Original Article

Comparison of the Reliability of the House–Brackmann, Facial Nerve Grading System 2.0, and Sunnybrook Facial Grading System for the Evaluation of Patients with Peripheral Facial Paralysis

Erdem Mengi[1], Cüneyt Orhan Kara[1], Fazıl Necdet Ardıç[1], Bülent Topuz[1], Ulaş Metin[1], Uğur Alptürk[1], Gökçe Aydemir[1], Hande Şenol[2]

[1]Department of Otorhinolaryngology, Head and Neck Surgery, Pamukkale University, School of Medicine, Denizli, Turkey
[2]Department of Biostatistics, Pamukkale University, School of Medicine, Denizli, Turkey

ORCID iDs of the authors: E.M. 0000-0001-8813-2443, C.O.K. 00000-0003-2219-4283, F.N.A. 0000-0003-4230-3141, B.T. 0000-0001-7895-6190, U.M. 0000-0001-5517-628X, U.A. 0000-0002-1297-5310, G.A. 0000-0002-9780-4413, H.Ş. 0000-0001-6395-7924.

BACKGROUND: To compare the reliability of the House–Brackmann (HB), Facial Nerve Grading System 2.0 (FNGS 2.0), and Sunnybrook Facial Grading System (SB) which are widely used in the evaluation of peripheral facial paralysis (PFP) patients.

METHODS: Thirty-five video-recorded adult PFP patients were included in the study. The evaluators comprised 6 physicians. Evaluations were conducted twice independently, utilizing video recordings. Simultaneously, the evaluators were asked to keep time during the evaluation. For the analysis of reliability, Fleiss' kappa coefficient was used for the HB, and the intraclass correlation coefficient (ICC) was used for the FNGS 2.0 and SB.

RESULTS: The mean evaluation time of 1 patient was found to be 1.06 ± 0.24, 1.47 ± 0.23, and 2.32 ± 0.41 minutes for the HB, FNGS 2.0, and SB, respectively. For interrater reliability, Fleiss' kappa for the HB was 0.495 and 0.403; ICC for the FNGS 2.0 was 0.966 and 0.958; ICC for the SB was 0.960 and 0.967 for the first and second measurements, respectively. For intrarater reliability, Fleiss' kappa for the HB was 0.391, 0.446, 0.564, 0.502, 0.626, and 0.455; ICC for the FNGS 2.0 was 0.87, 0.982, 0.966, 0.929, 0.933, and 0.948; ICC for the SB was 0.935, 0.96, 0.895, 0.941, 0.96, and 0.94 for the 6 raters, respectively.

CONCLUSION: In the present study, statistically high intra- and interrater correlations were found for the FNGS 2.0 and SB, while a moderate correlation was found for the HB. Although the HB seems to be more practical, it has been concluded that the FNGS 2.0 and SB are more reliable.

KEYWORDS: Facial Nerve Grading System 2.0, facial palsy, House–Brackmann, Sunnybrook

INTRODUCTION

Otorhinolaryngologists frequently encounter peripheral facial paralysis (PFP) patients during routine clinical practice. Bell's palsy is the most frequent cause of PFP; however, tumors, infections, and trauma also play a significant role in the etiology. Prognosis in cases of PFP is related to the severity of the disease, as patients with mild paralysis are more likely to recover spontaneously than those with severe paralysis.[1] As a result, many grading systems for facial paralysis have been created to date to assess the extent and progress of the disease.

The House–Brackmann Facial Nerve Grading System (HB), which was first described in 1983 and accepted as the standard system for evaluating facial nerve function by the facial nerve disorders (FND) Committee of the American Academy of Otolaryngology, is the system that is currently most frequently used.[2] On this system, according to the degree of facial nerve function impairment, patients are rated from 1 to 6, with higher grades denoting a more serious disorder. As this system does not allow regional scoring and many facial movements are rated in the same grade, alternative grading systems have been developed over time.

In 2009, the FND committee modified the HB to the Facial Nerve Grading System 2.0 (FNGS 2.0), and it recommended that patients with facial paralysis adopt this system.[3] In contrast to the HB, this version individually assesses each of the 4 face regions and whether synkinesis is present. A total patient score is then generated by adding the 2 scores. The possible range of scores with FNGS 2.0 is 4-24 points, and, like the HB, a higher score indicates greater disease severity.

The Sunnybrook Facial Nerve Grading System (SB) is one of the most extensively used systems developed to date,[4] which was proposed as a standard system for the reporting of FND based on a study conducted by the Sir Charles Bell Society in 2015.[5] In the SB, the 3 facial areas—the eye, the cheek, and the mouth—were assessed independently at rest. Moreover, each of the facial nerve's motor branches was assessed separately during voluntary movements. Additionally, the same voluntary motions were used to rate the degree of synkinesis, and a composite SB score was calculated based on the total subscores. The facial paralysis severity is rated between 0 and 100 points on the SB scale, and in contrast to the HB and FNGS 2.0, a lower composite score indicated greater disease severity.

The most important advantage of facial paralysis grading systems is that they allow rapid and simple evaluation of facial nerve functions without the need for any special equipment. The disadvantages are that inconsistencies can be seen between and within the raters due to the subjective nature of facial movement assessments.[6] However, reliability is very important in such systems. The present study compares the intra- and interrater reliability of the HB, FNGS 2.0, and SB grading systems commonly used for the evaluation of PFP patients.

MATERIAL AND METHODS

Patient Selection

In a power analysis conducted before the study, the inclusion of at least 33 subjects for an effect size (r) of 0.4 yielded a power of 80% with a 95% CI, and so 35 voluntary adult patients diagnosed with PFP were included in the study. The ethics committee of Pamukkale University approved the project (Approval No: 60116787-020/92390). The study included both patients recently diagnosed with PFP (≤3 months) and those with chronic PFP (>3 months). After obtaining the consent of the patients, the same researcher video-captured their facial movements in a well-lit outpatient clinic setting. While at rest, the patients were instructed to close their eyes, lift their brows, show their teeth, wrinkle their noses, and smile, as per the standards in the system. All the evaluations in this study were made through video recordings.

MAIN POINTS

- The House-Brackmann facial grading system is a rapid tool that is suitable for clinical use.
- The Facial Nerve Grading System 2.0 and the Sunnybrook grading system have a high intra- and inter-rater correlation, while the House-Brackmann has a moderate correlation.
- The House-Brackmann is more practical, whereas the Facial Nerve Grading System 2.0 and the Sunnybrook grading system give more reliable data.

Data Collection

The rater group comprised researchers with various years of experience in the field of otorhinolaryngology to represent the system users. For this purpose, 3 otorhinolaryngology professors and 3 otorhinolaryngology residents participated in the study. Of the residents, 2 had 1 year of experience and 1 had 2 years of experience. Before proceeding with data collection, the HB, FNGS 2.0, and SB grading forms were introduced to the raters at a contact meeting, during which the participants were trained in the use of the systems, and carried out exercises on the video recordings of 3 patients who were excluded from the study.

During data collection, the raters independently evaluated the video recordings of the patients and filled out the forms in hard copy. In each session, the raters evaluated all 35 patients on a single system. At the same time, the raters were asked to keep time during evaluation and to record grading times for each system. After the completed forms were collected, the forms for the next system were distributed with a one-week break. The raters reassessed the video recordings of the same patients using the next system. After the initial grading was completed with 3 systems, the study was interrupted for 15 days for test–retest reliability analysis. Then, the patient videos were randomly reordered and the evaluation process with 3 systems was repeated in the same way. The data collection process was completed in 2 months.

Statistical Analysis

The data were analyzed using IBM SPSS Statistics version 24.0. (IBM SPSS Corp.; Armonk, NY, USA). Continuous variables were expressed as mean ± standard deviation and median with minimum and maximum values, while categorical variables were expressed as numbers and percentages. For the analysis of inter- and intrarater reliability, Fleiss' Kappa coefficient was used for the HB system, as the scale produced ordinal rather than interval (or continuous) data, while the intraclass correlation coefficient (ICC) was used for the evaluation of the SB and FNGS 2.0 findings, as these systems reflect the ratio of total variance within the grading group. ICC 95% CI was also presented. Spearman's correlation coefficient was used to evaluate the strength of the relationship among the HB, Sunnybrook, and FSES 2.0 systems. In all analyses, $P < .05$ was considered statistically significant.

The Fleiss' kappa values were interpreted, using the generally accepted criteria, where <0.20=poor; 0.21 to 0.4=fair; 0.41 to 0.60=moderate; 0.61 to 0.80=substantial; and 0.81 to 1.0=nearly perfect.[7] The ICC values were interpreted as follows: <0.4=poor, 0.4 to 0.75=acceptable to good, and ≥0.75=excellent.[8] The coefficient r was evaluated using the following standards: 0.00-0.25=little to no correlation; 0.26-0.49=low correlation; 0.50-0.69=moderate correlation; 0.70-0.89=high correlation; and 0.90-1.00=extremely high correlation.[9]

RESULTS

Patient Population

Of the patients included in the study, 23 (65.7%) were male and 12 (34.3%) were female, and the mean age was 51.65 ± 13.65 years (minimum–maximum: 18-71 years). Nineteen patients had acute (≤3 months) and 16 patients had chronic (>3 months) PFP. The most common cause of PFP was Bell's palsy, occurring in 25 patients (71.5%),

followed by Ramsay Hunt syndrome in 4 patients (11.4%), trauma in 2 patients (5.7%), cholesteatoma in 2 patients (5.7%), and parotid tumor in 2 patients (5.7%).

Evaluation Times
While the mean time of evaluation of one patient by the raters was 1.06 ± 0.24 minutes with the HB, this time was found to be 1.47 ± 0.23 minutes with the FNGS 2.0 and 2.32 ± 0.41 minutes with the SB.

The Interrater Reliability
The data of 6 raters were calculated separately for the 2 systems for the analysis of interrater reliability (agreement), the results of which are summarized in Table 1. The Fleiss' kappa for the HB was 0.495 for the first measurement and 0.403 for the second measurement; the ICC for the FNGS 2.0 was 0.966 for the first measurement and 0.958 for the second measurement; and the ICC for the SB was 0.960 for the first measurement and 0.967 for the second measurement. According to these results, there was a statistically significant, moderate correlation in both measures in the HB system, while a statistically significant, excellent correlation was found in the FSES 2.0 and SB systems.

The Intrarater Reliability
For the analysis of intrarater reliability (repeatability), the assessments made by the 6 raters at 2 separate time points were contrasted with one another. Table 2 displays the data obtained from the 6 raters. For the HB system, the Fleiss' kappa for the 6 raters was 0.391, 0.446, 0.564, 0.502, 0.626, and 0.455, respectively, and the ICC for the 6 raters was 0.87, 0.982, 0.966, 0.929, 0.933, and 0.948 for the FNGS 2.0, and 0.935, 0.96, 0.895, 0.941, 0.96, and 0.94 for the SB, respectively. As a result, although a significant intrarater correlation was found in all raters, a low correlation was found in one rater, moderate in 4 raters, and strong in one rater for the HB system, while an excellent correlation was found in all raters for the FNGS 2.0 and SB systems.

The Relationship Between the Systems
The study revealed a significant positive correlation between HB grade and total score with FNGS 2.0 ($P=.0001$; $r=0.895$) and a significant negative correlation between the HB grade and composite SB score ($P=.0001$; $r=-0.870$). Similarly, there was a strongly significant negative correlation between the total score on FNGS 2.0 and the composite SB score ($P=.0001$; $r=-0.868$). According to these results,

Table 1. The Interrater Reliability Results of the 3 Facial Grading Systems

House–Brackmann	Fleiss' Kappa	95% CI	P
Assessment 1	0.495	0.451-0.540	<.001
Assessment 2	0.403	0.359-0.447	<.001
FNGS 2.0	**ICC**	**95% CI**	**P**
Assessment 1	0.966	0.945-0.981	<.001
Assessment 2	0.958	0.932-0.976	<.001
Sunnybrook	**ICC**	**95% CI**	**P**
Assessment 1	0.96	0.935-0.977	<.001
Assessment 2	0.967	0.946-0.981	<.001

FNGS 2.0, Facial Nerve Grading System 2.0; ICC, Intraclass correlation coefficient.

Table 2. The Intrarater Reliability Results of the 3 Facial Grading Systems

House–Brackmann	Fleiss' Kappa	95% CI	P
Rater 1	0.391	0.208-0.575	<.001
Rater 2	0.446	0.272-0.620	<0.001
Rater 3	0.564	0.394-0.734	<.001
Rater 4	0.502	0.314-0.689	<.001
Rater 5	0.626	0.447-0.806	<.001
Rater 6	0.455	0.285-0.625	<.001
FNGS 2.0	**ICC**	**95% CI**	**P**
Rater 1	0.935	0.872-0.967	<.001
Rater 2	0.949	0.898-0.974	<.001
Rater 3	0.895	0.792-0.947	<.001
Rater 4	0.941	0.883-0.97	<.001
Rater 5	0.96	0.921-0.98	<.001
Rater 6	0.931	0.866-0.964	<.001
Sunnybrook	**ICC**	**95% CI**	**P**
Rater 1	0.87	0.742-0.934	<.001
Rater 2	0.982	0.965-0.991	<.001
Rater 3	0.966	0.933-0.983	<.001
Rater 4	0.929	0.859 - 0.964	<.001
Rater 5	0.933	0.866 - 0.966	<.001
Rater 6	0.949	0.899-0.974	<.001

FNGS 2.0, Facial Nerve Grading System 2.0; ICC, intraclass correlation coefficient.

a statistically significant high correlation was found among all 3 staging systems.

The analysis of the correlation between the components of the systems revealed the highest correlation for the voluntary movement scores and the lowest correlation for the synkinesis scores. In regards to the 3 components of the HB and SB systems, there was a high correlation in the voluntary movement component ($P=.0001$; $r=-0.864$), a moderate correlation in the resting score ($P=.0001$; $r=-0.584$), and a low correlation in the synkinesis score ($P=.0001$; $r=-0.439$). An analysis of the correlation between the total scores of FNGS 2.0 and the SB system showed a high correlation for voluntary movements ($P=.0001$; $r=-0.868$), a moderate correlation for resting scores ($P=.0001$; $r=-0.577$), and a low correlation for synkinesis scores ($P=.0001$; $r=-0.439$). An analysis of the correlation between the components of the HB and FNGS 2.0 showed the highest correlation for the eye movement scores and the lowest correlation for the synkinesis scores. There was a high positive correlation between the HB and FNGS 2.0 in terms of eye region scores ($r=0.908$), whereas the correlation coefficients (r values) for the other components of the FNGS 2.0, including eyebrow, nasolabial region, mouth region, and synkinesis score, were 0.661, 0.710, 0.747, and 0.475, respectively.

DISCUSSION
Peripheral facial paralysis grading systems are crucial for the accurate and consistent evaluation of diagnostic and treatment efficacy as well as for creating a global common language. An ideal facial paralysis grading system should be appropriate for clinical use, allow regional

evaluation both at rest and during movement, evaluate the sequels of facial paralysis, and have high intra- and interrater reliability.[6,10,11]

The most distinctive advantage of the HB, which continues to be the most widely used facial paralysis grading system today, is that it is very practical. In a study by Henstrom et al, the time required to evaluate a patient using the HB scale is only 25% of the time required for the FNGS 2.0.[12] In the present study, the mean evaluation time of a patient was 1.06 ± 0.24 minutes with the HB and 1.37 ± 0.23 minutes with the FNGS 2.0, while this time increased to 2.32 ± 0.41 minutes with the SB. Based on the available findings, the HB system can be considered a rapid instrument that is suitable for clinical use. The requirement for excessive calculations in the SB system limits its use in clinical practice.

A standard facial paralysis grading system requires high interrater agreement, as the same patient may be evaluated by various doctors over time, and treatment responses are assessed and reported using such scales. The present study found good interrater agreement among the 3 systems, although it was considerably better between the FNGS 2.0 and SB systems when compared to the HB system, for which we identified a moderate level of interrater agreement (Fleiss' kappa = 0.495). This finding is consistent with those reported by Volk et al (Fleiss' kappa = 0.214), Kanerva et al (Fleiss' kappa = 0.34), Vrabec et al (Fleiss' kappa = 0.386) and Henstrom et al (Fleiss' kappa = 0.58).[4,12-14] There was an excellent interrater agreement between the FNGS 2.0 and SB systems based on ICC values, and many other studies of the SB in literature have reported a high level of interrater agreement.[14-20] That said, few studies have been done to date evaluating FNGS 2.0. In a study involving 10 patients with PFP, Volk et al reported a high level of interrater agreement for the FNGS 2.0 scale (ICC = 0.837),[13] while the study by Mengi et al evaluated the validity of the Turkish version of the FNGS 2.0 reported a high level of interrater correlation (ICC = 0.969).[21]

The results obtained by a facial paralysis scale should also be repeatable. Simply put, the results of a grading assessment performed by the same rater at different time points must be consistent unless there has been a change in disease severity. The study conducted by Ahrens et al using 2 observers revealed moderate repeatability for the HB (Fleiss' kappa = 0.534) and excellent repeatability for the SB (ICC = 0.959).[22] Kanerva et al compared the HB and SB scales and reported high repeatability for the 2 systems.[15] The present study, therefore, compared the intrarater reliability of the grading systems and found an excellent correlation between the FNGS 2.0 and SB in all raters, compared to a significant correlation only in 1 rater for the HB system.

In the present study, a good correlation was identified between the scales, with the highest correlation between the components of the scales being observed for the voluntary movement component, and the lowest correlation for the synkinesis component. Furthermore, an analysis of the correlation between the components of the HB and FNGS 2.0 revealed the highest correlation to be in the eye movement component. Based on these results, it can be speculated that the assessments are based on voluntary movements rather than resting symmetry and synkinesis, and among the voluntary movements, evaluations of eye movements are predominant when using the HB system for the assessment of the patients. For this reason, using the

FNGS 2.0 or SB systems rather than the HB can produce more reliable results in cases where the branches of the facial nerve are affected, such as in patients with traumatic facial paralysis.

In conclusion, the present study found a high intra- and interrater correlation for the FNGS 2.0 and SB systems and a moderate correlation for the HB system. Based on the available results, the HB system can be said to be more practical, while the FNGS 2.0 and SB systems can be suggested to produce more reliable data. Therefore, the present study's authors recommend using FNGS 2.0 and the SB systems for more reliable results in the evaluation of PFP patients.

REFERENCES

1. Peitersen E. Bell's palsy: the spontaneous course of 2,500 peripheral facial nerve palsies of different etiologies. *Acta Otolaryngol Suppl*. 2002;549(549):4-30.
2. House JW, Brackmann DE. Facial nerve grading system. *Otolaryngol Head Neck Surg*. 1985;93(2):146-147. [CrossRef]
3. Vrabec JT, Backous DD, Djalilian HR, et al. Facial nerve grading System 2.0. *Otolaryngol Head Neck Surg*. 2009;140(4):445-450. [CrossRef]
4. Ross BG, Fradet G, Nedzelski JM. Development of a sensitive clinical facial grading system. *Otolaryngol Head Neck Surg*. 1996;114(3):380-386. [CrossRef]
5. Fattah AY, Gurusinghe ADR, Gavilan J, et al. Facial nerve grading instruments: systematic review of the literature and suggestion for uniformity. *Plast Reconstr Surg*. 2015;135(2):569-579. [CrossRef]
6. Katsumi S, Esaki S, Hattori K, Yamano K, Umezaki T, Murakami S. Quantitative analysis of facial palsy using a three-dimensional facial motion measurement system. *Auris Nasus Larynx*. 2015;42(4):275-283. [CrossRef]
7. Landis JR, Koch GG. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159-174. [CrossRef]
8. Fleiss LS. *The Design and Analysis of Clinical Experiments*. New York: Wiley and Son; 1986:1-32. Munro BH. *Statistical Methods for Health Care Research*. Philadelphia: Lippincott; 1997.
9. Akoglu H. User's guide to correlation coefficients. *Turk J Emerg Med*. 2018;18(3):91-93. [CrossRef]
10. de Ru JA, Braunius WW, van Benthem PP, Busschers WB, Hordijk GJ. Grading facial nerve function: why a new grading system, the MoReSS, should be proposed. *Otol Neurotol*. 2006;27(7):1030-1036. [CrossRef]
11. Kang TS, Vrabec JT, Giddings N, Terris DJ. Facial nerve grading systems (1985-2002): beyond the House-Brackmann scale. *Otol Neurotol*. 2002;23(5):767-771. [CrossRef]

12. Henstrom DK, Skilbeck CJ, Weinberg J, Knox C, Cheney ML, Hadlock TA. Good correlation between original and modified House Brackmann facial grading systems. *Laryngoscope*. 2011;121(1):47-50. [CrossRef]

13. Volk GF, Schaede RA, Thielker J, et al. Reliability of grading of facial palsy using a video tutorial with synchronous video recording. *Laryngoscope*. 2019;129(10):2274-2279. [CrossRef]

14. Hu WL, Ross B, Nedzelski J. Reliability of the Sunnybrook Facial Grading System by novice users. *J Otolaryngol*. 2001;30(4):208-211. [CrossRef]

15. Kanerva M, Poussa T, Pitkäranta A. Sunnybrook and House-Brackmann facial grading systems: intrarater repeatability and interrater agreement. *Otolaryngol Head Neck Surg*. 2006;135(6):865-871. [CrossRef]

16. Neely JG, Cherian NG, Dickerson CB, Nedzelski JM. Sunnybrook facial grading system: reliability and criteria for grading. *Laryngoscope*. 2010;120(5):1038-1045. [CrossRef]

17. Mengi E, Kara CO, Ardıç FN, et al. Validation of the Turkish version of the Sunnybrook facial grading system. *Turk J Med Sci*. 2020;50(2):478-484. [CrossRef]

18. Pavese C, Tinelli C, Furini F, et al. Validation of the Italian version of the Sunnybrook Facial Grading System. *Neurol Sci*. 2013;34(4):457-463. [CrossRef]

19. Neumann T, Lorenz A, Volk GF, Hamzei F, Schulz S, Guntinas-Lichius O. Validation of the German version of the Sunnybrook facial grading system. *Laryngorhinootologie*. 2017;96(3):168-174. [CrossRef]

20. Sanchez-Cuadrado I, Mato-Patino T, Morales-Puebla JM, et al. Validation of the Spanish version of the Sunnybrook facial grading system. *Eur Arch Otorhinolaryngol*. 2023;280(2):543-548. [CrossRef]

21. Mengi E, Kara CO, Ardıç FN, et al. Validation of the Turkish version of the facial nerve grading System 2.0. *Turk Arch Otorhinolaryngol*. 2020;58(2):106-111. [CrossRef]

22. Ahrens A, Skarada D, Wallace M, Cheung JY, Neely JG. Rapid simultaneous comparison system for subjective grading scales grading scales for facial paralysis. *Am J Otol*. 1999;20(5):667-671.