The Repeatability, Reproducibility, and Correlation of the Schirmer Test: A Comparison of Open versus Closed Eye.

Eren Ekici, MD¹, İsmail Çağatay Çağlar, MD², Esra Ün Akgümüş, RN³

¹Department of Ophthalmology, Ankara Ulucanlar Eye Training and Research Hospital
²Department of Ophthalmology, Çorum Hitit University, Faculty of Medicine,
³Sungurlu State Hospital, Eye Clinic

Corresponding Author: Eren Ekici, MD
Ankara Ulucanlar Eye Training and Research Hospital, Department of Ophthalmology,
Ankara/Turkey; Kale Mah. Ulucanlar Cad. No: 59, 06250 Altındağ
Phone: +903123126261
Fax: +903123123805
E-mail: opdrerenekici@gmail.com
ABSTRACT

**Purpose:** To compare the repeatability and reproducibility of the Schirmer test (ST) when the eyes are open (STo) and closed (STc) in previously undiagnosed patients with suggestive symptoms of dry eye.

**Methods:** In a comparative, observational series case study, 31 patients were included in the study. STo and STc were alternately applied for a total of 6 times. The ST was applied 2 times with the eyes open (S1) and closed (S2) respectively by a single ophthalmologist. Then the ST was repeated four times with the eyes open (S3, S5) and closed (S4, S6) respectively by a single nurse.

**Results:** S1, S3, and S5 were 23.4 mm, 23.7 mm, and 23.3 mm respectively. S2, S4, and S6 were 14.7 mm, 15.6 mm, and 16.6 mm respectively. STc scores were found to be statistically lower than the STo's in general (right: $t = 2.033, P = 0.048$; left: $t = 3.474, P = 0.004$). There was no statistically significant difference in the scores of the 3 tests with open eyes: S1, S3, and S5 ($p=0.462$). There was also no statistically significant difference in the scores of the 3 tests with closed eyes: S2, S4, and S6 ($p=0.05$).

**Conclusions:** Our study suggests that although administering the ST with the patient's eyes open produces higher readings than STc in patients with suggestive symptoms of dry eye; there was an acceptable reliability among tests performed open and closed. Moreover, intra-examiner reliability was higher than inter-examiner reliability for both with the eyes open and closed.

**Key Words:** Dry eye, Schirmer test, Repeatability, Reproducibility.
INTRODUCTION

Dry eye arises out of a variety of genetic, environmental and lifestyle conditions that produce changes in the tears and surface of the eyeballs. It leads to uncomfortableness, impairment in vision, and tear film instability with possible harmful effects to the ocular surface [1]. As reported by several researchers, dry eye is among the most frequently established diagnoses in an ophthalmology practice, severely reducing the patient’s quality of life [2]. Although the presence of some specific symptoms in the patient constitutes strong evidence for the diagnosis of dry eye disease, some tests should be performed in these patients. Each type of tests makes specific information available in connection with the condition of the ocular surface [3].

A “gold standard” diagnostic test is not present for dry eye disease. Plus, determining and being in unison on the most proper dry eye diagnostic tests for clinical practice is still debatable [4]. Despite the fact that a great variety of tests of tear production exist, the Schirmer test (ST) is one of the most employed methods peculiar to identify and assess the ocular tear production. In 1903, Otto Schirmer defined that uncomplicated test for the first time [5] which is still frequently applied in the office to evaluate aqueous tear production. Three variations related to ST have been described [4, 6-9]. ST-I has two branches: ST-I without anesthesia and with topical anesthesia. When performed without anesthesia, the ST-I evaluates basal tear secretion of the main lacrimal gland coupled with the trigeminal reflex tearing in which irritating nature of the filter paper makes the tear secretion develop. Whereas the function of the basal lacrimal secretion is measured by ST-I performed after topical anesthesia. ST-II test is primarily used for measuring the reflex tear secretion of the main lacrimal gland through causing an irritation on the nasal mucosa with a cotton-tipped applicator before evaluating tear production [9-11]. While this widely used test in practice is reliable and reproducible as to Prause et al. [12], it had been demonstrated to dearth of precision and reproducibility for detecting dry eye according to some other studies as well and the wide fluctuation in the test outcomes of the same person taken at the same time each day for several days has been shown [4, 6, 13]. On the other hand; in spite of it was essentially defined by Schirmer as a test to be performed with the patient seated, with open eyes, blinking freely [5], it is specified by some reports that the test is conducted with eyes open, others with eyes closed, while others remain unspecified. More recently, the effect of the eye position being open versus closed on the value of the ST has been measured, and upper ST values in the opened eye position were reported [14]. When the eyes are open, the effects of the upper/lower lid margins and eyelashes in stimulating tear secretion along with the influence of external factors such as evaporation, humidity, and temperature can be increased. All of these components can contribute to higher ST scores.

Consequently, this study aimed to contrast the repeatability and reproducibility of the Schirmer test without anesthesia when the eyes are open (STo) and closed (STc) and to investigate the correlation of these tests with each other.
METHODS

Sixty-two eyes of 31 patients who had not previously been diagnosed with dry eye and had complaints of stinging, burning, ocular fatigue, and grittiness were recruited to the study. Patients with these complaints were taken the Schirmer test in the outpatient clinic conditions. Prior to the testing sessions, informed written consent was provided from all participants. Study methods conformed to the ethical guidelines of the Declaration of Helsinki. The study was approved by the local Clinical Research and Ethics Committee (No: 232/5.5.2020).

A routine ophthalmic examination was performed for all individuals. Patients with anamnesis of dry eye diagnosis and artificial tears use, previous eye surgery history, ocular infection, ocular allergy, contact lens use, ocular medication, systemic medications (e.g., antidepressants, antihistamines, decongestants, hormone replacement therapy, drugs for acne and Parkinson's disease, etc.) known to affect tear production, previous punctal occlusion procedures, tear gland damage from inflammation or radiation, conjunctival concretions, and eyelid malpositions were excluded from the study.

To evaluate the reproducibility, the ST was performed for once by two separate examiners (doctor [S1, S2] and nurse [S3, S4]) with the eyes open and closed. In order to evaluate the repeatability, it was performed by the nurse (S5, S6) once again with eyes open and closed. In other words, ST without anesthesia was performed two times (S1, S2) with the eyes open and closed respectively by a single ophthalmologist (EE). Then, the ST was repeated four times with the eyes open and closed respectively by a single nurse (EUA) (S3, S4, and S5, S6) 6 times in total with 15 minutes space of time at a total of 3 visits. Before performing STs, any noticeable fluid was gently removed from the lower lid margin with a cotton swab each time. After a 1-minute waiting period, a Schirmer strip of filter paper was placed in the lower cul-de-sac within 2–3 mm from the lateral canthus of each eye with the patient seated and the eyes open and closed alternately without anesthesia. After 5 minutes, the strip was removed, and the amount of wetting was measured in millimeters. Patients who have readings above 40mm (i.e., the test paper was completely wetted) were excluded as a reason for the incapability to give an accurate measurement of the amount of wetting. To overcome environmental parameters such as light or temperature; all STs were performed in the same room with the nonexistence of airflow and a steady temperature (in the range 21 to 24 °C) at the same time interval of a day. To minimize the test anxiety, patients were informed well and made to be acquainted with the procedure. In order to reduce a sequence effect of the 2 tests, the participants were randomly assigned as to the order of the ST series (e.g., opened or closed). Moreover, the sequence was rotated for all patients. For instance, provided that the ST of the previous patient was initiated with open (STo); the following was initiated with closed (STc). STo and STc administered were named S1, S3, S5 (with the eyes open) and S2, S4, S6 (with the eyes closed) respectively. The patients were instructed to stay with both eyes open glancing at a higher point at the same time being permitted to blink without restriction in principle as well as requested to desist from blinking for as far as possible during STo. In addition, the patients were asked to keep the eyes closed during STc.
2.1 Statistical Analysis

Statistical Package for the Social Sciences (SPSS, Inc., Chicago, IL) version 22.0 was used for the data analysis. The "ggplot2" library was used in R package (version 3.5.0) for Bland-Altman plots. The authors presented the quantitative data as the mean ± standard deviation (SD). The normal distribution of the data was evaluated with the Shapiro-Wilks test. Since the data was not normally distributed, the non-parametric Friedman's test was used to compare 3 or more repeated measurements. Following Friedman's test, Post-hoc tests were performed as multiple comparison tests to determine the difference between groups. Following the results were received according to the sequence of the tests carried out as S1, S3, S5 (eyes open) and S2, S4, S6 (eyes closed); they were classified one by one to determine the intraclass correlation coefficient (ICC) of the 2 tests. These numerical amounts were computed from the estimates of within-and between-subject errors associated with the analysis of variance (ANOVA) using Bonferroni correction. With the purpose to ascertain the agreement between the test-retest reliability of the STo and STc measurements, the intraclass correlation coefficient (Two-way mixed model with Consistency type) with its 95% confidence interval and Bland-Altman plots with the 95% limits of agreement (LoA=mean difference ± 1.96 SD) were employed. The ICC values were interpreted as follows: poor reliability (< 0.5); moderate reliability (0.5–0.75); good reliability (0.75–0.9); excellent reliability (> 0.9). P<0.05 was considered significant.

RESULTS

Sixty-two eyes of 31 subjects, ranging in age from 18 to 69 years (average, 45.93 ± 15.93 years), were included in this study. Twenty-four of the participants were women, and 7 were men. There was no significant difference in the ages between males and females (p=0,085).

The Comparison Between the STc and STo

On each visit, STc scores were lower than STo scores to the greatest extent and were found to be statistically significant (Table 1). When the outcomes of STo and STc belonging to right and left eyes at each visit are examined, statistically significant difference in the scores of the 3 tests within right eyes open: S1 and S3 (p=1,000), S1 and S5 (p=1,000), S3 and S5 (p=1,000); within left eyes open: S1 and S3 (p=0,856), S1 and S5 (p=1,000), S3 and S5 (p=0,984); within right eyes closed: S2 and S4 (p=1,000), S2 and S6 (p=1,000), S4 and S6 (p=1,000); and within left eyes closed: S2 and S4 (p=1,000), S2 and S6 (p=0,260), S4 and S6 (p=0,580) was found respectively. There was no statistically significant difference between both inter-examiner and intra-examiner measurements when STo and STc were compared separately for the right and left eyes. The p value was above 0.05 in all measurements.

The Inter-correlation of STc and STo
The ICCs and their 95% CIs are shown in Table 2. The ICC values for STc (S2-S4, S2-S6, and S4-S6) were 0.622, 0.599, and 0.744 respectively, whereas the ICC values for STo (S1-S3, S1-S5, and S3-S5) were 0.694, 0.698, and 0.837 respectively.

The Correlation between STc and STo

The ICC values for STo and STc (S1-S2, S1-S4, S1-S6 and S3-S2, S3-S4, S3-S6 and S5-S2, S5-S4, S5-S6) were in the range of 0.270 to 0.556. All these comparisons between STo and STc had a statistically significant agreement (p<0.001) except S2-S3 (p=0.006) and S2-S5 (p=0.016). These tests were performed by two different examiners (S2-S3: ST doctor-closed and ST nurse-open [first visit]), S2-S5: ST doctor-closed and ST nurse-open [second visit], respectively). These tests also had a poor correlation (ICC was 0.314 and 0.270, respectively). The ICC of all other tests performed by the same examiners for STo and STc was around 0.5.

Inter-examiner (Reproducibility) and Intra-examiner (Repeatability) Reliability

The ICC values for inter-examiner reliability (doctor versus nurse) were S1-S3:0.694 and S2-S4:0.622. On the other hand, the ICC values for intra-examiner (nurse) reliability were S3-S5:0.837 and S4-S6:0.744.

The Agreement between STo and STc

Figures 1a and 1b present the Bland–Altman plots which investigate the inter-examiner agreement of the STo and STc. The mean of the differences and ±1.96 SD of these differences between the parameters are shown with the lines. In both plots, the measurement differences show a random distribution around zero. 95 % LoA was slightly narrower between the STo (doctor versus nurse) than the STc (doctor versus nurse). They were -14.98 to 14.34 and -17.42 to 15.68, respectively. Meanwhile two cases were not within the LoA for STo, four cases were not within the LoA for STc in the plots.

Figures 1c and 1d present the Bland–Altman plots which investigate intra-examiner agreement of the STo and STc. The mean of the differences and ±1.96 SD of these differences between the parameters are shown with the lines. In both plots, the measurement differences show a random distribution around zero. 95 % LoA was slightly narrower between the STo than the STc. They were -9.91 to 10.83 and -15.08 to 13.05, respectively. Two cases were not within the LoA both STo and STc in the plots.

DISCUSSION

The Schirmer tear test is still an important diagnostic test for the determination and evaluation of dry eye patients. The low cost, convenient accessibility, and simplicity of ST make itself the most frequently used screening test in everyday practice for the estimation of tear production under-examination patients [10, 14]. However ST was originally described with open eyes [5]; the results of practicing the test with closed eyes have been published by many authors [15, 16]. As far as we are
aware, the current study is the first to make comparisons between the repeatability and reproducibility of the 2 Schirmer tests without anesthesia.

In our study, it was revealed that mean STc scores (14.74 mm, 15.61 mm, and 16.62 mm respectively) were lower than the STo scores (23.43 mm, 23.75 mm, and 23.35 mm respectively) in general at each visit and there was a statistically significant difference (p < 0.001) between the results of two tests (Table 1). The fact that these values are all in the normal range may suggest that the patient cohort does not have dry eyes. Whereas, objective clinical signs often conflict with patient-reported symptoms [3]. A positive diagnosis of dry eye disease is often based heavily on the presence of symptoms, with the literature suggesting that symptoms are an essential component of the disease [1]. The results of our study were consistent with previous studies [8, 9] which report lower Schirmer scores in the closed eye circumstance than those in the open eye circumstance. Closing the eyes during ST may reduce the rate of blinking, ocular irritation due to eye movements over the paper strip or the impact regarding outside circumstances such as temperature, evaporation, and humidity. This one after another diminishes excess reflex tearing, which is a leading factor that endangers the reliability of ST [17]. Also, corneal sensitivity to different stimulus modalities (mechanical, thermal, and chemical, etc) were shown to significantly reduce in patients with dry eye when compared with age-matched normal subjects [18]. When the test was applied eyes closed, the physical characteristics of the stimulus delivered to eyes decreased as well and caused lower readings that might be thought to affect the reliability. The lower STc interpretations of the results which were found statistically significant in the current study confirm this decrease in reflex tearing while the eyes are closed throughout ST. When the results of three tests performed with the eyes open and closed were compared with each other, there was no statistically significant difference between them (S1-S3, S1-S5, S3-S5, S2-S4, S2-S6, S4-S6, p<0.05 for all measurements).

The ICCs estimated for STc and STo among themselves in the current study were found to be strong correlation. On the other hand, the ICC value was found to be highly statistically significant (ICC>0.6 and p<0.001 in all comparisons, Table 2) in all measurements made in the categories of both STo (S1-S3, S1-S5, S3-S5) and STc (S2-S4, S2-S6, S4-S6). Our results were inconsistent with Serin et al. [8] however only healthy patients without any symptoms were included in their study. In our study, we applied the Schirmer test to individuals with complaints that may be caused by dry eye. When STc and STo measurements were compared with each other, even though a weak correlation was found between STc and STo; Except for two of these correlations, the others were statistically significant (ICC<0.6 and p<0.001 in all comparisons except S2-S3 (p=0.006) and S2-S5 (p=0.016), Table 2). The aforementioned tests performed by two different examiners ([S2-S3: STc doctor-STo nurse], [S2-S5: STc doctor-STo nurse] respectively) also had a poor correlation (ICC was 0.314 and 0.270, respectively). The ICC value of the ST scores showed a weak correlation, but it was demonstrated in our study that there was a statistically significant correlation in these values. Despite the significant correlation, it was observed that the values found were low when the clinical practice was considered. Obtaining different scores even in consecutive tests suggested that a single ST score might be misleading in the clinic. Therefore, the ST score should be evaluated together with other findings and tests in the diagnosis of dry eye.
Although ST is used commonly in the daily ophthalmology practice, the inadequacy of accuracy and repeatability have been noted [13, 17]. In spite of the fact that any reading under 10 mm is recognized abnormal in general; according to other ophthalmologists, this test is an acceptable diagnostic tool only for severe dry eyes because of its moderate reproducibility [19], with values of less than 5 mm are being taken into account as significant by many practitioners [8]. In the current study, the reproducibility (doctor versus nurse; S1[STo Doctor]-S3[STo Nurse]: 0.694 and S2[STc Doctor]-S4[STc Nurse]: 0.622) was found to have a level of acceptable reliability. Similar results were observed for intra-examiner reliability, that is, the repeatability (nurse versus nurse; S3[STo Nurse]-S5[STo Nurse]: 0.837 and S4[STc Nurse]-S6[STc Nurse]: 0.744) was found to have a level of acceptable reliability for the Schirmer test as well. Besides, the reproducibility (0.694 versus 0.622) and repeatability (0.837 versus 0.744) of STo were found to be slightly higher than STc in the study (Table 2).

The Bland-Altman plots showed similar results with ICC in reliability. Intra-examiner reliability was higher than inter-examiner reliability. 95 % LoA was slightly narrower between the STo than the STc in the plots. The measurements within the LoA were lower in STo. According to the Bland-Altman plots as in ICC, measurements of STo were slightly more reliable than STc for both intra-examiner and inter-examiner estimations (Figure 1). In our study, the deduction that the inter-examiner reliability was lower than the intra-examiner reliability indicated that it would be appropriate for these tests to be performed by the same examiner both on repeated measurements on the same day and follow-up visits. The lower repeatability of the ST was one of the most crucial factors that reduced test reliability.

The greatness in the value of reliability in intra-examiner STs than inter-examiner's was consistent with Lee et al.'s study. It suggested that there is a significantly higher error in examinations by separate examiners than repeated examinations by one examiner [13]. The factors like the paper's contact with the eyelashes for a long period (5 minutes), the change in light or other environmental parameters (temperature, etc.), a reduction in reflex tearing in the second visit, test anxiety (produced more reflex tearing on the first test) or the disease status between visits were thought over to clarify the large inconsistencies in the reported repeatability of the ST [4, 17].

The limitation of our study was a relatively small number of participants owing to the strict criteria (e.g., absence of ocular allergy and drug use, no contact lens use, previous eye surgery etc.) and tiresome nature of the study.

In the current study, it has shown that administering the ST in patients with eyes open produced higher readings than ST with eyes closed. Considering that the study was conducted on individuals with complaints of dry eye, we would like to emphasize that the ST scores performed in patients with eyes closed seemed more realistic. But we would like to note that further studies should be conducted on which method will give a more accurate result in the clinical practice.

Although the correlations of the ST performed with either the eyes open or closed were found to be statistically significant, the correlation values were relatively low in clinical experience. The different outcomes even in consecutive tests suggested that a single ST score might be misleading in
the clinical practice. For this reason, the ST scores should be interpreted together with other clinical findings and tests for the correct diagnosis in dry eye.

Moreover, intra-examiner reliability (reproducibility) was found to be higher than inter-examiner reliability (repeatability) in the present study. Our findings proposed that the ST is recommended to be performed by the same examiner in repeated measurements on the same day and follow-ups visits.

**Conflicts of Interest:** The authors have no proprietary or commercial interest in any materials discussed in this article.

**Acknowledgements:** None.

**Funding:** None.
REFERENCES


Figure Legend

**Figure 1.** a. Bland–Altman plots which investigate the inter-examiner agreement of the STo. b. Bland–Altman plots which investigate the inter-examiner agreement of the STc. c. Bland–Altman plots which investigate intra-examiner agreement of the STo. d. Bland–Altman plots which investigate intra-examiner agreement of the STc.
Table 1. Schirmer test results with the eyes open and closed.

<table>
<thead>
<tr>
<th>Visit</th>
<th>STc</th>
<th>STo</th>
<th>p value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14.74 ± 9.80</td>
<td>23.43 ± 9.76</td>
<td>&lt; 0.001*</td>
</tr>
<tr>
<td>2</td>
<td>15.61 ± 9.62</td>
<td>23.75 ± 9.35</td>
<td>&lt; 0.001*</td>
</tr>
<tr>
<td>3</td>
<td>16.62 ± 10.42</td>
<td>23.35 ± 8.69</td>
<td>&lt; 0.001*</td>
</tr>
</tbody>
</table>

ST: Schirmer test; SD: standard deviation; STc: Schirmer test closed; STo: Schirmer test open; Visit 1: doctor’s Schirmer tests; Visit 2: nurse’s Schirmer tests (nurse’s first visit); Visit 3: nurse’s Schirmer tests (nurse’s second visit); * statistically significant.
### Table 2. Intraclass Correlation Coefficient (ICC) between Schirmer Test Results.

<table>
<thead>
<tr>
<th></th>
<th>S1</th>
<th>S2</th>
<th>S3</th>
<th>S4</th>
<th>S5</th>
<th>S6</th>
</tr>
</thead>
<tbody>
<tr>
<td>S1</td>
<td>1</td>
<td>0.450 (0.228-0.628)</td>
<td>0.694 (0.538-0.803)</td>
<td>0.474 (0.256-0.646)</td>
<td>0.698 (0.544-0.806)</td>
<td>0.470 (0.251-0.643)</td>
</tr>
<tr>
<td></td>
<td>P&lt;0.001</td>
<td>P&lt;0.001</td>
<td>P&lt;0.001</td>
<td>P&lt;0.001</td>
<td>P&lt;0.001</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td>S2</td>
<td>0.450 (0.228-0.628)</td>
<td>1</td>
<td>0.314 (0.072-0.521)</td>
<td>0.622 (0.442-0.754)</td>
<td>0.270 (0.024-0.486)</td>
<td>0.603 (0.412-0.738)</td>
</tr>
<tr>
<td></td>
<td>P&lt;0.001</td>
<td>P=0.006</td>
<td>P&lt;0.001</td>
<td>P&lt;0.001</td>
<td>P=0.016</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td>S3</td>
<td>0.694 (0.538-0.803)</td>
<td>0.314 (0.072-0.521)</td>
<td>1</td>
<td>0.535 (0.331-0.743)</td>
<td>0.837 (0.743-0.899)</td>
<td>0.491 (0.276-0.659)</td>
</tr>
<tr>
<td></td>
<td>P&lt;0.001</td>
<td>P=0.006</td>
<td>P&lt;0.001</td>
<td>P&lt;0.001</td>
<td>P&lt;0.001</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td>S4</td>
<td>0.474 (0.256-0.646)</td>
<td>0.622 (0.442-0.754)</td>
<td>0.535 (0.331-0.692)</td>
<td>1</td>
<td>0.567 (0.372-0.715)</td>
<td>0.744 (0.608-0.837)</td>
</tr>
<tr>
<td></td>
<td>P&lt;0.001</td>
<td>P&lt;0.001</td>
<td>P&lt;0.001</td>
<td>P&lt;0.001</td>
<td>P&lt;0.001</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td>S5</td>
<td>0.698 (0.544-0.806)</td>
<td>0.270 (0.024-0.486)</td>
<td>0.837 (0.743-0.899)</td>
<td>0.567 (0.372-0.715)</td>
<td>1</td>
<td>0.556 (0.357-0.706)</td>
</tr>
<tr>
<td></td>
<td>P&lt;0.001</td>
<td>P=0.016</td>
<td>P&lt;0.001</td>
<td>P&lt;0.001</td>
<td>P&lt;0.001</td>
<td>P&lt;0.001</td>
</tr>
<tr>
<td>S6</td>
<td>0.470 (0.251-0.643)</td>
<td>0.599 (0.412-0.659)</td>
<td>0.491 (0.276-0.659)</td>
<td>0.744 (0.608-0.837)</td>
<td>0.556 (0.357-0.706)</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>P&lt;0.001</td>
<td>P&lt;0.001</td>
<td>P&lt;0.001</td>
<td>P&lt;0.001</td>
<td>P&lt;0.001</td>
<td>P&lt;0.001</td>
</tr>
</tbody>
</table>

*S1: Schirmer test doctor-open; S2: Schirmer test doctor-closed; S3: Schirmer test nurse-open (first visit); S4: Schirmer test nurse-closed (first visit); S5: Schirmer test nurse-open (second visit); S6: Schirmer test nurse-closed (second visit); P < 0.05

*Two-way mixed model with Consistency type was used when calculating the ICC of Schirmer test results.*