Read all three of our prestigious publications, each offering high-quality content to keep you informed with the latest developments in the field.

**The Laryngoscope**

Founded in 1896

Editor-in-Chief: Michael G. Stewart, MD, MPH

The leading source for information in head and neck disorders.

[Laryngoscope.com](http://Laryngoscope.com)

**Laryngoscope Investigative Otolaryngology**

Editor-in-Chief: D. Bradley Welling, MD, PhD, FACS

Rapid dissemination of the science and practice of otolaryngology-head and neck surgery.

[InvestigativeOto.com](http://InvestigativeOto.com)

**ENTtoday**

A publication of the Triological Society

Editor-in-Chief: Alexander Chiu, MD

Must-have timely information that Otolaryngologist-head and neck surgeons can use in daily practice.

[Enttoday.org](http://Enttoday.org)
Reliability of Grading of Facial Palsy Using a Video Tutorial With Synchronous Video Recording

Gerd Fabian Volk, MD; Rebecca Anna Schaede; Jovanna Thielker, MD; Luise Modersohn, MSc; Oliver Mothes, MSc; Charles C. Nduka, MD; Jodi Maron Barth, PT; Joachim Denzler, PhD; Orlando Guntinas-Lichius, MD

Objective: To determine the intrarater, interrater, and retest reliability of facial nerve grading of patients with facial palsy (FP) using standardized videos recorded synchronously during a self-explanatory patient video tutorial.

Study Design: Prospective, observational study.

Methods: The automated videos from 10 patients with varying degrees of FP (5 acute, 5 chronic FP) and videos without tutorial from eight patients (all chronic FP) were rated by five novices and five experts according to the House-Brackmann grading system (HB), the Sunnybrook Grading System (SB), and the Facial Nerve Grading System 2.0 (FNGS 2.0).

Results: Intrarater reliability for the three grading systems was very high using the automated videos (intraclass correlation coefficient [ICC]: SB: ICC = 0.967; FNGS 2.0: ICC = 0.931; HB: ICC = 0.931). Interrater reliability was also high (SB: ICC = 0.921; FNGS 2.0: ICC = 0.837; HB: ICC = 0.736), but for HB Fleiss kappa (0.214) and Kendell W (0.231) was low. The interrater reliability was not different between novices and experts. Retest reliability was very high (SB: novices ICC = 0.979; experts ICC = 0.964; FNGS 2.0: novices ICC = 0.979; experts ICC = 0.969). The reliability of grading of chronic FP with SB was higher using automated videos with tutorial (ICC = 0.845) than without tutorial (ICC = 0.538).

Conclusion: The reliability of the grading using the automated videos is excellent, especially for the SB grading. We recommend using this automated video tool regularly in clinical routine and for clinical studies.

Key Words: Facial nerve, facial paralysis, facial grading, assessment, outcome measure, interrater reliability.

Level of Evidence: 4

Laryngoscope, 129:2274–2279, 2019

INTRODUCTION

Photographs and video recordings are part of the assessment of patients with facial palsy (FP). It has been demanded to take standardized photography series and videos, including the face at rest and with attempts to perform a set of important mimic functions (e.g., eye closure, smiling, pursing the lips, brow raising).1–3 This documentation should be repeated at regular intervals after initiation of therapy and be part of the outcome assessment.4 However, there still is no consensus on the sequence of photographic and video documentation required for patients with FP. The assessment also includes the use of a clinician-graded system of facial function. Most commonly used systems such as the House-Brackmann system (HB),5 the Sunnybrook Facial Grading system (SB),5 or the Facial Nerve Grading System 2.0 (FNGS 2.0)7 can be applied when the patient is sitting in front of the therapist or applied subsequently using the photographic or video documentation.8 Software tools such as the FACE-gram definitively need photographs or a video.9

Reliability of all clinician-graded systems for facial palsy is influenced by interrater and intrarater variability. The variability is not only based on the raters’ subjectivity; the variations caused by the patients and performance of the mimic movements also influence the accuracy of subjective and computer-based objective ratings.

Therefore, we have recently introduced an easy-to-use, fast, self-explanatory video tutorial presented to the patient without intervention of the therapist.10 The patient is filmed synchronously. Thereby, a standardized video is recorded that allows for an assessment with standard grading systems. In the present study, we tested the reliability of clinician-graded rating of facial function with this new standardized video tool when using the HB, SB, and FNGS 2.0.
MATERIALS AND METHODS

Patients and Raters

The study protocol was approved by the institutional ethics committee. All patients enrolled in the study provided written informed consent. Patients were selected from the Facial Nerve Center Jena, Jena University Hospital, Germany, with the intention to recruit a broad spectrum of severities of peripheral acute and chronic facial palsy. The inclusion criteria consisted of proven facial palsy upon completion of the diagnostics, age ≥ 18 years, and written informed consent. Diagnostics included a clinical examination and electrodiagnostics of the facial nerve and facial muscles on the affected side. Ten patients were included (6 female, median age: 54 years) to record videos with the standardized video tutorial at the Facial Nerve Center Jena. Five consecutive patients with an acute FP (≤ 3 months after onset) and five consecutive patients with a chronic palsy (> 3 months after onset) were included. Five patients were recorded two times within 1 hour to analyze the retest reliability. In addition, eight further patients (6 female; all chronic palsy) were recorded at Queen Victoria Hospital, East Grinstead, United Kingdom, with a standardized video setting; however, no patient tutorial with synchronous automated recording. The etiologies varied between idiopathic (n = 8), tumorous (n = 5), traumatic (n = 3), and infectious (n = 2). Five medical students with no experience in facial grading (novices) and five experts (staff members of the Facial Nerve Center Jena; > 3 years of experience with facial nerve grading systems) rated the video material.

Setting of the Video Tutorial With Synchronous Video Recording

The development and first experience with patient video tutorial with synchronous recording of facial movements of the patients was published recently.10 The video tutorial is available in German and English. The German version was used for the present study because German was the native language of all participants. A roll cart with two monitors and a personal computer (PC 5000-2011V3, Jecosys, Jena, Germany; Zotac graphics card GTX970 and sound system, 4 GB PCIe, Linux operating system) with a webcam and autofocus function (C930e, Logitech, Munich, Germany) were used. One monitor with integrated loudspeaker was used for the patient, and the other one was turned away on the opposite side of the roll cart for the videographer. The videographer used the second monitor to control the positioning of the patient in front of the webcam and the other monitor, as well as to start the video tutorial and synchronous recording. The recording were stored...


at 30 frames per second in .png image format and then converted into a video file in .avi format. The patient’s monitor was framed with LED-light panels (4 × 40 cm, 15 W/750 lm, Müller-Licht, Lilienthal, Germany) to illuminate the patient’s face in an otherwise darkened room. The webcam in the middle of the upper edge of the monitor had a standard distance of 26 cm to the nasion of the sitting patient. Free webcam software (Cheese; version 3.14.2, Google LLC, Mountain View, CA) was used to control the optimal distance. The video tutorial was displayed on the upper half of the screen to guarantee that the patient was in an upright position during the video recording (Fig. 1, Fig. 2). The video lasted 11 minutes and 10 seconds (Supporting Video S1). It was divided in four parts: 1) introduction; 2) face at rest and 11 different facial muscle movements (Supporting Fig. S1), each movement repeated three times; 3) repeating sentence; and 4) spontaneous emotions. For the standard setting without tutorial, the sitting patients were recorded by a photographer with a video camera (DSR-PD170P, Sony, Tokio, Japan) in a distance of 105 to 110 cm, always in the same room with neon ceiling lighting in front of a gray photography backdrop.

Facial Nerve Grading of the Videos

The videos were used to classify the facial palsy of the patients with three standard grading systems: HB,5 SB,6 and FNGS 2.0.7 HB is a gross six-point facial grading system (I = normal; VI = total paralysis). SB is a regional weighted system that rates three subscores: 1) resting symmetry, 2) the degree of voluntary facial muscle movement, and 3) involuntary muscle contraction (synkinesis). The three subscores are used to calculate a composite score (0 = total paralysis; 100 = normal function). FNGS 2.0 is a revision of HB. FNGS 2.0 is a regional system of four regions (brow, eye, nasolabial fold, and oral region) and gives each region a score of 1 (= normal) to 6 (= no movement). The total score reaches from 4 (= normal) to 24 (= complete paralysis). Synkinesis is scored additionally across the face on a scale of 0 (= none) to 3 (= disfiguring synkinesis). The novices received a manual and were introduced to the three grading systems. Each video could be regarded as long as wanted, but after change to the next video it was not possible to go back to the previous one. The novices classified the same videos 4 weeks later for a second time to analyze the intrarater reliability. To compare the grading of videos recorded while instructing the patient with the video tutorial with standard nonautomated videos, the same five novices graded also the standard nonautomated videos according to SB. To increase the comparability between the two patient samples, the video tutorial group was divided into two subgroups. For the comparison with the nonautomated videos, we only considered the subgroup containing the patients with chronic facial palsy and synkinesis (4 female, 1 male).

Statistical Analysis

All statistical analyses were performed using IBM SPSS, version 25.0.0.0 (Chicago, IL). Spearman’s correlation was used to analyze the bivariate correlation of the results of the three grading systems to each other. Intraclass correlation coefficient (ICC) statistics expressed with confidence intervals (CI) was used to analyze the intrarater reliability in between the novices as well as for the interrater reliability for the interval scaled ratings of the SB and FNGS 2.0 gradings. ICC was also used to analyze the retet reliability of the repeated SB and FNGS 2.0 ratings. HB represents an ordinal scale. Therefore, a retet reliability testing was not feasible. Cohen’s weighted kappa was additionally used analyze the intrarater reliability of the ordered scaled HB gradings. To analyze the interrater reliability of the repeated HB gradings, Kendall coefficient of concordance (W) and also Fleiss kappa were calculated. P values < 0.05 were considered significant.

RESULTS

Facial Nerve Grading

A rating using the HB, SB, and FNGS 2.0 gradings was feasible for all videos. The average gradings are summarized in Supporting Table S1. Reflecting the variability of patients with acute incomplete to complete facial palsy, and also long-term facial palsy without and with severe
synkinesis, the full range of scoring was represented. The rating of the three grading systems correlated highly to each other (SB vs. FNGS 2.0: $\rho = -0.939$, $P < 0.001$; SB vs. HB: $\rho = -0.957$, $P < 0.001$; HB vs. FNGS 2.0: $\rho = 0.963$, $P < 0.001$).

### Intrarater Reliability In Between Novices

The intrarater reliability in between the novices was overall very high ($ICC > 0.9$) (Table I). The intrarater reliability was slightly higher for the grading with the SB than with the FNGS 2.0 or HB.

### Interrater Reliability for Novices and Experts

The interrater reliability both for novices and experts was overall good to very high ($ICC = 0.8$ to $>0.9$) (Table II). Again, the results were best when using the SB for the rating. The ICCs for the FNGS 2.0 were better than for HB. The concordance coefficients Fleiss’ kappa and Kendall $W$ were low for HB (kappa < 0.3; $W < 0.4$). The SB subscores were also analyzed (Supporting Table SII): ICC for the assessment of the resting symmetry was moderate to good ($ICC = 0.6$ to $0.7$). The subscore for voluntary movement showed excellent results ($ICC > 0.9$). The assessment of the synkinesis was good ($ICC = 0.8$ to $1.0$).

### Retest Reliability of Repeated Video-Based Ratings

The retest reliability showed very good correlations both for ratings by the novices and by the experts ($ICC > 0.9$) (Table III).

### Comparison of the Results Using Automated Videos With Patient Tutorial Versus Videos Without Tutorial

The intrarater reliability was better for the ratings with the automated videos taken following the tutorial
TABLE III.
Retest Reliability for the Sunnybrook and FNGS 2.0 Grading Systems Rated by Novices and Facial Palsy Experts.

<table>
<thead>
<tr>
<th>Sunnybrook</th>
<th>ICC</th>
<th>95% CI [lower; upper]</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novice 1</td>
<td>0.980</td>
<td>[0.059; 0.999]</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Novice 2</td>
<td>0.940</td>
<td>[0.126; 0.998]</td>
<td>0.037</td>
</tr>
<tr>
<td>Novice 3</td>
<td>0.987</td>
<td>[0.644; 1.000]</td>
<td>0.009</td>
</tr>
<tr>
<td>Novice 4</td>
<td>0.998</td>
<td>[0.959; 1.000]</td>
<td>0.001</td>
</tr>
<tr>
<td>Novice 5</td>
<td>0.992</td>
<td>[0.874; 1.000]</td>
<td>0.004</td>
</tr>
<tr>
<td>All novices</td>
<td>0.979</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expert 1</td>
<td>0.981</td>
<td>[0.148; 1.000]</td>
<td>0.002</td>
</tr>
<tr>
<td>Expert 2</td>
<td>0.998</td>
<td>[0.953; 1.000]</td>
<td>0.001</td>
</tr>
<tr>
<td>Expert 3</td>
<td>0.980</td>
<td>[0.403; 0.999]</td>
<td>0.014</td>
</tr>
<tr>
<td>Expert 4</td>
<td>0.913</td>
<td>[0.040; 0.998]</td>
<td>0.022</td>
</tr>
<tr>
<td>Expert 5</td>
<td>0.950</td>
<td>[0.013; 0.999]</td>
<td>0.034</td>
</tr>
<tr>
<td>All experts</td>
<td>0.964</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FNGS 2.0</th>
<th>ICC</th>
<th>95% CI [lower; upper]</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novice 1</td>
<td>0.970</td>
<td>[0.472; 0.999]</td>
<td>0.009</td>
</tr>
<tr>
<td>Novice 2</td>
<td>0.986</td>
<td>[0.685; 1.000]</td>
<td>0.004</td>
</tr>
<tr>
<td>Novice 3</td>
<td>0.985</td>
<td>[0.589; 1.000]</td>
<td>0.004</td>
</tr>
<tr>
<td>Novice 4</td>
<td>0.995</td>
<td>[0.923; 1.000]</td>
<td>0.003</td>
</tr>
<tr>
<td>Novice 5</td>
<td>0.960</td>
<td>[0.507; 0.999]</td>
<td>0.020</td>
</tr>
<tr>
<td>All novices</td>
<td>0.979</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Expert 1</td>
<td>0.959</td>
<td>[0.502; 0.999]</td>
<td>0.019</td>
</tr>
<tr>
<td>Expert 2</td>
<td>0.965</td>
<td>[0.551; 0.999]</td>
<td>0.018</td>
</tr>
<tr>
<td>Expert 3</td>
<td>1.000</td>
<td>[1.000; 1.000]</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Expert 4</td>
<td>0.983</td>
<td>[0.564; 1.000]</td>
<td>0.004</td>
</tr>
<tr>
<td>Expert 5</td>
<td>0.938</td>
<td>[0.324; 0.998]</td>
<td>0.031</td>
</tr>
<tr>
<td>All experts</td>
<td>0.969</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Retest reliability analysis was not feasible for the House-Brackmann grading because the House-Brackmann grading represents ordinal scale.

**ICC = intraclass correlation coefficient.**

Disability grading is essential for the initial, follow-up, and outcome assessment of patients with FP. Typically, the grading is performed by the therapist while sitting in front of the patient or is performed later based on photographic or video documentation. Standardized documentation of the facial mimic function allows an independent assessment of the palsy. Recording of a video in the clinical routine setting is time-consuming and binds the capacity of one therapist during the recording process. Furthermore, the introductions and motivation of therapist influences the performance of the patient with FP. The presented patient video tutorial with synchronous recording avoids these problems and optimally standardizes the video sequences. The financial expenses (standard computer, screen, webcam) are manageable and in any case are widely available. The tutorial explains and records the face at rest and during all important facial movement, as well as while the patient articulates standard phrases and during involuntary emotional smiling; that is, it fulfills all recommendations for an optimal video allowing all conceivable outcome measures.

For grading, the HB was used because it probably still is the most commonly used system for facial grading. SB and FNGS 2.0 were chosen because these two gradings have been evaluated as the most robust grading instruments in regard of reproducibility, low interrater, and intrarater variability. The present study proved that facial grading with these three tools is easy and very feasible with automated standardized videos for facial nerve experts as well as novices. The sample size was too small to allow a statistical meaningful comparison of experts and novices. Confirming the results of a recent systematic review, SB was the most robust grading system. Therefore, we also recommend SB as the standard in reporting outcomes of facial nerve disorders for acute FP and chronic FP with and without synkinesis.

Overall, the intrarater reliability for the SB grading with the automated videos was very high (average ICC = 0.967). The same results are reported for video-recorded patients in Finland and the United States. Also in accordance to prior studies, the intrarater reliability for the HB grading (ICC, 0.931, kappa = 0.845) is typically lower. Data for the intrarater reliability of the FNGS 2.0 was calculated for the first time, showing that the intrarater reliability was also high when using this grading system (average ICC = 0.931). The intrarater reliability was also very high for the SB when using the new automated video recording system (average ICC = 0.921). Similar results were reached in previous studies with standard videos. The same can be stated for the FNGS 2.0. The intrarater reliability was high (average ICC = 0.931). Similar values were

TABLE IV.
Grading Using Videos With Tutorial and Synchronous Recording Versus Standard Videos.

<table>
<thead>
<tr>
<th>Rater</th>
<th>ICC</th>
<th>95% CI [lower; upper]</th>
<th>P Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Novices, 1st rating</td>
<td>0.934</td>
<td>[0.832; 0.981]</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>Novices, 2nd rating</td>
<td>0.902</td>
<td>[0.762; 0.971]</td>
<td>&lt; 0.001</td>
</tr>
<tr>
<td>All</td>
<td>0.918</td>
<td>0.845</td>
<td></td>
</tr>
</tbody>
</table>

CI = confidence interval; ICC = intraclass correlation coefficient; NA = not applicable.
reached in a study by Vrabec et al. when establishing the FNGS 2.0 instrument (ICC = 0.985).\(^7\) The interrater reliability for the HB was much less concordant (Fleiss’ kappa = 0.214). The same was shown in older studies by Kanerva et al. (Fleiss’ kappa = 0.34) and by Vrabec et al. (Fleiss’ kappa = 0.386).\(^7,13\) Finally, the retest reliability was also very high for the SB (average ICC = 0.972) and the FNGS 2.0 (average ICC = 0.974). We could not identify any other retest data in the literature. The sample size did not allow a statistical comparison of interrater reliability of the ratings in patients with chronic facial palsy with automated videos (ICC > 0.8; CI = 0.6 to 0.98) versus classical videos (ICC = 0.5; CI = 0.3 to 0.85). Although the average ICC was higher when using automated videos, the CIs show a considerable overlap. If there is a meaningful statistical difference, it could only be answered with a large sample size and re-sampling tests. It can be concluded that the introduced approach with automated videos is at least not worse than standard approaches dependent on interaction with a therapist. In the present study, the reliability was even better when the same raters compared automated videos to standard videos without tutorial. A limitation is that we did not take automated videos and standard videos of the same patients. The standard videos were taken in a typical clinical photography setting. We could not detect any other factor influencing the outcome except for the video tutorial. Furthermore, the comparison was limited to the SB grading. From a formal point of view, conclusion for the HB and the FNGS 2.0 cannot be drawn. A disadvantage of the automated video setting is that the tutorial needs a screen to be shown, and thus far an upright position of the patient. Although the video trolley is movable, it cannot be easily used in the hospital ward, for instance, to evaluate a bedridden patient the day after vestibular schwannoma surgery.

At minimum, and in particular for scientific purpose, photographic documentation or ideally videography should be prerequisite for objective assessment of facial function using automatic image analysis software.\(^15,16\) The small sample size limited the statistical power of the study. As a next step, we must validate the video tutorial in a multicenter setting with larger sample size. The implementation of an automated analysis is ongoing.

For daily use, we also created a short version of the tutorial (7:59). This version is intended for use in patients who are familiar with the original tutorial. It is also available for free download on: https://vimeo.com/211376483. Its use is intended for patients who are familiar with the original version.

**CONCLUSION**

The automated standard video recordings of facial movements synchronous to a patient tutorial without intervention by a therapist allow a very reliable facial nerve grading with various grading instruments. In particular, grading with the SB allows assessment with excellent intrarater, interrater, and retest reliability. The videos can also be used in the future for automated grading algorithms.

**Acknowledgments**

The authors thank Tino Tschieche and Benjamin Loschek, Multimedia Center, Friedrich-Schiller-University Jena, Jena, Germany, for the technical support creating the video tutorial. We thank Helene Kreyda, Department of Psychology, General Psychology and Cognitive Neuroscience, Friedrich-Schiller-University Jena, Jena, Germany, for translating and revising the English version of the video tutorial. We thank Lisa Wedekind, Department for Medical Statistics, Computer Science and Documentation, Jena University Hospital, Jena, for the support with statistical questions. We thank all novice raters (Tabitha Schüler, Lisa Geist, Wiebke Pittack-Puls, Wengelawit Misikire, and Benjamin Rödiger) and the expert raters not being coauthors (Katharina Geißler and Thordis Granitzka) who took part in the study.

**BIBLIOGRAPHY**


