Creation and validation of a photonumeric scale for assessment of lip fullness

Patrick Trevidic MD1 | Wayne Carey MD, FRCP, FAAD2 | Anthony V. Benedetto DO, FACP3 | John H. Joseph MD4 | Laura Eaton RN, BSN5 | Stéphanie Antunes PhD6 | Pauline Maffert MSc6

1Expert 2 Expert, Paris, France
2Siena Medical Research Corporation, Montreal, QC, Canada
3Department of Dermatology, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania, USA
4Clinical Testing of Beverly Hills, Beverly Hills, California, USA
5UltaMed Corporation, Fort Lauderdale, Florida, USA
6Teoxane SA, Geneva, Switzerland

Abstract

Background: Validated, objective clinical scales are needed to assess aesthetic improvement of the lips after augmentation with dermal fillers.

Objective: To develop a lip fullness rating scale and establish its reliability for grading subjects in clinical trials or routine practice, and sensitivity for detecting clinically meaningful changes.

Methods: The Teoxane Lip Fullness Scale (TLFS), a proprietary, 5-grade photonumeric scale, was developed by clinical experts based on real subject photographs and was validated through both photographic and live subjects’ evaluation.

Results: Clinician intra- and inter-rater agreement for the TLFS was substantial to almost perfect. Mean intra-rater weighted Kappa score between the two rounds of photographic validation was 0.92, and inter-rater agreement was substantial with an ICC of 0.93 for the combined rounds. Average intra-rater weighted Kappa score and inter-rater ICC for the live validation were equally high, reaching 0.91 and 0.89 respectively. Additionally, evaluators identified clinically significant differences between photographs of subjects presenting a 1-grade or 2-grade difference on the scale in 90% and 98% of cases, respectively.

Conclusions: The intra-rater Kappa scores and inter-rater ICC met their predetermined acceptance criteria of >0.70 in the photographic and live validation. The TLFS was shown to be a repeatable and reproducible Clinician Reported Outcome (Clin-RO) for healthcare providers to classify lip fullness both in clinical trials and in routine patient care. A 1-grade difference on the TLFS can detect a clinically meaningful difference in lip fullness.

Keywords
dermal filler, facial, measuring methods
The lips play a significant role in facial esthetic perception and are important for social interaction. Lip augmentation is one of the most requested minimally invasive cosmetic procedures. With the availability of reversible hyaluronic acid (HA) dermal fillers, lip enhancement, and augmentation procedures have become increasingly popular. According to the American Society of Plastic Surgeons (ASPS), soft tissue fillers were number two among the top five cosmetic minimally invasive procedures in the United States, accounting for over 3.4 million esthetic procedures in 2020, of which more than 2.6 million were hyaluronic acid (HA) filler procedures.

With the large volume of procedures requested annually, dedicated clinical trials are needed to confirm effectiveness and safety of dermal fillers in their target indications. Using objective effectiveness endpoints in these trials is essential, especially in esthetic medicine where subjectivity and lack of standardized measurements may substantially bias the outcomes. This need is usually addressed by the development of photonumeric clinical scales, illustrating several grades of the esthetic feature of interest (e.g., lip fullness, wrinkles, folds, and volume loss) through representative photographs combined with short descriptors. Scale validation aims to confirm that the developed clinical scale is adequate for measuring the esthetic endpoint of interest and guarantees repeatable and reproducible subject evaluations. United States FDA requirements set for IDE clinical trials have made scale validation a mandatory part of any premarket clinical study design.

The first publication on a Wrinkle Severity Rating Scale (WSRS) validation study, focusing on nasolabial folds, has paved the way for many other studies developing and validating clinical scales in a wide range of facial and body areas. Several lip fullness scales have been developed and validated over the past few years, all grading the upper and lower lips separately.

This article presents the development and validation of the proprietary Teoxane Lip Fullness Scale (TLFS), a Clinician Reported Outcome created for grading the overall fullness of upper and lower lips of clinical subjects using a unique score.

## Methods

### Scale development and validation process

The TLFS was developed and validated following a creation and validation scheme (Figure 1) similar to processes used in the past in the development of esthetic scales as described in the literature. A Clinical Significance Exercise (CSE) was also conducted on photographs to verify the scale’s ability to detect a clinically meaningful difference between subjects or before/after treatment. The scale was validated both photographically and in live subjects.

### Collection of photographs

A collection of photographs taken from 208 consented, volunteer subjects was obtained from QuantifiCare. Subjects were male and female, at least 21 years of age, with a variety of races/ethnicities, Fitzpatrick Skin Phototypes, and degrees of lip fullness. Photos from 196 of the 208 consented subjects were used in the scale development process. Ten subject photographs were included in the Scale itself, and 60 other subject photographs, distinct from those defining Scale, were used to create the photographic Validation Sets. The CSE utilized 78 subject photographs (39 pairs of lips). Remaining subject photographs, which were not utilized in the Scale or the Validation Sets, were used for evaluator training.

All subject photographs were taken with 2D digital photography using standardized equipment, parameters, and subject positioning. The subject photographs were at rest, with the lips not parted. Subjects were required to have general symmetry of the mouth from side to side without facial drooping. Frontal and 90-degree right photos were obtained from each subject and cropped to draw the attention to the mouth/lips. Due to the difficulty encountered in recruiting subjects with Thin and Very Thin lips, a few photographs of the Validation Sets were professionally photo-morphed.

### Image selection and scale creation

The Scale Working Group (SWG) was composed of three clinical experts (two Dermatologists and one Plastic Surgeon) who agreed on the “true value” severity grades of each subject’s overall lip fullness, grading them from 1 to 5 (Very Thin to Very Full). The group then agreed on the most representative photos and on a short text describing the key features of each grade to be used for the final layout of the 5-point TLFS (Figure 2).

### Preparation of photographic validation sets

Sixty subject photographs—different from the ten reference photographs chosen for the scale—were selected by the SWG to constitute the scale Validation Sets. Subjects were of different ages, genders, races/ethnicities, Fitzpatrick Skin Phototypes, and were distributed across all grades of lip fullness, ranging from Very Thin (Grade 1) to Very Full (Grade 5).

Photographs selected for the photographic validation were randomized and compiled into two electronic Validation Sets. Both Validation Sets were comprised of the same photographs, with a different randomization order. A third set of photographs was prepared for the CSE. The CSE booklet contained 39-paired...
subject photographs randomly paired side-by-side, some with identical clinical grades, and others with 1- or 2-point differences on the TLFS.

3.1.4 | Training of the evaluators

A member of the SWG trained the Scale Validation Group (SVG) evaluators on how to use the scale, during 1:1 WebEx trainings sessions prior to the photographic and live validation sessions.

3.1.5 | Photographic scale validation process

The scale was photographically validated by a panel of 4 expert clinicians (1 Dermatologist and 3 Plastic Surgeons). An electronic evaluation system developed by QuantifiCare was utilized for the photographic scale validation. For each of the two validation rounds at least 14 days apart, SVG members were provided with a USB stick loaded with the electronic evaluation system and the first or second Validation Set.

3.1.6 | Determination of clinical significance

The objective of the CSE was to further validate the TLFS by confirming the clinical significance of differences in adjacent grades, and the scale’s sensitivity to detect clinically meaningful changes in lip fullness.

Evaluators from the photographic SVG group were provided with an electronic copy of the CSE and were required to answer the following question for all side-by-side photographs: "Using the provided Lip Fullness Scale, please indicate if there is a clinically meaningful difference between the two photographs? (Yes/No)".

The existence of a clinically meaningful difference between each clinical grade was assessed based on the percentage of "Yes" and "No" answers calculated for side-by-side photographs showing 0-, 1- or 2-point differences between subjects. The following acceptance
criteria were set: at least 80% of evaluators had to answer “Yes” when there was a 1-grade difference; and at least 80% had to answer correctly when there was either no difference (ie, true answer = “No”) or a 2-point difference (ie, true answer = “Yes”).

3.1.7 | Live validation process

To better reflect real conditions of use of the scale, a live scale validation was performed with the assistance of a qualified vendor (Canfield Scientific, Parsippany, NJ).14,16 The validation process followed similar live validation steps as reported in the literature.3,17,18 Adult subjects with a mean age of 45.54 ± 15.61 years, with a broad diversity of ethnicity/race, Fitzpatrick Skin Phototypes, and degrees of lip fullness were recruited and photographed as part of a prescreening process. Institutional review board approval was not required as study procedures were limited to obtaining photographic images and performing live evaluations (ie, minimal risk). All subjects signed a photographic release form. 70.1% of subjects were female, with 69% being white, 16.1% Black or African American, and 14.9% other. 67% of subjects were Fitzpatrick skin phototypes I–III, while 33% were Fitzpatrick skin phototypes IV–VI.

After reviewing photographs from 164 potential candidates, the SWG confirmed the eligibility of 100 subjects, of whom 87 were enrolled for the validation sessions and an additional 10 subjects were invited to attend as training subjects.

Despite the numerous difficulties and logistical challenges brought on by the COVID-19 pandemic, two live validation sessions were held 3 weeks apart in New Jersey with the same subjects being assessed by the same five live evaluators (3 Dermatologists and 2 Plastic Surgeons).

3.2 | Statistical analysis

Kappa statistics and ICC were calculated using SAS 9.4 computer software.

To consider the TLFS validated for use in clinical studies, the acceptance criteria for Kappa statistics and ICC determining intra-rater and inter-rater agreement was set to >0.70 for both photographic and live validation.

Intra-rater reliability was assessed using weighted Kappa coefficient statistics with the formula described and validated by Cohen in 1968,19 representing individual evaluator’s rating consistency between validation rounds 1 and 2.
To determine the level of agreement between evaluators at each validation round and overall, Intraclass Correlation Coefficient (ICC) were calculated using the Shrout and Fleiss formula.20 Using live validation data, supplementary analyses of intra- and inter-rater agreements were repeated within subgroups determined by Fitzpatrick Skin Phototypes (Types I-III and IV-VI) to verify the applicability of the TLFS in diverse skin types representative of the US population.

4 | RESULTS

4.1 | Demographics

The collection of photographs for scale creation was obtained from 208 individuals, including 166 female subjects (79.8%), with a substantial part (21.2%) of darker skin phototypes (Fitzpatrick V or VI). Likewise, the 87 subjects recruited for the live validation sessions were adults aged 45.5 ± 15.6 years, mostly female (70.1%), representative of diverse ethnicities and skin types as detailed in Table 1.

4.2 | Photographic validation results

Five evaluators utilized the 5-point photonumeric TLFS to rate 60 subject photographs randomized in Validation Sets #1 and #2. Intra-rater weighted Kappa scores reached a mean value of 0.92 (Table 2). Inter-rater ICC scores reached a value of 0.94, 0.93, and 0.91 for round 1, round 2 and combined rounds, respectively (Table 3). A supplementary analysis was performed to assess the percentage of agreement with the “true value” grades provided by the SWG. The average of agreement across all five Raters was 76.3%, 80.3%, and 78.3% for Validation Sets #1, #2, and combined, respectively.

4.3 | Clinical significance exercise results

Analysis of responses to the CSE indicated that evaluators were able to detect a 1-point difference on the TLFS scale in 90% of side-by-side photographs. Additionally, evaluators provided an overall correct assessment 92% of the time when there was either no difference or a 2-point difference between subjects. These scores exceed the minimal acceptable threshold for use in a clinical trial, which was set at 80%.

4.4 | Live validation results

In the live validation, the mean weighted Kappa value for intra-rater reliability reached 0.91 (Table 4), while the ICC score for inter-rater reliability was of 0.89 for combined rounds, corresponding to a high strength of agreement between raters (Table 5). Therefore, the acceptance criterion of >0.70 to pass was met. Despite differences in sample sizes, intra-rater and inter-rater agreements within Fitzpatrick Skin Phototypes’ subgroups were equally high for fair and darker-skinned individuals, with weighted intra-rater Kappa scores of 0.83 and 0.79 (Table 4), and inter-rater ICC of 0.78 and 0.77, for Types I–III and Types IV–VI, respectively (Table 5).

### Table 1
Demographics of subjects enrolled in the TLFS live validation

| Age     | N  | Mean ± SD |
|---------|----|-----------|
|         | 87 | 45.54 ± 15.61 |

| Gender  | N  |          |
|---------|----|----------|
| Female  | 61 | (70.1%) |
| Male    | 26 | (29.9%) |

| Fitzpatrick skin type  | N  |          |
|------------------------|----|----------|
| I                      | 1  | (1.1%)   |
| II                     | 32 | (36.8%) |
| III                    | 25 | (28.7%) |
| IV                     | 12 | (13.8%) |
| V                      | 5  | (5.7%)   |
| VI                     | 12 | (13.8%) |

| Ethnicity               | N  |          |
|-------------------------|----|----------|
| American Indian or Alaska Native | 3  | (3.4%) |
| Black or African American | 14 | (16.1%) |
| Chinese                 | 5  | (5.7%)   |
| Filipino                | 2  | (2.3%)   |
| Korean                  | 2  | (2.3%)   |
| Native Hawaiian or Other Pacific Islander | 1  | (1.1%) |
| White/Caucasian         | 60 | (69.0%) |

### Table 2
TLFS photographic validation, intra-rater reliability

| Rater | Weighted Kappa (95% CI) | Percentage of agreement |
|-------|-------------------------|-------------------------|
| Rater 1 | 0.95 (0.92–0.98)       | 81.7%                   |
| Rater 2 | 0.87 (0.82–0.92)       | 63.3%                   |
| Rater 3 | 0.92 (0.88–0.96)       | 73.3%                   |
| Rater 4 | 0.93 (0.89–0.97)       | 75.0%                   |
| Rater 5 | 0.94 (0.90–0.97)       | 76.7%                   |
| All  | 0.92 (0.91–0.94)       | 74.0%                   |

### Table 3
TLFS, photographic validation, inter-rater interclass correlation coefficients (ICC)

| Validation set | Inter-rater ICC (95% CI) |
|----------------|--------------------------|
| #1             | 0.94 (0.91–0.95)         |
| #2             | 0.93 (0.90–0.95)         |
| Combined       | 0.93 (0.90–0.95)         |
TABLE 4  TLFS, live scale validation, intra-rater results by Fitzpatrick skin type

|               | Fitzpatrick I–III (n = 58) | Fitzpatrick IV–VI (n = 29) | Total (n = 87) |
|---------------|-----------------------------|-----------------------------|----------------|
|               | Weighted Kappa (95% CI)     | Weighted Kappa (95% CI)     |                 |
|               | Time 1 vs. time 2           | Time 1 vs. time 2           |                 |
|               | Percentage of agreement     | Percentage of agreement     |                 |
| Rater 1       | 0.89 (0.83–0.96)            | 0.76 (0.63–0.89)            | 0.92 (0.89–0.96) |
|               | 84.5%                       | 65.5%                       | 78.2%           |
| Rater 2       | 0.71 (0.60–0.82)            | 0.86 (0.75–0.98)            | 0.90 (0.86–0.94) |
|               | 70.7%                       | 79.3%                       | 73.6%           |
| Rater 3       | 0.88 (0.82–0.95)            | 0.61 (0.41–0.82)            | 0.90 (0.86–0.94) |
|               | 81.0%                       | 62.1%                       | 74.7%           |
| Rater 4       | 0.80 (0.70–0.90)            | 0.81 (0.68–0.94)            | 0.90 (0.85–0.94) |
|               | 75.9%                       | 75.9%                       | 75.9%           |
| Rater 5       | 0.83 (0.75–0.91)            | 0.84 (0.73–0.95)            | 0.91 (0.87–0.95) |
|               | 72.4%                       | 79.3%                       | 74.7%           |
| All           | 0.83 (0.75–0.91)            | 0.79 (0.73–0.85)            | 0.91 (0.89–0.93) |
|               | 76.9%                       | 72.4%                       | 75.4%           |

Inter-rater ICC (95%CI)

|               | Fitzpatrick I–III (n = 58) | Fitzpatrick IV–VI (n = 29) | Total (n = 87) |
|---------------|-----------------------------|-----------------------------|----------------|
| Validation set #1 | 0.78 (0.70–0.85)            | 0.71 (0.58–0.81)            | 0.87 (0.83–0.91) |
| Validation set #2 | 0.80 (0.74–0.85)            | 0.83 (0.71–0.89)            | 0.90 (0.86–0.93) |
| Validation sets combined | 0.78 (0.72–0.84)            | 0.77 (0.65–0.84)            | 0.89 (0.85–0.92) |

5  | DISCUSSION

This study allowed to successfully validate the TLFS for evaluation of the effectiveness of lip augmentation after dermal filler injection. The TLFS was created with 5 grades of lip fullness, like other available scales in this indication.18

For photographic validation, statistical analyses demonstrated repeatability and reproducibility of assessments performed using this objective measurement tool, as all intra-rater weighted Kappa scores and inter-rater ICC met the acceptance criteria of >0.70. Intra-rater and inter-rater agreements were “substantial” to “almost perfect” and comparable to those obtained for other scales in this indication.3,17,18 Furthermore, evaluators generally agreed with “true value” ratings of the subject photographs given by the SWG, with ≥78% of agreement throughout the combined Validation Sets.

Importantly, the validation cycle was repeated in live subjects by a different group of evaluators, and all validation scores were maintained in these “real-life” conditions. The average intra-rater and inter-rater agreement statistics exceeded the acceptance criteria of >0.70, indicating that the scale is adequate to assess subjects “live” in clinical studies.

Finally, results of the CSE confirmed that each grade on the TLFS represents a clinically meaningful difference in lip fullness from the adjacent grades, and that the scale has adequate sensitivity for measuring changes in lip volume.

Challenges encountered in this validation study included difficulty to recruit subjects with Thin or Very Thin lips, as these phenotypes tend to be rare. This required photo-morphing of at least one lip to generate enough photographs representative of the thinner grades in the photographic Validation Set.

Another limitation of this study is that the scale rates both lips collectively, assuming general symmetry side-to-side, which is not always the case in clinical practice; therefore, subjects with >1-grade asymmetry should be excluded from any clinical study relying on the scale. On the other hand, in eligible subjects this overall grading system may provide a more convenient and easily readable scale for clinical evaluators.

The TLFS has significant strengths, including the fact that it was successfully validated both photographically and in live subjects. This is the first to date esthetic lip fullness scale to be entirely live validated among other published scales. The live validation provides an adequate representation of real-life subjects presenting to the clinician during clinical trials or in routine practice. Another key advantage of the TLFS is the inclusion of the whole diversity of Fitzpatrick Skin Types both in the validation studies and in the final layout of the scale. Moreover, substantial intra-rater and inter-rater agreements were obtained throughout Fitzpatrick Skin Phototypes’ subgroups. This confirmed the suitability of the scale for grading lip fullness regardless of ethnicity.

Further scale validation steps will aim to correlate patient’s perception with changes on the TLFS during clinical trials. Indeed, patient-reported outcomes are increasingly considered in esthetic medicine, where no therapeutic benefit is intended hence subject satisfaction is the first treatment objective. An ongoing prospective, multicenter, double blind, randomized control trial using the TLFS as primary effectiveness endpoint will provide a comprehensive suite of measurements to help confirm the scale’s construct validity.
6 | CONCLUSIONS

The TLFS has been shown to demonstrate repeatability and reproducibility; therefore, the scale should be considered a reliable instrument for healthcare providers to classify lip fullness both in clinical trials and in routine patient care. A 1-point score difference can detect a clinically meaningful difference in lip fullness. Additionally, results of a sub-group analysis of Fitzpatrick Skin Phototypes during the live validation showed the pertinence of the TLFS in diverse skin types representative of the US and worldwide populations.

ACKNOWLEDGEMENTS

We are grateful to the following physicians for their contribution to this study, grading the photos: Drs Steven Dayan, Derek Jones, Lilly Lee and Rady Rahban; or the subjects during the live scale validation sessions: Drs David Bank, Deanna Robinson, Hilton Kaplan, Paul Lorenec and Robert Weiss, and to Axonal-Biostatem (Castres, FRANCE) for conducting statistical analyses.

CONFLICT OF INTEREST

Materials and funding for this project were provided by TEOXANE. Drs Trevidic, Carey, Benedetto and Joseph received financial compensation for their participation in this study, either as scale developers or validators. Laura Eaton is a paid consultant for TEOXANE. Stéphanie Antunes and Pauline Maffert are employees of TEOXANE.

AUTHOR CONTRIBUTIONS

All authors have read and approved the final manuscript. P.T., W.C., A.B., J.J., and L.E. performed the research. S.A. and L.E. designed the research study. P.M. contributed to data analysis and manuscript development.

ETHICAL APPROVAL

Institutional review board approval was not required as study procedures were limited to obtaining photographic images and performing live evaluations (i.e., minimal risk). All subjects signed a photographic release form.

DATA AVAILABILITY STATEMENT

Research data are not shared.

ORCID

Anthony V. Benedetto https://orcid.org/0000-0002-4280-9380

REFERENCES

1. Yazdanparast TAS, Hournaz H, Saman Ahmad N, Alireza F, Mansour NK. Assessment of the efficacy and safety of hyaluronic acid gel injection in the restoration of fullness of the upper lips. J Cutan Aesthet Surg. 2017;10(2):101-105.
2. de Maio MRB. The upper and lower lips. In Injectable fillers in aesthetic medicine, 2nd edn. Springer; 2014:114.
3. Werschler WPFS, Thomas J, Paradkar-Mitragotri D, Rotunda A, Beddingfield FC. Development and validation of a photographic scale for assessment of lip fullness. Aesthetic Surg J. 2015;35(3):294-307.
4. American Society of Plastic Surgeons. 2020 plastic surgery statistics report. https://www.plasticsurgery.org/documents/News/Statistics/2020/plastic-surgery-statistics-full-report-2020.pdf. Published 2021 Accessed July 26, 2021.
5. Day DJ, Littler CM, Swift RW, Gottlieb S. The wrinkle severity rating scale: a validation study. Am J Clin Dermatol. 2004;5(1):49-52.
6. Narins RS, Carruthers J, Flynn TC, et al. Validated assessment scales for the lower face. Dermatol Surg. 2012;38(2 Spec No.):333-342.
7. Flynn TC, Carruthers A, Carruthers J, et al. Validated assessment scales for the upper face. Dermatol Surg. 2012;38(2 Spec No.):309-319.
8. Moradi A, Lin X, Allen S, Fagien S, Norberg M, Smith S. Validation of photonumeric assessment scales for temple volume deficit, infraorbital hollows, and chin retrusion. Dermatol Surg. 2019;46(9):1148-1154.
9. Haxsel D, Fabi SG, Sattler G, et al. Validated assessment scales for cellular dimples on the buttocks and thighs in female patients. Dermatol Surg. 2019;45(Suppl 1):S2-s11.
10. Kamner MS, Casabona G, Peeters W, et al. Validated assessment scales for skin laxity on the posterior thighs, buttocks, anterior thighs, and knees in female patients. Dermatol Surg. 2019;45(Suppl 1):S12-s21.
11. Rossi AB, Nkengne A, Stamatas G, Bertin C. Development and validation of a photonumeric grading scale for assessing lip volume and thickness. J Eur Acad Dermatol Venereol. 2011;25(5):523-531.
12. Carruthers A, Carruthers J, Hardas B, et al. A validated lip fullness grading scale. Dermatol Surg. 2008;34(5):S161-S166.
13. Buchner L, Vamvakias G, Rom D. Validation of a photonumeric wrinkle assessment scale for assessing nasolabial fold wrinkles. Plast Reconstr Surg. 2010;126(2):596-601.
14. Donofrio LCJ, Hardas B, Murphy DK, et al. Development and validation of a photonumeric scale for evaluation of infraorbital hollows. Dermatol Surg. 2016;42(42):S241-S258.
15. Carruthers J, Donofrio L, Hardas B, et al. Development and validation of a photonumeric scale for evaluation of facial fine lines. Dermatol Surg. 2016;42(52):S227-S234.
16. Moradi A, Lin X, Allen S, Fagien S, Norberg M, Smith S. Validation of photonumeric assessment scales for temple volume deficit, infraorbital hollows, and chin retrusion. Dermatol Surg. 2020;46(9):1148-1154.
17. Rossi AB, Nkengne A, Stamatas G, Bertin C. Development and validation of a photonumeric grading scale for assessing lip volume and thickness. J Eur Acad Dermatol Venereol. 2011;25(5):523-531.
18. Kane MAC, Lorenz CL, Lin X, Smith SR. Validation of a lip fullness scale for assessment of lip augmentation. Plast Reconstr Surg. 2012;129(5):822e-828e.
19. Cohen J. Weighted kappa: nominal scale agreement provisi on for scaled disagreement or partial credit. Psychol Bull. 1968;70(4):213-220.
20. Shrout PE, Fleiss J. Intraclass correlations: uses in assessing rater reliability. Psychol Bull. 1979;86(2):402-428.