</td>
<td>101.87</td>
<td>9.73</td>
<td>101.77</td>
<td>11.64</td>
<td>0.641</td>
<td></td>
</tr>
<tr>
<td>n</td>
<td>-1.00</td>
<td>0.15</td>
<td>-1.00</td>
<td>0.16</td>
<td>0.641</td>
<td></td>
</tr>
</tbody>
</table>

Mean values and the standard deviations for the two measurements are reported for each parameter as are the F and the P values for T - RT factor and T - RT × sex interaction.
velocity and duration were available, only the latter were considered in the discussion.

Another important point concerns the discrepancy observed for the precision’s \( R \) and average subject coefficient of variation values. The average subject coefficient of variation does not express an absolute evaluation of variability as \( R \) does, but it allows a relative ranking among different saccade parameters. In our data, the two relative rankings are similar, with the exception of the precision, which showed the first (best) average subject coefficient of variation but only the fourth \( R \) value. This discrepancy suggests that the reliability of precision could have been underestimated by \( R \) because \( R \) is derived, not only from WMS, but also from BMS. A parameter like precision, which is stable among differing subjects (low interindividual variability), can possibly show low \( R \) values as a result of the low weight of BMS.

Concerning the coefficients of the target velocity/gain relationship, they both showed excellent \( R \) values despite the fact that the strength of this relationship \( (R^2 \text{ usually approximately } 0.7) \) was less than that of the amplitude/duration and amplitude/peak velocity \( (R^2 \text{ usually approximately } 0.95) \). This last feature is particularly important because, in general, the evaluation of smooth pursuit eye movements per se cannot distinguish between differing possible sites of the lesion, but it seems to be sensitive in detecting subclinical involvement of the central nervous system. In many reports, smooth pursuit eye movements were able to discriminate between two groups of subjects, not only when the discriminant factor was important (ie, patients with neurologic disease versus control subjects\(^{20,22-29}\)) but also when the discriminant factor was more subtle: psychophysiologic\(^{30} \) and physiologic (age) variables,\(^ {31-33} \) and differing pharmacologic treatments.\(^ {1,3,5,7,8} \)

The normal variability values deserve our final comment. To our knowledge, only one previous article has reported normal variability values to evaluate changes observed in individual subjects.\(^ {10} \) In many longitudinal studies, such changes are evaluated by comparing differing groups of subjects, rather than by considering the subjects individually. However, we believe that the evaluation of individual modifications is relevant to the outcome and the possible correlation with other clinical or instrumental data. During each recording session, a subject makes several saccades and several smooth pursuit ramps; therefore, individual evaluation could use a statistical approach based either on the t-test or on covariance analysis. However, with this approach, the changes that cannot be attributed to a chance fluctuation in a stochastic experiment would seem significant, but it would not be possible to determine whether these changes are biologically relevant. Our method is based on the WMS value which, in our opinion, represents the best estimate of intraindividual variability. In addition, the WMS computed on our data can be upgraded by additional testing of the same subjects or by repeated testing of new subjects. In addition, it provides adequate evaluation of any pair or triad of repeated measurements in which the interval between two measurements should be no longer than 4 mo or shorter than 1 week. An increased variance observed in three measurements can be understood better by studying the differences between pairs of measurements.

The standard approach to evaluate the differences between pairs of measurements considers the mean and the standard deviation of the differences. This approach implies that three measurements lead to three possible samples of differences, any of which has its own mean (slightly different from zero) and its own standard deviation. However, because none of the means of the differences significantly differs from zero (analysis of variance) and the variances are homogeneous (Bartlett-Box F and Cochran’s C tests), there is no reason to use differing ranges of normal variability for each pair of measurements.

An important question would be how the data about the reliability and upper normal limits for T–RT

---

**TABLE 3. Saccade Parameter Reliability**

<table>
<thead>
<tr>
<th></th>
<th>WMS</th>
<th>BMS</th>
<th>( \bar{R} ) (%)</th>
<th>ACV (%)</th>
<th>( s^2 ) max</th>
<th>d max</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>2.13</td>
<td>23.59</td>
<td>79.56</td>
<td>6.44</td>
<td>16.11</td>
<td>5.67</td>
</tr>
<tr>
<td>b</td>
<td>0.014</td>
<td>0.28</td>
<td>87.70</td>
<td>4.39</td>
<td>0.11</td>
<td>0.46</td>
</tr>
<tr>
<td>c ( \times 10^4 )</td>
<td>137.34</td>
<td>787.41</td>
<td>64.63</td>
<td>7.14</td>
<td>1038.63</td>
<td>45.59</td>
</tr>
<tr>
<td>d ( \times 10^4 )</td>
<td>2.53</td>
<td>13.94</td>
<td>87.55</td>
<td>10.88</td>
<td>19.13</td>
<td>6.18</td>
</tr>
<tr>
<td>P ( \times 10^2 )</td>
<td>1.47</td>
<td>12.93</td>
<td>75.01</td>
<td>1.36</td>
<td>11.11</td>
<td>4.71</td>
</tr>
<tr>
<td>L</td>
<td>89.34</td>
<td>2374.32</td>
<td>91.39</td>
<td>2.99</td>
<td>690.25</td>
<td>35.50</td>
</tr>
<tr>
<td>m</td>
<td>12.81</td>
<td>246.14</td>
<td>87.55</td>
<td>3.40</td>
<td>96.87</td>
<td>13.91</td>
</tr>
<tr>
<td>n</td>
<td>0.006</td>
<td>0.047</td>
<td>71.92</td>
<td>7.03</td>
<td>0.05</td>
<td>0.30</td>
</tr>
</tbody>
</table>

WMS, within-subjects mean square. BMS, between-subjects mean square. \( \bar{R} \) (%), estimate of the intraclass coefficient of reliability expressed as a percentage. ACV (%), average of individual’s coefficient of variation expressed as a percentage. \( s^2 \) max, maximum normal limit of variability for the variances. d max, maximum normal limit of variability for the differences.
variability would be modified by different protocols of stimulation, recording procedures and equations applied to raw data. It is likely that the reliability would increase when using recording techniques less noisy than electro-oculography (infrared reflection or search coil oculography) and would not be affected by the kind of protocol, given both an adequate control of the subject's arousal during the recording session and a not difficult task to accomplish. Concerning the use of different equations, we are not able to suggest how reliability would change; nevertheless, it is evident that reliability depends on how well the equation fits the raw data.

In conclusion, our data demonstrate that the quantitative analysis of both saccadic and smooth pursuit eye movements is reliable. The statistic WMS was used to ascertain the reliability of eye movement parameters; it also enables the determination of normal variability values that can be usefully employed for neurophysiologic longitudinal studies.

**Key Words**

saccades, smooth pursuit, reliability

**References**


25. Schalten L, Henriksson N, Pykkö L. Quantification of...
Reliability of Eye Movement Measures


