ABSTRACT

Purpose: To investigate the repeatability of the Taylor Visagraph II in analyzing eye movements made by children during reading. Methods: Twenty-two children in grades 3 to 8 (mean grade=5.1 and SD=1.57) with vision correctable to 20/20, ability to read English fluently with no diagnosis of a reading disability or a neurological disorder were recruited. Visagraph II testing with grade appropriate paragraphs was performed four times and repeated for an additional four trials on a separate visit. The repeatability of the Visagraph II was calculated, for 3 consecutive readings following a practice trial, using the intra-class correlation coefficient (ICC) and the 95% limits of agreement (LoA) for seven parameters measured by the Visagraph II. Results: The values for within and between session repeatability were similar. The ICC values for between session repeatability were as follows: number of fixations (0.69), number of regressions (0.60), span of recognition or words per fixation (0.85), duration of fixation in seconds (0.89), reading rate in words per minute (0.92), percent correct on comprehension questions (0.56), and grade equivalent (0.88). The 95% limits of agreement were as follows: number of fixations (-70.16, 59.94), number of regressions (-28.33, 22.87), span of recognition or words per fixation (-0.0418, 0.0422), duration of fixation in seconds (-0.0418, 0.0422), reading rate in words per minute (-60.58, 61.52), percent correct on comprehension questions (-28.67, 19.29), and grade equivalent (-3.967, 3.973). Conclusions: The ICC showed good to excellent repeatability for most measures on the Visagraph II. In contrast, the 95% LoA indicated large intra-subject variability for all measures. For example, the LoA for grade equivalent was approximately +/- 4 grades. The primary limiting factor for repeatability appears to be the relatively small separation between normative values for different grade levels used in this study.

KEY WORDS
Visagraph, saccades, reliability, children, reading, ocular motility

INTRODUCTION

The Visagraph II is an eye movement device produced by Taylor Associates that uses an infrared emitter and detectors mounted in goggles to evaluate eye movements made during silent or oral reading. It is one of the few clinically based devices that objectively measures eye movements during silent reading. Eye movements are recorded while the patient reads a grade appropriate passage and then a software program develops an idealized trace of the eye movements. After completing the passage the patient answers 10, yes or no, comprehension questions about the paragraph. The program then calculates several parameters including; fixations, regressions, duration of fixation, span of recognition, reading rate, comprehension questions, and grade equivalent. (see Table 1).

Research has evaluated eye movement patterns in good and poor readers. The studies generally agree that poor readers make more fixations and regressions...
when reading than good readers.\textsuperscript{2-4} However, the etiology of frequent eye movements in poor readers remains controversial.\textsuperscript{2-4} An important question to address when evaluating eye movements during silent reading is the repeatability of the tests. Without an idea of the typical individual variability when reading a grade appropriate passage, interpretation of test results becomes difficult. To date only two studies have assessed the repeatability of the Visagraph II and only in adult subjects.

Ciuffreda et al\textsuperscript{5} evaluated 30 visually normal adults (mean age of 28 years) without a history of reading disability using the Visagraph II. Each subject read 5 paragraphs standardized for high school/college adult level. For the group analysis Ciuffreda et al calculated the standard error of the mean for grade level (ranging from 1 to 18) determined by converting the relative efficiency to a grade equivalent using norms provided by Taylor. This is a non-linear conversion.

<table>
<thead>
<tr>
<th>Table 1. Definitions of data produced by the Visagraph</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fixations</strong> is the number of eye pauses per 100 words read</td>
</tr>
<tr>
<td><strong>Regressions</strong> is the number of significant right-to-left eye movements per 100 words read (excluding return sweeps)</td>
</tr>
<tr>
<td><strong>Span of recognition</strong> is the number of words read divided by the number of fixations.</td>
</tr>
<tr>
<td><strong>Duration of fixation</strong> is the duration of fixations in seconds (total reading time divided by number of fixations)</td>
</tr>
<tr>
<td><strong>Reading rate</strong> with comprehension is the reading rate (in words per minute) determined for all lines in the paragraph, excluding the first and last.</td>
</tr>
<tr>
<td><strong>Comprehension questions</strong> correct is the percentage of correct answers on 10 comprehension questions</td>
</tr>
<tr>
<td><strong>Grade equivalent</strong> is an academic grade (ranging from 1 to 18) determined by converting the relative efficiency to a grade equivalent using norms provided by Taylor. This is a non-linear conversion.</td>
</tr>
</tbody>
</table>

Colby et al\textsuperscript{6} evaluated the repeatability of the Taylor Visagraph II in analyzing eye movements in children during silent reading on two separate occasions.

## METHODS

### Subject Selection

Children, ages 8 to 15, were recruited from the clinic population at the Southern California College of Optometry Eye Care Center and from children of employees at the college. Consent was obtained from the parent or guardian of each child and assent was obtained from each child. The study protocol was approved by the Institutional Review Board at the Southern California College of Optometry.

All subjects met the following inclusion criteria: ability to read English fluently, no diagnosis of reading disability (per parental report), best corrected visual acuity of 20/20 at 40 cm, no neurological or ocular condition (per parental report and chart review) that would interfere with accurate recording on the Visagraph II.

### Procedure

The grade level used for selecting reading passages from the Visagraph II reading booklet was based on each subject’s independent or instructional reading level which was obtained from the San Diego Quick Assessment.\textsuperscript{10} This test uses lists of 10 single words for each grade level and the child is instructed to read the list. Words are scored correct if read within 2 seconds and we used a cut off value of the highest grade where the child correctly read 90% (9/10) or more of the words as the starting reading level for Visagraph II testing.
The Visagraph II goggles were fitted for alignment over the child’s habitual near correction. Subjects were seated approximately 40 cm away from the Visagraph reading booklet which was placed on a slant board. Paragraphs were randomly selected from the 10 paragraphs available for each grade level and no paragraph was read twice. For a given subject, a different paragraph was used at each trial across the two test sessions. Subjects were then instructed to silently read a paragraph at their predetermined grade level and answer 10 comprehension questions. If the child answered less than 7 questions correctly on the comprehension section then the grade level was dropped by one level. Eye movements were recorded during the reading task. Four trials, separated by one-minute resting intervals, were completed at each session. The first trial at each session served as practice and was not included in the data analysis. Subjects were asked to return in approximately one week to repeat the above procedure.

**Data Analysis**

For each subject, we used the idealized recording from the Visagraph II and recorded the fixations, regressions, duration of fixation, span of recognition, reading rate, grade equivalent, and comprehension questions (see Table 1 for definitions). We evaluated both the within and between session repeatability of the subject responses using statistical methods that are well accepted for repeatability studies. ^7^-^11^ Within session repeatability calculations were based on the three measurements taken at either session one and two. Within session repeatability was assessed by calculating the mean range and the 95th percentile of the ranges (R95). The mean range would represent the typical within session difference in the three measurements, where as, the R95 yields the upper limit of the range. In addition, we determined the intra-class correlation coefficient (ICC) using within and between subject variability obtained from a one-way analysis of variance. The ICC is an overall index of repeatability with values ranging from 0 to 1. A value of 1 indicates perfect repeatability and a value of 0 indicates no repeatability. The ICC values are commonly interpreted as follows: ICC < 0.4 indicates poor repeatability; 0.4 < ICC < 0.75 indicates fair to good repeatability; and ICC > 0.75 indicates good to excellent repeatability. ^9^, ^12^ The use of ICC’s allows the reader to compare the reliability of measures that may have different units of measurement (e.g., words/min versus number of fixations).

The between session calculations is based on differences between session 1 and 2 and uses the mean of the three measurements for each session. To assess between session variability we calculated the ICC between session 1 and 2 based on using the mean of three measures for each session. We then evaluated the distribution of the differences between test and retest measurements using the method developed by Bland and Altman which calculates the 95% limits of agreement (LoA). That is, the difference between measurement session 1 and 2 is plotted versus the mean of session 1 and 2. In addition, we calculated the median absolute differences (MAD), which is the 50th percentile of the absolute differences between session 1 and 2 measurements. The MAD provides a measure of the typical patient variability between testing on session 1 and 2, while the 95% LoA provides the worst case of the differences.

**RESULTS**

**Subjects**

Twenty-two subjects in grades 3 to 8 (mean grade=5.1, SD=1.57, mean age=11.05 yrs, SD=1.44) were recruited and completed both session 1 and 2. Two subjects only completed session 1 and were excluded from the data.
The distribution by grade was as follows: grade 3 (2), grade 4 (8), grade 5 (6), grade 7 (2), and grade 8 (3). The average time between session 1 and 2 was 8 days (SD=3.83) with a range from 1 to 15. There were 12 females and 10 males. The mean independent or instructional reading levels determined from the San Diego Quick Assessment was 4.5 (SD=1.14). The size of the sample did not allow for assessing repeatability by gender or by grade of the student and all analysis was for the entire sample of 22 children.

**Within Session Repeatability**

The mean and standard deviation for three readings for each of the seven measures of the Visagraph II taken at session 1 and 2 are presented in Tables 2 and 3. The mean range and the R95 values are also presented in Tables 2 and 3. In general, the ICC values are good to excellent for all measures except for comprehension questions where the values are poor to fair. The poor ICC result for comprehension questions may in part be attributed to the limited variability in this measure which in turn adversely impacts the ICC value. That is, when a child is reading at the independent or instructional reading level the comprehension scores are typically 70% or higher which limits the possible variation between subjects and can lead to low ICC values.

**Between Session Repeatability**

We first looked for significant biases between the mean values for session 1 and 2 and found small differences between sessions for most measures. For example, there was a small increase in fixations and regressions between session 1 and 2. None of the differences were significantly different using

### Table 4 Between session repeatability.

<table>
<thead>
<tr>
<th></th>
<th>MAD</th>
<th>95% limits of agreement (LoA)</th>
<th>ICC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixations</td>
<td>19.5</td>
<td>-70.16, 59.94</td>
<td>0.69</td>
</tr>
<tr>
<td>Regressions</td>
<td>5.0</td>
<td>-28.33, 22.87</td>
<td>0.60</td>
</tr>
<tr>
<td>Span of Recognition</td>
<td>0.08</td>
<td>-0.274, 0.282</td>
<td>0.85</td>
</tr>
<tr>
<td>Duration of Fixation (sec)</td>
<td>.015</td>
<td>-0.0418, 0.0422</td>
<td>0.89</td>
</tr>
<tr>
<td>Reading Rate (Words per Minute)</td>
<td>13.67</td>
<td>-60.58, 61.52</td>
<td>0.92</td>
</tr>
<tr>
<td>Comprehension Questions (percent correct)</td>
<td>10</td>
<td>-28.67, 19.29</td>
<td>0.56</td>
</tr>
</tbody>
</table>

Figure 1. Scatterplot of session 1 and session 2 differences vs. the mean of session 1 and session 2 for fixations. Mean is the mean difference between session 1 and 2; L is the lower bound of the 95% limits of agreement; and U is the upper bound of the 95% limits of agreement.

Figure 2. Scatterplot of session 1 and session 2 differences vs. the mean of session 1 and session 2 for regressions. Mean is the mean difference between session 1 and 2; L is the lower bound of the 95% limits of agreement; and U is the upper bound of the 95% limits of agreement.
Measures of repeatability for the Taylor Visagraph II were found for children in grades 3 through 8. These values were compared to previous studies using the Visagraph II and to normative data published in the manual of the Visagraph II.

It is difficult to compare our results to previous studies due to different data analysis procedures. Ciuffreda et al.\textsuperscript{5} reported a standard error of the mean of +/- 0.6 grades for grade equivalent score and +/- 10 words per minute for reading rate in adult patients for within session measurements. We also calculated the SEM for grade equivalent (+/- 0.5 grades) and reading rate (9.0 words per minute) which were similar to the results of Ciuffreda et al.\textsuperscript{5}

However, the SEM relates to variation of mean values based on the standard deviation and the total number of observations and does not adequately describe the within session repeatability of the measure for individual subjects. A more appropriate description of within session repeatability can be found by looking at the mean range and R95 values. We found a mean range (typical within session variability) of approximately 2 grades for grade equivalent and 38 (words per minute) for reading rate. The R95 values (worst case within session variability) were 5 grades for grade equivalent and 64 (words per minute) for reading rate.

Ciuffreda et al.\textsuperscript{5} also performed a trend analysis of the subjects by looking at variability of individual data to the entire sample and identified five sub-groups. For example, on the reading rate measure, one subgroup was stable over time where as the other groups tended to either increase or decrease in the number of words read per minute over 5 consecutive measurements. We were unable to duplicate this analysis because our data was collected over two different sessions. However, there were no significant mean biases between sessions 1 and 2 for grade level or reading rate, indicating that overall subjects were not changing across the two sessions. In addition, when looking at the scatter plots for reading rate and grade equivalent a similar number of subjects

a t-test (p < .086 for all tests). The ICC values were again good to excellent, for all the of Visagraph II measures, except for comprehension when looking at between session measurements. The LoA and the MAD for each of the seven measurements of the Visagraph II measures are also shown in Table 4. In contrast, to the good to excellent ICC values, the measures of the Visagraph II show large intra-subject variability (see discussion section for further explanation). For example, the LoA’s for grade equivalent variability was -3.97, 3.97. Figures 1 to 7 show the difference vs. mean plot for each of the measures of the Visagraph II. The lower and upper bounds for the 95% LoA and mean difference between test sessions are indicated on each plot.
increased as decreased across the two sessions. Thus, we agree with Ciuffreda et al. that looking at individual results are important to see if the subjects responses have typical variation characteristic of the MAD or higher levels of variability.

We also compared our results to Colby et al., but again, the different methods of analysis makes comparisons difficult. A comparison of our ICC values to the split half correlation coefficients used by Colby et al. show similar values for between session repeatability. Both studies show that comprehension questions had the lowest repeatability (0.56 for ICC and 0.29 for split-half). For the measures assessed in our study Colby et al. correlation values ranged from 0.69 to 0.88 and our values ranged from 0.69 to 0.92. Thus, the two studies show similar measures of between session repeatability.

A seemingly contradictory finding in this study is that the ICC values indicate good to excellent repeatability for most measures but the LoA values show large variability. For example, the ICC for grade equivalent was 0.88 (excellent repeatability), where as, the LoA was -3.97, 3.97. That is, a finding of grade 5 equivalent on the Visagraph II could vary between grades 1 to 9 and this would certainly make accurate diagnosis difficult. The problem probably lies in determining the clinical relevance from the statistical analysis of repeatability. One of the primary rationales for calculating repeatability is to determine the relationship between individual variation and values needed for making clinical decisions. If the individual variability is large compared to diagnostic values, then making clinical judgments becomes more difficult. For example, reading rate on the Visagraph II has normative data for grade 4 (158), grade 5 (173), and grade 6 (185), and grade 7 (195), and shows little separation between individual grades. The repeatability of reading rate shows an excellent ICC value of 0.92, but the MAD of 13.67 (words per minute) and the LoA of -60.58 to 61.52 (words per minute) show that individual variation is large when compared to differences between grades. Thus, making distinction between grades for reading rate is difficult because the individual variation is large when compared to differences between grade levels. This is another reason to use the LoA method because the ICC data can be misleading when applying the analysis to clinical decision making.

The difficult issue becomes how to use the repeatability data, to help with clinical decision making, when using the Visagraph II. For grades 4 to 8 most of the grade level values for measures on the Visagraph II have small separations compared to the individual variation. The relatively small difference between grades probably represents the relatively minor changes that occur in measures of reading fluency from grades 4 to 8 which were represented in our study. Studies of both silent and
oral reading fluency show rapid improvements in grades 1 to 4 with much smaller changes in grades 4 to 8. Thus, reading fluency, as assessed by the Visagraph II, may need to be evaluated using different criteria than grade equivalent. Looking at absolute values maybe more appropriate than grade equivalent. For example, when looking at reading rate, the average value for fifth grade is 173 words per minute. If a fifth grade child, who has an independent reading level of fifth grade, has a reading rate of 100 words per minute then the clinician is pretty sure that this is a clinical problem (173-61 = 112 words per minute). In contrast, if the same fifth grader has a reading rate of 155 words per minute then the clinician would probably be unsure if this is a clinical problem because this is within the normal variability for reading rate.

The repeatability data can help clinicians determine whether changes in reading eye movements made during therapy are real or accounted for by normal variability. For example, when monitoring a patient’s reading rate on the Visagraph II a change of over 61 words per minute would be necessary to be outside of normal variability as assessed by the LoA and would be considered a significant change in performance.

Our results should be viewed with some caution due to a sample that was on average in fifth grade. Our results may not apply to children in the earlier grades or for high school and college students. Future studies should assess repeatability in these groups and also evaluate methods for improving repeatability of the Visagraph II instrument. We would also recommend that the Visagraph II normative data be updated. The current normative data is from a study done in 1960 and neither the standard deviations nor the standard error of measurements for the individual measures for each grade were reported.

References