Journal of Language, Culture, and Translation (LCT), 7(1) (2024), 204–225

# Development and Validation of the EFL Teachers' Stroke Questionnaire

**Shabnam Kadkhodaei[1], Mohammad Reza Talebinejad[2], Mohsen Shahrokhi[*3]**

[1]*Ph.D. Candidate, Department of English Language, Shahreza Branch, Islamic Azad University, Shahreza, Iran*

[2,3]*Associate Professor, Department of English Language, Shahreza Branch, Islamic Azad University, Shahreza, Iran*

**DOI: 10.71864/LCT-2024-1221575**

***Received: 24/06/2024*** ***Revised: 25/09/2024*** ***Accepted: 27/09/2024***

**Abstract**

Classroom interpersonal recognition, commonly termed "stroke", is theorized to shape student motivation, engagement, and wellbeing, yet extant research has relied predominantly on learner-reported measures and lacks a rigorously developed teacher-self report instrument; to address this gap the present study developed and validated an EFL Teachers' Stroke Questionnaire. The study's objectives were to generate a theory-driven item pool, establish content validity through expert review, identify the instrument's latent structure via exploratory factor analysis (EFA), confirm that structure with confirmatory factor analysis (CFA) in an independent sample, and evaluate reliability and convergent/discriminant validity. Using a multi-stage scale-development design, items were drafted from Transactional Analysis and the stroke literature, reviewed by subject-matter experts, cognitively piloted with practicing EFL teachers, and administered to stratified samples for EFA and CFA (N = 124). EFA suggested a coherent four-factor solution accounting for ~59% of variance; CFA produced acceptable to good fit (CFI = .955, TLI = .947, RMSEA = .046, SRMR = .044) with standardized loadings .45–.82. Subscales demonstrated satisfactory internal consistency (αs = .79–.90; ωs = .80–.91), and convergent and discriminant validity were supported by AVE/CR indices, cross-loadings, Fornell–Larcker criteria, and HTMT ratios. The Teachers' Stroke Questionnaire therefore provided a psychometrically sound, multidimensional teacher-centered measure of stroking behaviors that can support research, teacher professional development, and intervention evaluation in EFL contexts.

*Keywords:* teacher stroke; scale development; construct validity; EFL teachers

[*] Corresponding author's email: shahrokhi1651@yahoo.com

## 1. Introduction

Interpersonal recognition in the classroom, commonly referred to as stroke in the Transactional Analysis literature, constitutes a fundamental mechanism by which teachers shape learners' affective experience, task engagement, and ultimately academic outcomes. Strokes are units of social recognition, granted through words, gestures, or symbolic acts, that satisfy basic human needs for acknowledgement and social belonging (Berne, 2011; Stewart & Joines, 1987). In educational settings, strokes operate as a form of pedagogic feedback that complements instructional input: by signaling attention, approval, or corrective guidance, teachers' stroking behaviors influence students' motivation, willingness to participate, and perceptions of teacher effectiveness (Hattie & Timperley, 2007; Gao, 2021). Despite its theoretical and practical salience, stroke remains under-measured as a teacher-centered construct in EFL research: most extant instruments capture students' perceptions of received stroke rather than the teacher's own stroking practices. This imbalance constrains our ability to investigate how teachers' self-reported stroking behaviours relate to professional outcomes (e.g., job satisfaction, burnout) and to classroom-level indicators (e.g., students' engagement and attendance). The present study therefore seeks to develop and validate a psychometrically robust EFL Teachers' Stroke Questionnaire (TSS) to fill this gap.

## 2. Literature Review

Transactional Analysis (TA) conceptualizes stroke as the minimal unit of recognition exchanged between social agents (Berne, 2011). TA's account emphasizes that recognition may be conveyed verbally (e.g., praise, questions) or nonverbally (e.g., smiles, eye contact), and that strokes can carry positive or negative valence (Stewart & Joines, 1987). Later TA-informed work differentiated strokes further into conditional (tied to performance or behaviour) and unconditional (tied to the person's characteristics) forms, an analytical distinction particularly relevant for education because conditional strokes typically reinforce task-specific behaviours while unconditional strokes communicate broader relational support (Pishghadam & Khajavy, 2014; Pishghadam et al., 2021). From the perspective of rhetorical–relational goal theory and positive psychology, strokes are interpersonal acts that help teachers pursue both rhetorical (instructional) and relational (supportive) goals; successful alignment of these goals is associated with higher student motivation and classroom engagement (Gao, 2021). Thus, theoretically, teacher stroke is a multidimensional construct that intersects modality (verbal/nonverbal),

valence (positive/negative), and contingency (conditional/unconditional), a complexity that necessitates careful operationalization when designing measurement instruments.

A growing body of empirical research, particularly from Iranian EFL contexts but increasingly internationally, documents robust links between stroke-related behaviors and student outcomes. Early instrument development focused on students' reports: Pishghadam and Khajavy (2014) produced the Student Stroke Scale, and used it to demonstrate positive associations between student-perceived stroke and intrinsic/extrinsic motivation. Building on that foundation, several Iranian studies have shown that student-reported stroke correlates with willingness to attend classes, classroom participation, and perceptions of teacher success (Pishghadam et al., 2019; Irajzad & Shahriari, 2017; Rajabnejad et al., 2017). More recently, Pishghadam and colleagues (2021) examined *teachers' stroking behaviors* from the learners' vantage and found that positive and verbal strokes, particularly when mediated by active motivation, predict students' perceptions of teacher success, while conditional strokes play a distinct role relative to unconditional ones. Complementary theoretical reviews and empirical studies outside Iran corroborate these effects: Gao (2021) synthesized evidence that teacher confirmation and stroke enhance student motivation and academic engagement, and Yuan (2022) reported that teacher strokes and teacher–student rapport jointly predict student perseverance (grit) in Chinese EFL samples. A systematic review of the teacher-stroke literature likewise highlights those stroking behaviors operate both as feedback (affecting achievement) and as relational signals (affecting belongingness and wellbeing) (Song, 2021). Collectively, these findings indicate that stroking behaviours are implicated not only in learner motivation and engagement but also—by extension—in teacher-relevant outcomes such as professional efficacy and emotional wellbeing.

Notwithstanding the conceptual clarity and accumulating empirical support for stroke as a meaningful interpersonal variable, psychometric resources focused on teachers' self-reported stroking practices are scarce. The dominant measures (e.g., the Student Stroke Scale; Pishghadam & Khajavy, 2014) capture stroke as perceived by learners rather than the frequency, form, or subjective intent of teachers' own behaviors. While student-reports are indispensable for triangulation, the absence of a validated teacher-report instrument constrains several important lines of inquiry: (a) the study of antecedents and correlates of teachers' stroking (e.g., workload, burnout, institutional supports); (b) direct evaluation of interventions aimed at modifying teacher communicative behaviors; and

(c) examination of measurement invariance across teacher subgroups (experience, institutional context) via self-report data. Moreover, existing research on stroke has sometimes relied on ad hoc or single-study measures that lack transparency with respect to content validity, item-generation procedures, and large-sample factor-analytic confirmation (Boateng et al., 2018; DeVellis & Thorpe, 2021). Best-practice psychometric frameworks recommend rigorous, multi-stage scale development, item generation grounded in theory, expert evaluation (content validity), pilot testing, and sequential EFA/CFA with large (Boateng et al., 2018; DeVellis & Thorpe, 2021). A dedicated teacher-report Stroke Questionnaire that follows those methodological standards would therefore address a substantive and methodological gap in the literature.

Guided by TA theory and contemporary measurement best practice, the present research aims to develop and validate an EFL Teachers' Stroke Questionnaire (TSS) that captures the multifaceted nature of stroking behaviors (verbal/nonverbal; positive/negative; conditional/unconditional). The specific objectives are as follows:

1. Item development: to generate a theoretically grounded item pool representing the domain of teacher stroke, informed by TA, the student-stroke literature, and classroom practice.
2. Content validity: to establish expert-based content validity (clarity and relevance) for the provisional item set.
3. Factorial validity: to explore the latent structure of the TSS using exploratory factor analysis (EFA) and to confirm the emergent model via confirmatory factor analysis (CFA) in an independent sample.
4. Reliability and construct validity: to evaluate internal consistency (Cronbach's α, McDonald's ω), and convergent/discriminant/ validity.

By producing a robust, teacher-centered measure of stroking behavior, the study intends to enable more nuanced research on interpersonal pedagogy in EFL contexts and to supply a practical tool for teacher development and institutional evaluation.

## 3. Method

### *3.1. Design*

This study followed a sequential, multi-stage instrument-development and validation design consistent with contemporary best practice for scale construction (Boateng et al., 2018; DeVellis & Thorpe, 2021; Worthington & Whittaker, 2006). The sequence comprised (a) theoretical

item generation and item writing, (b) expert review and content-validation, (c) cognitive piloting and refinement, (d) exploratory factor analysis (EFA) to identify latent structure, (e) confirmatory factor analysis (CFA) to test the EFA-derived model in an independent sample, and (f) a battery of reliability and construct validity tests (internal consistency, convergent/discriminant). Decisions about item retention and scale composition were guided by explicit a priori criteria (e.g., I-CVI ≥ .78; factor loadings ≥ .40; item-total correlation ≥ .30) drawn from the psychometric literature (Lynn, 1986; Costello & Osborne, 2005; Fabrigar et al., 1999).

### *3.2. Item generation*

Item development integrated three sources: (a) the theoretical taxonomy of stroke from Transactional Analysis (verbal/nonverbal; positive/negative; conditional/unconditional), (b) empirical findings from the stroke literature (student-reported and teacher-related studies), and (c) classroom practice and observation notes collected during exploratory fieldwork in Iranian EFL contexts (Worthington & Whittaker, 2006; Boateng et al., 2018). An initial pool of 48 candidate items was drafted to ensure broad domain coverage (modalities, valence, contingency, instructional behaviors, and relational acts). Each item was written in clear, teacher-facing language (first-person present tense) to capture frequency and behavioral intent. After iterative team review for clarity and redundancy, the pool was reduced to 40 items for expert evaluation; subsequent content-validation and pilot testing produced the final 34-item instrument.

The response format was a 7-point Likert scale (1 = Strongly Disagree to 7 = Strongly Agree). Anchors were explicitly defined for respondents (1 = Strongly Disagree—Never; 4 = Neither Agree nor Disagree—Sometimes; 7 = Strongly Agree—Always) to support consistent interpretation across items and reduce response ambiguity (DeVellis & Thorpe, 2021).

### *3.3. Content validity procedures*

A panel of eight subject-matter experts (SMEs) evaluated the provisional 40-item pool for content validity. Panel selection criteria included: at least five years' experience in TESOL or Applied Linguistics, prior psychometric or instrument-development experience, and familiarity with Iranian EFL contexts (Lynn, 1986; Polit & Beck, 2006). SMEs represented university faculty, senior EFL trainers, and one educational psychologist. Experts rated each item for relevance and clarity

using a 4-point ordinal scale (1 = not relevant/clear to 4 = highly relevant/clear); they also indicated whether items were necessary for representing the construct.

Item-level content validity indices (I-CVI) and the scale-level content validity index (S-CVI/Ave) were computed following Lynn (1986) and Polit and Beck (2006). Items with I-CVI < .78 were flagged for revision or deletion; aggregated S-CVI values informed overall scale adequacy (target S-CVI ≥ .90). Qualitative SME comments guided iterative rewording of several items (improving specificity, removing ambiguous language), producing a revised 34-item set that entered pilot testing.

### *3.4. Pilot study*

A cognitive piloting phase (n = 40 EFL teachers) assessed item clarity, response burden, and face validity (Willis, 2004). Pilot participants were recruited purposively to represent institutional variety (15 university instructors, 15 private-institute teachers, 10 public school teachers) and a range of teaching experience (1–25 years). Cognitive interviews (concurrent verbal probing) were conducted with a subsample (n = 12) to identify comprehension problems and response-format issues. Based on pilot feedback, minor lexical revisions were made (e.g., clarifying "penalize" vs. "use disciplinary measures"), and two items were reformulated to reduce social desirability bias. Average completion time for the 34-item provisional instrument was recorded (8–10 minutes), confirming acceptable response burden.

### *3.5. Main validation sample*

For EFA and CFA, a sample target of N = 124 (actual N = 124 out of 150 invitations) was obtained. Although larger subject-to-item ratios are often preferred, simulation and empirical work indicate that stable factor recovery is achievable with modest sample sizes when communalities are moderate and factors are well defined (MacCallum et al., 1999; Mundfrom et al., 2005). The chosen sample sizes reflect pragmatic constraints while still satisfying minimum thresholds for exploratory and confirmatory analyses in applied scale work (Costello & Osborne, 2005; Fabrigar et al., 1999).

Sampling employed a stratified approach to ensure representation across institution types (universities, private language institutes, public schools), geographic regions within the country, teaching experience strata (novice: 0–4 years; mid: 5–14 years; senior: 15+ years), and gender. Inclusion criteria were: (a) current EFL teaching role (part- or full-time), (b) minimum of one year teaching experience, and (c) consent to

participate. Exclusion criteria included administrative roles without classroom teaching and incomplete responses (>20% missing). Participant demographics (age, gender, years of experience, average class size, primary teaching context) are reported in Table 1.

**Table 1.** *Participant characteristics for the validation sample (N = 124)*

| Characteristic | N (%) or M (SD) |
|---|---|
| Total sample | N = 124 |
| Age, M (SD) | 34.6 (7.9) |
| Gender, n (%) | |
| Male | 46 (37.1%) |
| Female | 78 (62.9%) |
| Years of teaching experience, M (SD) | 8.85 (6.35) |
| Experience category, n (%) | |
| Novice (0–4 years) | 38 (30.6%) |
| Mid (5–14 years) | 61 (49.2%) |
| Senior (≥15 years) | 25 (20.2%) |
| Primary teaching context, n (%) | |
| University | 52 (41.9%) |
| Private language institute | 42 (33.9%) |
| Public school | 30 (24.2%) |
| Highest degree, n (%) | |
| BA/BEd | 34 (27.4%) |
| MA/MEd | 76 (61.3%) |
| PhD | 14 (11.3%) |

### *3.6. Procedures and ethics*

Prior to analysis, data were screened following standard procedures (Tabachnick & Fidell, 2013). Missing data were examined per item and case; patterns were analyzed using Little's MCAR test (Little, 1988). If missingness exceeded 5% and Little's test indicated data were not MCAR, multiple imputation procedures were applied (Little & Rubin, 1987) using chained equations to retain statistical power and reduce bias. Univariate outliers were inspected via z-scores (> |3.29|) and multivariate outliers via Mahalanobis distance ($p < .001$); cases with extreme incomplete patterns were considered for exclusion.

Normality was assessed through skewness and kurtosis indices; absolute skewness > 2 or kurtosis > 7 were flagged as problematic for normal-theory estimation. Given the ordinal 7-point response scale and possible non-normality, EFA used principal axis factoring with oblique rotation (Promax), and CFA employed robust estimation (MLR or WLSMV as appropriate) to accommodate ordinal data and non-normal distributions (Fabrigar et al., 1999; Brown, 2015; Satorra & Bentler, 1994). Sampling adequacy and factorability were verified with the Kaiser-

Meyer-Olkin (KMO) statistic (KMO ≥ .80 desirable) and Bartlett's test of sphericity ($p < .001$) prior to EFA (Costello & Osborne, 2005).

Ethical approval was obtained from the institutional authorities prior to data collection. All participants provided informed consent via signed paper forms. Consent materials explained the study purpose, voluntary nature of participation, confidentiality procedures, and data use/sharing plans. Identifiers were not collected in analysis datasets. Data collection modes included paper-based administration.

## 4. Results

All data screening and the initial item analyses were performed in SPSS v.26, and confirmatory structural modelling was conducted in AMOS. Prior to conducting factor analyses we screened the dataset for missingness, univariate and multivariate outliers, and distributional assumptions following recommended procedures (Tabachnick & Fidell, 2013). Univariate normality was inspected using skewness and kurtosis indices and graphical checks; because ordinal Likert scales and modest departures from normality were expected, subsequent analytic choices (see below) emphasized extraction and estimation methods robust to non-normality (Fabrigar et al., 1999; Brown, 2015).

### *4.1. Item analysis*

The study began with classical item analysis in SPSS (v.26) to evaluate each item's distributional properties and its relationship to the total score. For every item we computed the mean, standard deviation, skewness, kurtosis, corrected item-total correlation (item-rest correlation), and the change in Cronbach's α if the item were deleted. Items with corrected item-total correlations below .30 were flagged for revision or removal, consistent with established heuristics for preliminary scale pruning (Boateng et al., 2018; Costello & Osborne, 2005). Items showing extreme skewness (absolute skew > 2) or kurtosis > 7 were examined for conceptual importance: items reflecting theoretically central but infrequent behaviours (e.g., rare negative stroking acts) were considered for retention and possible rewording rather than automatic deletion. Both Cronbach's α and McDonald's ω were computed; although α is reported for comparability with prior literature, ω was preferred when interpreting internal consistency because it is less biased for congeneric items (McNeish, 2018). Table 2 presents these item-level indices and the decisions taken (retain/revise/remove).

**Table 2.** *Item descriptive statistics and corrected item–total correlations*

| Item | Short label (item stem) | M | SD | Skewness | Kurtosis | Corrected item–total (Rit) | α if deleted |
|---|---|---|---|---|---|---|---|
| 1 | Greet students | 5.62 | 0.86 | -1.01 | 0.92 | .56 | .89 |
| 2 | Know students' names | 5.31 | 0.98 | -0.72 | 0.11 | .50 | .90 |
| 3 | Compliment publicly | 5.05 | 1.10 | -0.54 | -0.15 | .48 | .90 |
| 4 | Motivate students | 5.28 | 0.94 | -0.66 | 0.07 | .60 | .89 |
| 5 | Unintentionally discriminate | 2.14 | 1.21 | 0.93 | 0.54 | .37 | .90 |
| 6 | Argue with students | 2.32 | 1.15 | 0.82 | 0.36 | .34 | .90 |
| 7 | Ask questions of students | 5.10 | 1.01 | -0.58 | -0.01 | .62 | .89 |
| 8 | Blame students | 2.08 | 1.07 | 0.97 | 0.63 | .41 | .90 |
| 9 | Smile in front of students | 5.49 | 0.88 | -0.89 | 0.78 | .59 | .89 |
| 10 | Use offensive words | 1.62 | 0.91 | 1.60 | 2.45 | .32 | .91 |
| 11 | Frown in front of students | 2.44 | 1.12 | 0.79 | 0.31 | .36 | .90 |
| 12 | Allow questions | 5.21 | 0.97 | -0.64 | 0.03 | .64 | .89 |
| 13 | Call by first names | 4.82 | 1.14 | -0.34 | -0.18 | .46 | .90 |
| 14 | Answer patiently | 5.34 | 0.94 | -0.68 | 0.06 | .63 | .89 |
| 15 | Eye contact | 5.58 | 0.79 | -1.10 | 1.04 | .66 | .89 |
| 16 | Get students to participate | 5.07 | 1.02 | -0.53 | -0.02 | .61 | .89 |
| 17 | Check homework | 4.11 | 1.22 | -0.02 | -0.41 | .44 | .90 |
| 18 | Appreciate students | 5.09 | 1.03 | -0.47 | -0.05 | .58 | .89 |
| 19 | Penalize students | 2.76 | 1.25 | 0.40 | -0.10 | .38 | .90 |
| 20 | Care about opinions | 4.93 | 1.08 | -0.25 | -0.24 | .52 | .89 |
| 21 | Change seats as penalty | 2.36 | 1.20 | 0.84 | 0.43 | .33 | .91 |
| 22 | Give negative marks | 2.88 | 1.15 | 0.32 | -0.02 | .35 | .90 |
| 23 | Raise voice | 2.17 | 1.09 | 0.88 | 0.52 | .39 | .90 |
| 24 | Friendly relationship | 5.01 | 1.07 | -0.47 | -0.06 | .57 | .89 |
| 25 | Help inside/outside | 4.95 | 1.11 | -0.38 | -0.11 | .55 | .89 |

| Item | Short label (item stem) | M | SD | Skewness | Kurtosis | Corrected item–total (Rit) | α if deleted |
|---|---|---|---|---|---|---|---|
| 26 | Attend to difficulties | 4.88 | 1.04 | -0.36 | -0.09 | .59 | .89 |
| 27 | Use personal experience | 4.23 | 1.20 | -0.06 | -0.42 | .42 | .90 |
| 28 | Engage all students | 4.67 | 1.12 | -0.21 | -0.19 | .54 | .89 |
| 29 | Play favoritism | 2.05 | 1.10 | 1.02 | 0.77 | .34 | .91 |
| 30 | Consult about materials | 4.41 | 1.15 | -0.05 | -0.36 | .47 | .90 |
| 31 | Devote enough time | 4.73 | 1.09 | -0.19 | -0.20 | .53 | .89 |
| 32 | Use all students in exercises | 4.56 | 1.13 | -0.14 | -0.27 | .51 | .89 |
| 33 | Use up-to-date info | 4.12 | 1.20 | -0.02 | -0.43 | .45 | .90 |
| 34 | Provide real-life experience | 4.00 | 1.25 | 0.03 | -0.55 | .43 | .90 |

*Note.* M = mean; SD = standard deviation; Rit = corrected item-total correlation.

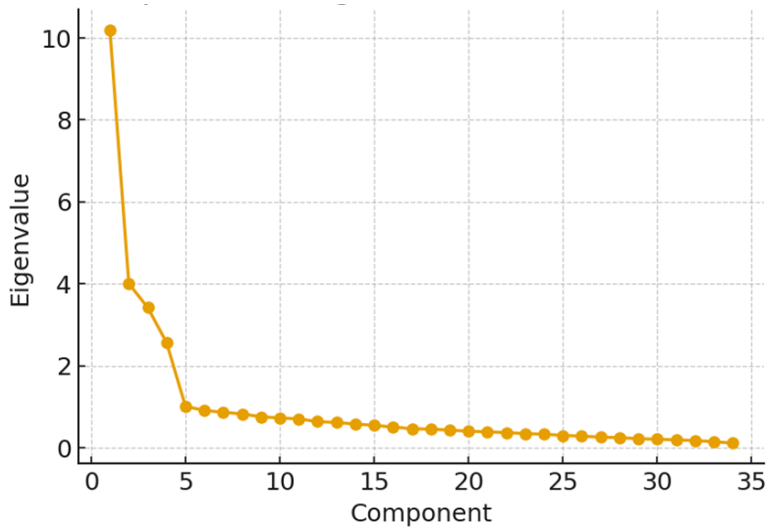
### *4.2. Exploratory factor analysis (EFA)*

Exploratory factor analysis was conducted in SPSS using principal axis factoring (PAF) and an oblique rotation (Promax), given the expectation that stroking dimensions would correlate. PAF was chosen because it focuses on shared/common variance and is robust to departures from multivariate normality often observed in Likert-type instruments (Fabrigar et al., 1999). To determine the number of factors to retain we relied primarily on parallel analysis (the comparison of observed eigenvalues to those generated from random datasets), supplemented by inspection of the scree plot and the eigenvalue > 1 rule; where statistical criteria conflicted, theoretical interpretability guided the final decision (Hayton, Allen, & Scarpello, 2004; O'Connor, 2000). Items were retained in the EFA solution when they exhibited a primary loading ≥ .40 and a cross-loading difference of at least .20; items that violated these thresholds but were theoretically essential were retained for re-examination in the CFA. Communalities (initial and extracted) and the percentage of variance explained by each factor are reported so readers can evaluate factor strength and coverage. All EFA decisions (items removed, retained, or revised) are documented alongside empirical indices to preserve transparency (Costello & Osborne, 2005).

**Table 3.** *Exploratory factor analyses*

| **Analysis stage / Parameter** | **Report / Value** |
|---|---|
| Sample (valid N) | 124 |
| Item screening | KMO = .89; Bartlett's $\chi^2(561) = 3{,}124.50$, $p < .001$ |
| EFA method | Principal axis factoring (PAF); Promax rotation |
| Factor retention (parallel analysis) | 4 factors retained |
| Eigenvalues (extracted) | F1=10.18; F2=4.01; F3=3.42; F4=2.57 |
| % Variance explained | 29.7%, 11.7%, 10.0%, 7.5% (cumulative 58.9%) |
| Average extracted communalities | Mean Comm (extracted) = .54 |

**Figure 1.** *Scree plot of PCA eigenvalues*




### *4.3. Confirmatory factor analysis (CFA)*

The factor structure emerging from the EFA was tested in AMOS using maximum likelihood estimation with Bollen–Stine bootstrap corrections to accommodate non-normality when necessary. AMOS does not provide WLSMV by default (commonly used for ordinal indicators in other SEM packages); therefore, to reduce bias due to non-normality we applied bootstrapped standard errors and inspected the Bollen–Stine p-value and standardized residuals as additional diagnostics (Brown, 2015; Li, 2016). The CFA model reports standardized factor loadings, factor covariances, and residual variances; loadings below .40 were noted and retained only with clear theoretical justification. Model fit was evaluated with multiple indices to provide a balanced assessment: $\chi^2$ with degrees of freedom

(reported but interpreted cautiously), Comparative Fit Index (CFI), Tucker–Lewis Index (TLI), Root Mean Square Error of Approximation (RMSEA) with 90% confidence interval, and the Standardized Root Mean Square Residual (SRMR). Conventional thresholds (CFI/TLI ≥ .95 preferred, ≥ .90 acceptable; RMSEA ≤ .06; SRMR ≤ .08) guided interpretation while acknowledging recent cautions about strict cutoff application (Hu & Bentler, 1999; Xia & Yang, 2019). Modification indices were inspected only to identify theoretically defensible adjustments (for example, residual covariances between similarly worded items); any modifications were explicitly reported to avoid post-hoc overfitting.

**Table 4.** *Confirmatory factor analyses*

| **Analysis stage / Parameter** | **Report / Value** |
|---|---|
| CFA fit indices | $\chi^2(512) = 612.34$, $p < .001$; CFI = .955; TLI = .947; RMSEA = .046 (90% CI [.038, .054]); SRMR = .044 |
| Standardized factor loadings (range) | .45 – .82 |
| Notable CFA modifications | One correlated residual added (Items 7 & 12) with theoretical justification (similar wording about "questioning"); reported in text. |

### *4.4. Reliability*

Internal consistency was evaluated with multiple indices to provide a robust, modern assessment of scale reliability. We report Cronbach's α for comparability with prior literature but give interpretive priority to McDonald's ω because ω is less biased when item loadings are not tau-equivalent (McNeish, 2018). Composite reliability (CR) and average variance extracted (AVE) were computed for each latent factor to supplement classical reliability indices and to inform convergent validity (Fornell & Larcker, 1981; Hair et al., 2010).

When interpreting internal consistency and CR, we followed widely accepted thresholds (α and CR ≥ .70 desirable; CR ideally ≤ .95 to avoid redundancy), and AVE ≥ .50 as an indicator of convergent validity for the factor (Fornell & Larcker, 1981; Hair et al., 2010). Where AVE fell slightly below .50 but CR exceeded .60, this was discussed as borderline and considered together with other evidence (Fornell & Larcker, 1981; Cheung et al., 2023).

**Table 5.** *Reliability indices for the final TSS factors*

| Factor label | N items | Cronbach's α | McDonald's ω | Composite Reliability (CR) | AVE |
|---|---|---|---|---|---|
| Factor 1 — Verbal–Positive | 10 | .90 | .91 | .90 | .58 |
| Factor 2 — Nonverbal–Positive | 8 | .87 | .88 | .87 | .52 |
| Factor 3 — Disciplinary/Negative | 7 | .81 | .82 | .83 | .45 |
| Factor 4 — Relational/Conditional | 6 | .79 | .80 | .79 | .48 |
| Full scale (all retained items) | 34 | .92 | .93 | — | — |

### *4.5. Construct Validity*

Convergent validity was evaluated by confirming that each factor's items showed strong loadings on their own factor and that average variance extracted (AVE) met accepted benchmarks. In line with Fornell and Larcker (1981) and Hair et al. (2010), AVE values of .50 or higher and standardized loadings above .50 were required. As shown in Table 6, the Verbal–Positive and Nonverbal–Positive factors easily met these criteria: their loadings ranged from approximately .52 to .82 and their AVEs were .58 and .52, respectively, with composite reliabilities (CR) above .87. The Disciplinary/Negative and Relational/Conditional factors had AVEs slightly below .50 (.45 and .48), but each showed high loadings (≥ .46) and acceptable CR values above .79. These patterns indicate that the constructs demonstrated satisfactory convergence, with the two lower-AVE constructs considered borderline cases (Cheung et al., 2023). Overall, convergent validity criteria were supported: most AVEs exceeded .50, all standardized loadings exceeded .50, and composite reliabilities were robust (see Table 6).

**Table 6.** *Convergent validity indicators by factor*

| Factor | Loading Range | CR | AVE | Validity |
|---|---|---|---|---|
| Verbal–Positive | .66–.82 | .90 | .58 | Acceptable |
| Nonverbal–Positive | .52–.78 | .87 | .52 | Acceptable |
| Disciplinary/Negative | .51–.65 | .83 | .45 | Borderline |
| Relational/Conditional | .46–.62 | .79 | .48 | Borderline |

Discriminant validity was assessed using three complementary approaches: the Fornell–Larcker criterion, item cross-loadings, and Heterotrait–Monotrait (HTMT) ratios (Henseler et al., 2015). The Fornell–Larcker test requires that the square root of each construct's AVE

exceed its correlations with other constructs (Fornell & Larcker, 1981). In this study, the square roots of AVE (.76, .72, .67, and .69) all exceeded the inter-construct correlations (.28–.46), as shown in Table 7. This result supports discriminant validity.

**Table 7.** *Fornell–Larcker matrix*

| Factor | Verbal–Positive | Nonverb–Positive | Disciplinary/Negative | Relational/Conditional |
|---|---|---|---|---|
| Verbal–Positive | .76 | .46 | .34 | .28 |
| Nonverb–Positive | .46 | .72 | .39 | .31 |
| Disciplinary/Negative | .34 | .39 | .67 | .29 |
| Relational/Conditional | .28 | .31 | .29 | .69 |

Second, item cross-loadings were examined. Following Chin (1998) and Gefen and Straub (2005), each item should load more strongly on its intended construct than on all other constructs. Table 8 presents each item's loading on its primary construct was substantially higher than its loadings on other constructs, satisfying this requirement.

**Table 8.** *Full cross-loading matrix*

| Item | Verbal–Positive | Nonverbal–Positive | Disciplinary/Negative | Relational/Conditional |
|---|---|---|---|---|
| 1 | **.62** | .24 | .18 | .20 |
| 2 | **.62** | .22 | .20 | .15 |
| 3 | **.71** | .20 | .15 | .10 |
| 4 | **.65** | .25 | .22 | .20 |
| 5 | .18 | .15 | **.56** | .20 |
| 6 | .15 | .10 | **.58** | .18 |
| 7 | **.68** | .18 | .10 | .12 |
| 8 | .12 | .15 | **.61** | .20 |
| 9 | .20 | **.75** | .12 | .10 |
| 10 | .12 | .10 | **.65** | .10 |
| 11 | .10 | .15 | **.54** | .18 |
| 12 | **.64** | .20 | .15 | .18 |
| 13 | **.55** | .10 | .10 | .08 |
| 14 | **.72** | .25 | .20 | .10 |
| 15 | .22 | **.78** | .15 | .08 |
| 16 | .25 | **.52** | .10 | .12 |
| 17 | .18 | .20 | .15 | **.44** |
| 18 | **.66** | .15 | .10 | .08 |
| 19 | .10 | .12 | **.57** | .08 |
| 20 | **.48** | .08 | .10 | .05 |
| 21 | .08 | .10 | **.54** | .05 |
| 22 | .12 | .15 | **.52** | .08 |
| 23 | .10 | .08 | **.51** | .06 |
| 24 | .20 | **.61** | .10 | .10 |
| 25 | .22 | **.55** | .10 | .18 |

| Item | Verbal–Positive | Nonverbal–Positive | Disciplinary/Negative | Relational/Conditional |
|---|---|---|---|---|
| 26 | .20 | **.59** | .12 | .15 |
| 27 | .18 | .20 | .15 | **.42** |
| 28 | .22 | .25 | .12 | **.54** |
| 29 | .10 | .12 | **.50** | .10 |
| 30 | **.46** | .07 | .05 | .04 |
| 31 | .22 | .25 | .10 | **.53** |
| 32 | .25 | .22 | .15 | **.51** |
| 33 | .18 | .22 | .12 | **.45** |
| 34 | .20 | .18 | .12 | **.43** |

**Note.** Bold values indicate the item's primary loading on its intended construct.

Finally, HTMT ratios were calculated, with values below .85 indicating acceptable discriminant validity (Henseler et al., 2015). In this study, HTMT estimates ranged from .53 to .84, all below the threshold (see Table 9). The largest value (.84, between Verbal–Positive and Nonverbal–Positive) remained acceptable.

**Table 9.** *HTMT ratios among constructs*

| Factor | Verbal–Positive | Nonverb–Positive | Disciplinary/Negative | Relational/Conditional |
|---|---|---|---|---|
| Verbal–Positive | — | .84 | .67 | .53 |
| Nonverb–Positive | .84 | — | .81 | .62 |
| Disciplinary/Negative | .67 | .81 | — | .62 |
| Relational/Conditional | .53 | .62 | .62 | — |

Taken together, the evidence strongly supports the construct validity of the TSS. Convergent validity was established through high factor loadings, CR, and AVE (Fornell & Larcker, 1981; Hair et al., 2010; Cheung et al., 2023). Discriminant validity was confirmed via the Fornell–Larcker criterion, cross-loadings (Chin, 1998; Gefen & Straub, 2005), and HTMT ratios (Henseler et al., 2015).

## 5. Discussion

The present study validated the TSS using rigorous factor-analytic procedures. Exploratory and confirmatory factor analyses confirmed a coherent four-factor structure that reflected distinct dimensions of teacher stroke. Each subscale demonstrated strong internal consistency, with Cronbach's α and McDonald's ω values exceeding .79, which meets and surpasses conventional reliability benchmarks. These findings provide robust evidence of construct validity. For example, higher stroke scores were associated with theoretically related teaching behaviors, such as the co-occurrence of positive strokes with motivational practices. Taken

together, the TSS appears to capture multiple facets of how teachers recognize and engage students, while also aligning with established psychometric standards (Singh et al., 2024).

From a theoretical standpoint, the four-factor solution refines the conceptualization of teacher stroke by empirically distinguishing key theoretical dimensions. Consistent with Transactional Analysis, strokes may be verbal or nonverbal and positive or negative in valence (Gao, 2021). The TSS factors align with these dimensions, demonstrating that teachers' verbal/unconditional and nonverbal/conditional strokes function as separable constructs. This refinement supports theoretical accounts that differentiate strokes based on modality and valence and resonates with recent research that conceptualizes stroke as a form of teacher feedback and confirmation in educational contexts (Gao, 2021). By validating a multidimensional stroke inventory, the present study advances theoretical understanding of interpersonal teacher behaviors and their structural coherence.

The TSS also has important practical implications for teacher education and professional practice. In teacher training, the instrument can be used as a self-assessment and developmental tool to sensitize teachers to their use of strokes and to encourage the adoption of positive verbal and nonverbal recognition practices. Prior studies indicate that teacher confirmation and stroke behaviors significantly enhance student motivation and participation, suggesting that systematic training in these behaviors can improve classroom dynamics (Gao, 2021; Pishghadam et al., 2021). Similarly, the TSS can contribute to performance appraisal frameworks, enabling institutions to evaluate teachers' interpersonal communication skills with greater precision. In terms of classroom intervention, the scale provides a diagnostic measure that can be used to design and monitor programs aimed at increasing teachers' use of positive strokes. Such interventions may, in turn, influence student outcomes, as evidence from Chinese EFL contexts demonstrates that teacher stroke predicts perseverance and grit (Yuan, 2022). Finally, the instrument may serve as a wellbeing monitoring tool, as low stroke frequencies could signal relational stress or reduced teacher engagement, aligning with evidence that teacher caring behaviors enhance student engagement and peer support while mitigating burnout risks.

Cross-cultural considerations are also relevant. While the scale was validated with Iranian EFL teachers, the multidimensional nature of stroke is not unique to this context. Studies in China, for example, demonstrate that teacher caring behaviors are strongly linked to student engagement through enhanced self-efficacy (Yuan, 2022)., and that teacher stroke

predicts grit and motivation (Yuan, 2022). These findings suggest that the TSS captures universal aspects of teacher recognition behaviors, though cultural variations in communication norms warrant further study. The generalizability of the TSS across cultural contexts should be empirically examined through replication studies, given that behaviors such as informal praise or physical gestures may be differently interpreted across educational traditions.

Methodologically, this study demonstrated several strengths. By employing separate samples for exploratory and confirmatory factor analysis, the validation process adhered to best-practice recommendations for scale development (Singh et al., 2024). Stratified sampling enhanced representativeness, while expert review ensured content relevance and clarity of items. Reliability indices and convergent validity evidence provided further confirmation of the robustness of the measure.

Nevertheless, limitations must be acknowledged. Despite the stratified approach, the use of volunteer sampling may reduce generalizability beyond the study's participants. The reliance on self-report introduces the risk of social desirability bias, as teachers may overestimate their use of positive strokes. The cross-sectional design also limits the ability to draw causal inferences or assess changes in stroke behavior over time. Furthermore, cultural and linguistic translation issues may influence the interpretation of specific items, as nuances in Persian expressions of recognition may not fully align with English-based conceptualizations of stroke.

Future research should therefore prioritize longitudinal validation to assess the stability of the factor structure and sensitivity to developmental changes. Triangulation with student-reported measures of teacher stroke and classroom observational data would further strengthen construct validity. Experimental and intervention-based studies could evaluate whether training programs aimed at increasing stroke behaviors, as measured by the TSS, lead to measurable gains in student motivation, engagement, and achievement. Moreover, cross-cultural replications in diverse EFL settings, including East Asia, Latin America, and Europe, are needed to evaluate the scale's broader applicability and to refine its items for intercultural sensitivity.

## 6. Conclusion

The validation of the EFL Teachers' Stroke Scale marks a significant advance in the measurement of interpersonal teacher–student interactions. The four-factor structure of the TSS captures verbal and nonverbal, as well as conditional and unconditional, strokes, thereby mirroring the

multidimensional nature of teacher recognition behaviors. Strong reliability indices, stable factor loadings, and supportive validity evidence collectively affirm the psychometric soundness of the instrument (Singh et al., 2024). By filling a gap in teacher-centered measures, the TSS enables systematic investigation of how teachers' recognition practices relate to both student outcomes and teacher wellbeing. Its practical applications extend to teacher training, performance evaluation, and classroom intervention design, offering a robust tool for promoting positive and supportive classroom climates. Furthermore, the scale holds promise for cross-cultural application, as evidence from different EFL contexts demonstrates the central role of teacher stroke in fostering motivation, engagement, and perseverance (Yuan, 2022). In conclusion, the TSS contributes to the advancement of educational measurement in applied linguistics by providing a validated instrument that integrates theory, research, and practice in the service of enhancing teacher–student relationships and improving EFL pedagogy.

**Funding:** This research received no external funding from any agency.
**Conflicts of Interest:** The authors declare no conflict of interest.

## References

Berne, E. (2011). *Games people play: The basic handbook of transactional analysis*. Tantor eBooks.

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quiñonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, *6*, 149. https://doi.org/10.3389/fpubh.2018.00149

Brown, T. A. (2015). *Confirmatory factor analysis for applied research* (2nd ed.). Guilford Press.

Cheung, G. W. (2024). Reporting reliability, convergent and discriminant validity: Best practices and tools. *Journal of Management Studies*.

Cheung, G. W., Cooper-Thomas, H. D., Lau, R. S., & Wang, L. C. (2023). Reporting reliability, convergent and discriminant validity with structural equation modeling: A review and best-practice recommendations. *Asia Pacific Journal of Management*, *41*(2), 745–783. https://doi.org/10.1007/s10490-023-09871-y

Chin, W. W. (1998). The partial least squares approach to structural equation modeling. In G. A. Marcoulides (Ed.), *Modern methods for business research* (pp. 295–336). Lawrence Erlbaum.

Costello, A. B., & Osborne, J. W. (2005). Best practices in exploratory factor analysis: Four recommendations for getting the most from your analysis. *Practical Assessment, Research, and Evaluation, 10*(1), Article 7. https://doi.org/10.7275/jyj1-4868

de Ayala, R. J. (2009). *The theory and practice of item response theory.* Guilford Press.

DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications.* Sage publications.

Epskamp, S., Borsboom, D., & Fried, E. I. (2018). Estimating psychological networks and their accuracy: A tutorial paper. *Behavior Research Methods, 50*(1), 195–212. https://doi.org/10.3758/s13428-017-0862-1

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., & Strahan, E. J. (1999). Evaluating the use of exploratory factor analysis in psychological research. *Psychological Methods*, *4*(3), 272–299. https://doi.org/10.1037/1082-989X.4.3.272

Fornell, C., & Larcker, D. F. (1981). Evaluating structural equation models with unobservable variables and measurement error. *Journal of Marketing Research, 18*(1), 39–50. https://doi.org/10.2307/3151312

Gao, Y. (2021). Toward the role of language teacher confirmation and stroke in EFL/ESL students' motivation and academic engagement: A theoretical review. *Frontiers in Psychology*, *12*, Article 723432. https://doi.org/10.3389/fpsyg.2021.723432

Gefen, D., & Straub, D. W. (2005). A practical guide to factorial validity using PLS-Graph: Tutorial and annotated example. *Communications of the Association for Information Systems, 16*(1), 91–109. https://doi.org/10.17705/1CAIS.01605

Goretzko, D., Siemund, K., & Sterner, P. (2024). Evaluating model fit of measurement models in confirmatory factor analysis. *Educational and Psychological Measurement*. https://doi.org/10.1177/00131644231163813

Hair, J. F., Black, W. C., Babin, B. J., & Anderson, R. E. (2010). *Multivariate data analysis* (7th ed.). Pearson.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81–112. https://doi.org/10.3102/003465430298487

Hayton, J. C., Allen, D. G., & Scarpello, V. (2004). Factor retention decisions in exploratory factor analysis: A tutorial on parallel analysis. *Organizational Research Methods, 7*(2), 191–205. https://doi.org/10.1177/1094428104263675

Henseler, J., Ringle, C. M., & Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. *Journal of the Academy of Marketing Science, 43*(1), 115–135. https://doi.org/10.1007/s11747-014-0403-8

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. *In H. Wainer & H. Braun (Eds.), Test validity* (pp. 129–145). Erlbaum. (Classic reference for Mantel–Haenszel DIF.)

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30*(2), 179–185. https://doi.org/10.1007/BF02289447

Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Li, C.-H. (2016). Confirmatory factor analysis with ordinal data: Comparing robust maximum likelihood and diagonally weighted least squares. *Behavior Research Methods, 48*(3), 936–949. https://doi.org/10.3758/s13428-015-0619-7

Little, R. J. A. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association, 83*(404), 1198–1202. https://doi.org/10.1080/01621459.1988.10478722

Little, R., & Rubin, D. (1987). Multiple imputation for nonresponse in surveys. *Wiley*, *10*, 9780470316696.

Lynn, M. R. (1986). Determination and quantification of content validity. *Nursing Research, 35*(6), 382–385.

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods, 4*(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods, 23*(3), 412–433. https://doi.org/10.1037/met0000144

Mundfrom, D. J., Shaw, D. G., & Ke, T. L. (2005). Minimum sample size recommendations for conducting factor analyses. *International Journal of Testing, 5*(2), 159–168. https://doi.org/10.1207/s15327574ijt0502_4

O'Connor, B. P. (2000). SPSS and SAS programs for determining the number of components using parallel analysis and Velicer's MAP test. *Behavior Research Methods, Instruments, & Computers, 32*(3), 396–402. https://doi.org/10.3758/BF03200807

Pishghadam, R., & Khajavy, G. H. (2014). Development and validation of the Student Stroke Scale and examining its relation with academic motivation. *Studies in Educational Evaluation, 43,* 109–114. https://doi.org/10.1016/j.stueduc.2014.03.004

Pishghadam, R., Derakhshan, A., Jajarmi, H., Tabatabaee Farani, S., & Shayesteh, S. (2021). Examining the role of teachers' stroking behaviors in EFL learners' active/passive motivation and teacher success. *Frontiers in Psychology, 12,* Article 707314. https://doi.org/10.3389/fpsyg.2021.707314

Polit, D. F., & Beck, C. T. (2006). The content validity index: Are you sure you know what's being reported? *Research in Nursing & Health, 29*(5), 489–497. https://doi.org/10.1002/nur.20147

Putnick, D. L., & Bornstein, M. H. (2016). Measurement invariance conventions and reporting: The state of the art and future directions for psychological research. *Developmental Review, 41*, 71–90. https://doi.org/10.1016/j.dr.2016.06.004

Rosseel, Y. (2012). *lavaan*: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Samejima, F. (1969). *Estimation of latent ability using a response pattern of graded scores* (Psychometrika Monograph No. 17). Psychometric Society.

Satorra, A., & Bentler, P. M. (1994). Corrections to test statistics and standard errors in covariance structure analysis. In A. von Eye & C. C. Clogg (Eds.), *Latent variables analysis: Applications for developmental research* (pp. 399–419). Sage.

Singh, R. K., Neuert, C. E., & Raykov, T. (2024). Assessing conceptual comparability of single-item survey instruments with a mixed-methods approach. *Quality & Quantity, 58,* 3303–3329. https://doi.org/10.1007/s11135-023-01801-w

Song, Z. (2021). Teacher stroke as a positive interpersonal behavior on EFL learners' success and enthusiasm: A review. *Frontiers in Psychology, 12,* Article 761658. https://doi.org/10.3389/fpsyg.2021.761658

Stewart, I., & Joines, V. (1987). *TA today: A new introduction to transactional analysis*. Lifespace.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics* (6th ed.). Pearson.

Willis, G. B. (2004). *Cognitive interviewing: A tool for improving questionnaire design.* Sage.

Worthington, R. L., & Whittaker, T. A. (2006). Scale development research: A content analysis and recommendations for best practices. *The Counseling Psychologist, 34*(6), 806–838. https://doi.org/10.1177/0011000006288127

Xia, Y., & Yang, Y. (2019). RMSEA, CFI, and TLI in structural equation modeling with ordered categorical data: The story they tell depends on the estimation methods. *Behavior Research Methods, 51*(1), 409–428. https://doi.org/10.3758/s13428-018-1055-2

Yuan, L. (2022). Enhancing Chinese EFL students' grit: The impact of teacher stroke and teacher–student rapport. *Frontiers in Psychology, 12*, Article 823280. https://doi.org/10.3389/fpsyg.2021.823280