

Two-Consonant Cluster Errors and the Perception-Production Relationship among Vietnamese Non-English-Major Freshmen

Authors Details

Truong Thi My Linh^{1*}

Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam.

Corresponding Author: *Truong Thi My Linh*

ABSTRACT: English two-consonant (CC) clusters play an important role in pronunciation, yet remain persistently arduous for EFL learners because of the structural mismatch between English and Vietnamese syllable structure. Under establishing models of Contrastive Analysis Hypothesis, the Markedness Differential Hypothesis and the Speech Learning Model, this explanatory sequential mixed-methods study investigated CC cluster errors and the perception-production relationship among first-year non-English majors at Nguyen Tat Thanh University. First, 110 participants administered a 24-item identification test covering 12 cluster contrasts distributed initially and finally. A volunteer sample of 11 freshmen, subsequently had controlled-reading tasks in isolated words and in sentences, totalling 528 tokens. These recordings were then evaluated for intelligibility on a four-rating scale by two native English-speaking raters and classified into error categories. Praat was also used to verify 60 of the 120 disagreement tokens, while the remaining 60 tokens were excluded, yet it did not change the error distribution by sensitivity analysis. 117 of the 468 classified tokens had errors, mostly concentrating in coda clusters, with deletion being the dominant type (67.5%), followed by substitution (22.2%) and epenthesis (5.1%). The results also show the positive correlation between perception and intelligibility at the learner level ($r=.81$), although this relationship was sensitive to one learner ($\rho=.55$). However, there was no significant contrast-level link, and

several contrasts exhibited perception-production mismatches. The findings are theoretically predominant to support the predictions of the CAH and MDH and the co-evolution view of SLM-r. Overall, the findings reinforce the significance of cluster-focused instruction in pronunciation teaching, prioritizing coda clusters and combining perception training with explicit production practice in L2 speech learning models.

KEYWORDS: *Consonant clusters, Intelligibility, Perception-production relationship, Phonological errors, Syllable position, Vietnamese EFL learners*

INTRODUCTION

English is widely used as a lingua franca, and its non-native speakers now outnumber its native speakers [1, 2]. For learners of English as a foreign language (EFL), pronunciation therefore matters less as a means of sounding native than as a condition for being understood [3, 4]. Among segmental features, consonant clusters pose a particular challenge for learners whose first language (L1) has simple syllable structures, because errors in clusters change the syllable structure of a word and can reduce intelligibility [3].

Vietnamese is a clear case. It is a largely monosyllabic, tonal language that allows no consonant clusters in any position and permits only a single coda consonant, limited to unreleased stops, nasals and glides [5, 6]. English, by contrast, permits up to three consonants in the onset and four in the coda, and its final stops are often released [7]. Research on Vietnamese learners has consistently reported deletion as the dominant repair, especially in coda clusters, together with substitution of segments absent from Vietnamese [8, 9, 10, 11, 12]. Studies of intelligibility have also suggested that final consonant errors are more damaging for Vietnamese speakers than for speakers of other L1s [3].

English consonant clusters and Vietnamese syllable structure

A consonant cluster is generally described as a sequence of two or more consonants within the same syllable with no intervening vowel [7, 13]. Such tautosyllabic clusters stand in the onset (e.g., /pl/ in *play*) or the coda (e.g., /sk/ in *ask*). Word-medial sequences such as /mfl/ in *pamphlet* usually span a syllable boundary and are better grouped in heterosyllabic sequences. English two-consonant onsets include /s/ + consonant and stop or fricative + approximant combinations, and two-consonant codas include pre-final + final sequences such as /-mp, -nt, -lp, -sk/ [7].

The Vietnamese syllable consists of an optional onset, an optional medial glide, a nucleus, an optional coda and a tone [14, 15]. Several apparent consonant correspondences between the two languages differ phonetically: Vietnamese *b* and *đ* are implosive stops, /p/ occurs initially only in loanwords, and Vietnamese contrasts unaspirated /t/ with aspirated /t^h/, whereas English [t] and [t^h] are allophones of one phoneme [6]. Vietnamese has no /θ, ð, ʃ, ʒ, tʃ, dʒ/ and no /g/, which is replaced by the fricative /ɣ/. Coda presents the largest difference, which in Vietnamese is restricted to unreleased [p̚ t̚ k̚], the nasals /m, n, ŋ/ and the glides /w, j/ [5, 6, 16]. English codas such as /-sp/, /-lp/ or /-lθ/ therefore violate two Vietnamese constraints at once: they contain more than one consonant, and they include consonants, such as /s/ and /l/, that never close a Vietnamese syllable.

Repair strategies and intelligibility

International studies show how strongly the repair depends on the L1. Saudi and Tunisian Arabic speakers insert a vowel to break up initial clusters (e.g., *split* → [səplɪt]) because most Arabic onsets contain a single consonant [17, 18]. Spanish speakers add a prothetic vowel before /s/-initial clusters (e.g., *speak* → [ɛspik]) [19], and Japanese speakers insert vowels to conform to the CV syllable (e.g., *dream* → [dorimu]) [20]. Several studies also report that coda clusters are more difficult than onset clusters, in line with markedness [21, 22]. Most of these studies, however, examined either onset or coda clusters, and few compared the same cluster in both positions.

L2 segmental errors are commonly classified as substitutions, distortions, insertions and deletions [4]. For clusters, learners may also simplify the structure by deleting a member or inserting a vowel, and several repairs may combine in a single word [3]. The preferred repair depends on the L1. Speakers of Arabic, Japanese, Korean and Spanish tend to insert vowels to create syllables that are legal in their L1 [17, 18, 19, 20]. Vietnamese speakers, whose language freely allows closed syllables, more often delete consonants in final clusters [3, 23, 24]. Vietnamese studies report deletion as the most frequent strategy for final clusters [8], accounting for up to 85.2% of errors [9] and 87% of coda errors [10], whereas substitution or feature change is more frequent in onset clusters [11] and epenthesis is reported far less often [12].

Not all errors affect intelligibility equally. Intelligibility, the extent to which a listener actually understands an utterance, is partly independent of accentedness [25]. Substitutions

that neutralize contrasts with a high functional load are more likely to cause misunderstanding [26], and changes to syllable structure are likely to affect intelligibility, particularly for Vietnamese speakers, whose final consonant errors have been associated with reduced intelligibility [3].

How intelligibility is measured also matters. Transcription tasks, in which listeners write down what they hear, provide the strictest measure, whereas rating scales are more practical for large numbers of tokens and allow raters to comment on the source of difficulty [25, 27]. Because raters who know the target word can often recover it from a degraded signal, ratings of this kind are best interpreted as rated intelligibility of the target rather than as intelligibility to naive listeners.

Previous studies of Vietnamese learners have summarized these patterns in Table 1, but differ widely in their targets and participants. Most examined English majors or small groups, mixed two-, three- and four-consonant clusters, and focused on either onset or coda position. None held the cluster constant across positions, and none examined perception alongside production. These differences make it difficult to separate the effect of syllable position from that of cluster length or segment identity, which the present design addresses through Set A.

Table 1. Previous studies of English consonant cluster errors by Vietnamese learners

Study	Targets	Main findings
Ha [28]	Onset /tr-/; codas in exam monologues	Substitution of /tr/ by /tʃ/; omission of final consonants; some insertion
Nguyen [8]	Two-member coda clusters	Deletion (mostly of the second consonant) most frequent; devoicing of voiced codas; epenthesis least used
Tran [9]	Onset and coda (CC, CCC)	Omission 85.2% of errors; substitution 14.8%
Tran & Nguyen [11]	Onset (CC, CCC)	CC: feature change 77.4%, deletion 22.2%; CCC: deletion 78.2%; epenthesis rare
Nguyen [10]	Coda (CC–CCCC)	Reduction 87% of errors (e.g., <i>help</i> → “hep”, <i>health</i> → “hel”); substitution 13%
Nguyen [29]	Onset and coda (CC–CCCC)	Accuracy fell with cluster length; deletion and feature change most frequent
Phi [12]	Onset and coda (CC–CCCC)	Deletion, substitution and vowel insertion

Theoretical framework

Three frameworks guide the study. The Contrastive Analysis Hypothesis (CAH) holds that structures that differ between the L1 and the L2 are difficult because learners transfer L1 patterns [30]. The Markedness Differential Hypothesis (MDH) refines this by predicting that differences which are also typologically more marked will be more difficult; because coda clusters are more marked than onset clusters, codas should be harder [21, 31]. Together, the CAH and MDH predict that Vietnamese learners will repair English clusters by deletion, substitution and epenthesis, with greater difficulty in codas.

The Speech Learning Model (SLM) addresses the link between perception and production. It proposes that L2 sounds are filtered through L1 categories by equivalence classification and that accurate production depends on accurate perceptual categories [32]. The revised SLM (SLM-r) abandons the strict precedence of perception and proposes that the two abilities co-evolve, so that mismatches between them are possible for individual sounds [33]. Figure 1 points to how the three frameworks are linked to the variables and research questions.

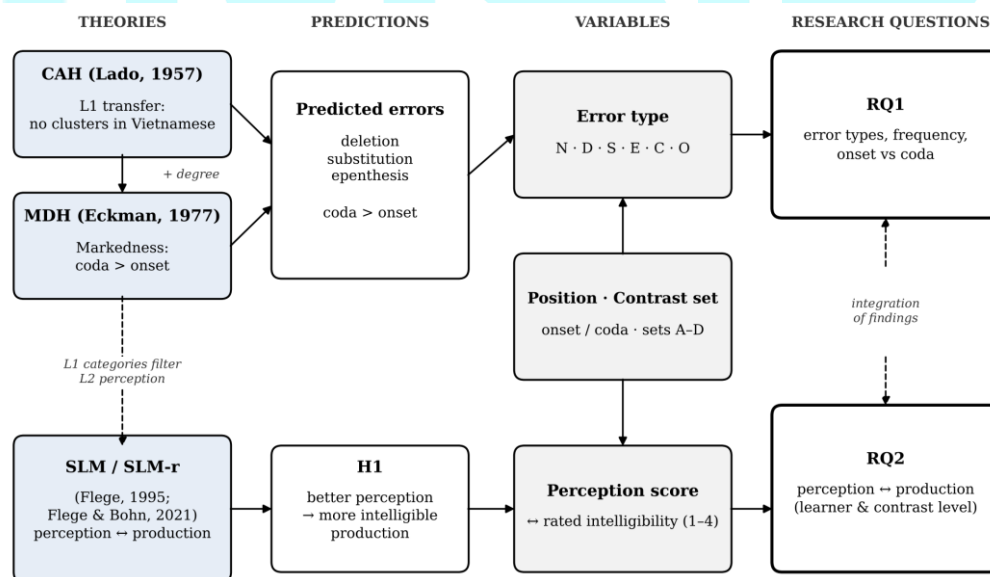


Figure 1. Conceptual framework of the study

The perception-production relationship

Empirical studies have found that L2 perception and production are related, but the relationship is neither universal nor uniform. Rauber et al. [34] found that the English vowel contrasts that Brazilian learners discriminated least accurately were also those they

produced with L1-like values. The strength of the link, however, varies with learners' L2 experience and with the properties of individual sounds [35], with word position, which may affect production more than perception [36], and with the tasks used to measure it [37, 38]. Some learners even produce a contrast more accurately than they perceive it [39, 40]. These findings are more consistent with the co-evolution view of SLM-r than with a strict precedence view. Almost all such studies, however, examined vowels or single consonants; consonant clusters, whose difficulty arises from syllable structure as well as from individual segments, remain largely unexplored..

Research gaps and the present study

Previous work, however, leaves four gaps. First, international studies have usually examined onset or coda clusters separately, and Vietnamese studies have mixed cluster lengths and positions without holding the cluster constant. Second, research on the relationship between L2 perception and production has been concentrated on vowels and single consonants [34, 35, 36], and no study has examined consonant clusters with Vietnamese learners. Third, most Vietnamese studies adopted listener judgments alone, without acoustic verification of the error classification. Fourth, non-English-major freshmen, who receive little explicit pronunciation instruction, have rarely been studied.

The present study addresses these gaps by comparing the same clusters in onset and coda positions, by measuring perception and production of the same target words in the same learners, and by verifying disputed error classifications acoustically. It asks two research questions:

RQ1: What types of two-consonant cluster errors do first-year non-English majors make, how frequently do they occur, and how do they differ between onset and coda positions?

RQ2: What is the relationship between learners' perceptual identification accuracy and the rated intelligibility of their production of English two-consonant clusters, at the learner level and at the contrast level?

Based on the original Speech Learning Model [32], one hypothesis was tested for RQ2 at the learner level. H1: learners' identification accuracy is positively correlated with the

rated intelligibility of their production. At the contrast level, following the revised model [33], the analysis was exploratory and examined whether individual contrasts show perception–production mismatches.

MATERIALS AND METHODS

Research design

The study adopted an explanatory sequential mixed-methods design [41]. In the first, quantitative phase, a large sample of 110 non-English freshmen completed a perception test. In the second phase, a volunteer subsample of 11 learners completed production tasks, which yielded quantitative data (intelligibility ratings and error codes) and qualitative data (the raters' open comments on what each token sounded like, and acoustic descriptions of disputed tokens). The two strands were integrated at the contrast level in a joint display of perception and production.

Participants

The identification test was administered with a total of 110 first-year non-English majors (60 female, 50 male; aged 18-19) from three intact general-English classes at a Nguyen Tat Thanh University, recruited by convenience sampling. Most were from the South (59.1%), followed by Central (29.1%) and Northern Vietnam (11.8%), and 63.6% had never received formal pronunciation training. Of these learners, 11 volunteered for the production phase (6 female, 5 male; five Southern, four Northern and two Central speakers). Seven of the 11 had no formal pronunciation training. Their mean perception score (19.36/24, 80.7%) was higher than that of the full sample (17.55/24, 73.1%). All participants gave informed consent and were identified only by codes (Table 2).

Table 2. Profile of the 11 production participants

Code	Gender	Region	Pronunciation training	Perception score (/24)	Mean rated intelligibility (1–4)
L01	M	South	None	24	3.20
L02	F	South	None	21	2.80
L03	F	Central	> 6 months	21	3.35
L04	M	North	None	23	3.08

Code	Gender	Region	Pronunciation training	Perception score (/24)	Mean rated intelligibility (1–4)
L05	F	North	None	20	3.06
L06	M	South	4–6 months	24	3.13
L07	F	North	1–3 months	20	3.40
L08	M	Central	None	18	3.16
L09	F	South	None	10	2.12
L10	M	North	> 6 months	17	2.86
L11	F	South	None	15	2.53

Note. Codes were reassigned for publication. Mean perception score: 19.36 (80.7%); full sample: 17.55 (73.1%).

Instruments

Target contrasts. Twelve contrasts (24 target words) were organized into four sets (Table 3). Set A holds the cluster constant and varies only its position, which isolates the effect of syllable position; Sets B and C vary the cluster within onset and coda position respectively; and Set D varies both cluster and position.

Table 3. Target contrasts

Set	Description	Contrasts (target words and clusters)
A	Same cluster, different positions	A1 list-stay /st/; A2 sport-clasp /sp/; A3 sky-mask /sk/
B	Different clusters, onset position	B1 glass-grass /gl-gr/; B2 play-pray /pl-pr/; B3 grow-slow /gr-sl/
C	Different clusters, coda position	C1 test-task /st-sk/; C2 bank-band /ŋk-nd/; C3 help-held /lp-ld/
D	Different clusters, different positions	D1 three-health /θr-lθ/; D2 twin-paint /tw-nt/; D3 fly-ask /fl-sk/

Perception test. The forced-choice A/B identification test included 24 items covering 24 consonant clusters across both onset and coda positions, each target words assigned in a semantically neutral sentence. Learners needed to listen to each sentence once before choosing which of two written words they had heard (e.g., *held* or *help*). There was a 10-second pause between sentences to make sure students could have time to complete. The

stimuli were recorded by a native speaker of American English at a natural pace. Each correct response scored one point (maximum 24). Internal consistency was acceptable (Cronbach's $\alpha = .760$) [42, 43].

Production tasks. The 11 learners volunteered to conduct controlled reading tasks, containing the list of 24 target words in isolation and then in sentences, with a total of 528 tokens (11 learners $\times$ 24 words $\times$ 2 tasks). They were recorded individually in a quiet room and the recordings were saved in WAV format for the subsequent acoustic speech analysis using Praat.

Procedure. The perception test was carried out in a regular classroom; the audio was played through a laptop, projector and loudspeakers, a ten-second pause followed each sentence, and learners responded on their own smartphones. The whole test took about 7-10 minutes. The production tasks were completed individually about one week later in a quiet room. Learners were informed to read at a natural pace without hyper-articulation, after a brief warm-up in which no pronunciation feedback was given. The raters received the recordings with learner codes only and worked independently.

Ratings and error coding

Two native English-speaking teachers experienced with Vietnamese-accented English (coded DS01 and SR02) rated each token independently on a four-point scale adapted from Niu et al. [27], from 1 (poor) to 4 (excellent; Table 4), according to how easily the target word and its cluster could be recognized. Because the raters knew the target words, this measure is referred to as *rated intelligibility* and is likely to overestimate intelligibility to listeners who do not know the target. The raters also sorted each token into no error (N), deletion (D), substitution (S), epenthesis (E), combination (C) or other (O). Each category had an auditory and an acoustic criterion: deletion demanded the absence of closure, burst or frication for the expected consonant; epenthesis revealed a vocalic interval with visible formant structure of at least 20 ms; and substitution required a change in voicing, noise source or formant transitions relative to a native model. An unreleased final stop with a visible closure was not counted as an error, because it is a normal allophone in English [3].

Table 4. Four-point rating scale (adapted from Niu et al. [27])

Score	Label	Description
4	Excellent	Fully intelligible; the target cluster is identified with no effort
3	Good	Mostly intelligible; minor deviations do not block recognition of the target
2	Fair	Moderately intelligible; errors cause some difficulty in recognizing the target
1	Poor	Hardly intelligible; the target cluster is missing, incorrect or unrecognizable

Acoustic verification

Of the 120 tokens on which the raters disagreed, 60 were analyzed in Praat [44]: the 54 tokens on the six targets with the most disagreement (*three*, *clasp*, *task*, *help*, *health* and *glass*) that had a mean rating of 2.5 or below, and the six tokens on which both raters heard an error but classified it differently. Each token was compared with a native-speaker production of the same word. The analysis was carried out by the researcher, who was not blind to the raters' codes; to limit this influence, each decision was based on the stated acoustic criteria. The remaining 60 disagreement tokens, in each of which one rater heard an error and the other did not, were excluded from the error counts.

Data analysis

Inter-rater reliability of the intelligibility ratings was estimated with two-way random-effects intraclass correlations (ICC) [45], and agreement on error categories with Cohen's κ [46]. At the learner level, each learner's perception score (0–24) was correlated with his or her mean rated intelligibility across 48 tokens. Because Pearson's r is sensitive to extreme values in small samples, the analysis also reports its 95% confidence interval, Spearman's ρ and a leave-one-out analysis. At the contrast level, perception was the number of correct responses to the two items of a contrast (0–2); because this measure is ordinal with many ties, Spearman's ρ was used, and the 12 p-values were adjusted with the Holm procedure [47]. Coefficients were interpreted following Schober et al. [48].

RESULTS

Perception of CC clusters

Mean identification accuracy was 73.1% ($M = 17.55/24$). Accuracy varied widely across contrasts, from 53.2% for A2 (*sport-clasp*) to 85.0% for B3 (*grow-slow*) (Table 5). Onset items were identified more accurately than coda items (82.9% vs 63.4% on average). The positional effect was clearest in Set A, where the same clusters were identified far more accurately in onset than in coda position: /sp/ (87.3% vs 19.1%), /st/ (87.3% vs 63.6%) and /sk/ (88.2% vs 70.9%). The lowest accuracies were for *clasp* (19.1%), which fell below chance because most learners chose its competitor *class*, and *health* (46.4%).

Descriptive comparisons across background variables showed small differences and no consistent pattern. Learners with 4–6 months of formal pronunciation training had the highest mean score ($M = 19.78$, $n = 9$), Northern speakers scored slightly higher ($M = 18.85$, $n = 13$) than Southern ($M = 17.40$, $n = 65$) and Central speakers, and self-rated proficiency was not related to accuracy in a consistent way. Because several subgroups were small, these comparisons were not tested statistically.

Table 5. Identification accuracy by contrast ($N = 110$)

Contrast	Item 1 (%)	Item 2 (%)	Mean (%)
A2 sport - clasp /sp/	87.3	19.1	53.2
D1 three - health /θr - lθ/	76.4	46.4	61.4
C3 help - held /lp - ld/	67.3	56.4	61.8
C2 bank - band /ŋk - nd/	73.6	69.1	71.4
D2 twin - paint /tw - nt/	73.6	70.0	71.8
A1 list - stay /st/	63.6	87.3	75.5
C1 test - task /st - sk/	68.2	83.6	75.9
A3 sky - mask /sk/	88.2	70.9	79.5
D3 fly - ask /fl - sk/	87.3	72.7	80.0
B1 glass - grass /gl - gr/	80.0	81.8	80.9
B2 play - pray /pl - pr/	88.2	74.5	81.4
B3 grow - slow /gr - sl/	84.5	85.5	85.0

Production intelligibility and rater reliability

Mean rated intelligibility was 2.98 in word reading and 2.97 in sentence reading. The averaged ratings were reliable overall (Table 6; ICC(2,2) = .868 for words and .848 for sentences), although reliability was poor for four contrasts (A1, C1, D1 and D3), whose means should be read with caution. SR02 rated words slightly more leniently than DS01 (+0.20, $t(10) = 3.10$, $p = .011$), but not sentences. Agreement on error categories was moderate ($\kappa = .445$; 77.3% identical codes), and disagreement ran in both directions (McNemar exact $p = .223$).

Table 6. Inter-rater reliability

Measure	Word reading	Sentence reading
ICC(2,1), single rater	.766	.736
ICC(2,2), average of two raters	.868	.848
Rater mean difference (SR02 - DS01)	+0.20 ($p = .011$)	+0.12 ($p = .177$)
Cohen's κ , six error categories	.387	.496

Note. Over all 528 tokens, $\kappa = .445$ (77.3% agreement); for error vs. no error, $\kappa = .512$ (81.6%).

Onset clusters were produced more intelligibly than coda clusters (mean of target means 3.36 vs 2.58), and ten of the eleven learners showed this pattern. Table 7 shows the same asymmetrical data within Set A: *stay* (3.93) vs *list* (3.00), *sport* (3.25) vs *clasp* (1.25), and *sky* (3.95) vs *mask* (2.43). The least intelligible targets were *clasp* (1.25), *help* (2.09) and *three* (2.27).

Reading context made little difference overall (2.98 vs 2.97), but it affected some coda targets. The C1 contrast fell from 2.91 in words to 2.48 in sentences, mainly because *task* was more often reduced in connected speech, whereas onset targets such as *sky*, *slow* and *fly* stayed near ceiling in both tasks as learners could identify these targets more accurately.

Table 7. Mean rated intelligibility by target (1–4; 11 learners, two raters, both tasks)

Set	Target 1: M (SD)	Target 2: M (SD)	Contrast M
A1	stay (O) 3.93 (0.18)	list (C) 3.00 (0.67)	3.47
A2	sport (O) 3.25 (0.72)	clasp (C) 1.25 (0.40)	2.25
A3	sky (O) 3.95 (0.15)	mask (C) 2.43 (1.24)	3.19
B1	glass (O) 2.61 (1.01)	grass (O) 3.30 (0.73)	2.95

Set	Target 1: M (SD)	Target 2: M (SD)	Contrast M
B2	play (O) 3.61 (0.55)	pray (O) 3.57 (0.78)	3.59
B3	grow (O) 3.05 (0.99)	slow (O) 3.98 (0.11)	3.51
C1	test (C) 2.98 (0.72)	task (C) 2.41 (0.81)	2.69
C2	bank (C) 3.07 (0.90)	band (C) 2.77 (0.74)	2.92
C3	help (C) 2.09 (0.93)	held (C) 2.48 (1.05)	2.28
D1	three (O) 2.27 (1.02)	health (C) 2.98 (0.79)	2.63
D2	twin (O) 2.91 (1.08)	paint (C) 2.52 (0.92)	2.72
D3	fly (O) 3.89 (0.31)	ask (C) 3.02 (0.92)	3.45

Note. O = onset; C = coda.

Error types and syllable position (RQ1)

In the Praat verification, the acoustic decision matched DS01 in 20 tokens, SR02 in 20 and neither rater in 20, so neither rater was systematically closer to the signal. After verification, 468 tokens were classified, of which 117 contained an error (Table 9). Deletion was the dominant error type (67.5%), followed by substitution (22.2%) and epenthesis (5.1%); combinations and metathesis were rare (treated in other errors). Coda clusters were in error more than three times as often as onset clusters (39.2% vs 11.6% of tokens), and the two positions were repaired differently. In codas, 82.0% of errors were deletions of one member of the cluster (e.g., *clasp* heard as “clas” or “clap”, *task* as “tas”, *help* as “hep”); in onsets, substitution was most frequent, mainly /θ/ → [t] in *three* and /g/ → [k] in *glass*, which produced the competitor *class*. All six epenthesis tokens occurred in onset obstruent + liquid clusters (e.g., *grow* → [gərou]), and three *clasp* tokens showed metathesis ([klæps]).

Table 8 summarizes the acoustic verification for the six priority targets. Deletion was the dominant acoustic decision for the coda targets *clasp* and *task*, where the spectrogram showed /s/ frication followed by no closure or release burst for the final stop. Substitution dominated for *three* and *health*, where /θ/ was realized with a stop closure and burst instead of dental frication, and for *glass*, where the onset /g/ lacked prevoicing. For *three*, the acoustic decision matched neither rater in 8 of 12 tokens, mainly because the raters had classified a shared percept differently, which shows the value of an independent acoustic criterion for disputed tokens.

Table 8. Acoustic verification for the six priority targets

Target	Disagreements	Analyzed	Dominant acoustic decision	Matched neither rater
three /θr/	14	12	Substitution	8
clasp /sp/	11	11	Deletion	2
task /sk/	11	10	Deletion	1
help /lp/	9	5	Deletion / substitution	2
health /lθ/	9	9	Substitution	5
glass /gl/	8	7	Substitution	1
Total	62	54		19

Table 9. Final distribution of error types by syllable position

Error type	All errors n (%)	Onset n (%)	Coda n (%)
Deletion	79 (67.5)	6 (21.4)	73 (82.0)
Substitution	26 (22.2)	16 (57.1)	10 (11.2)
Epenthesis	6 (5.1)	6 (21.4)	0
Combination	3 (2.6)	0	3 (3.4)
Other (metathesis)	3 (2.6)	0	3 (3.4)
Total	117	28	89

Note. 468 classified tokens (408 rater-agreed, 60 acoustically resolved).

A sensitivity analysis examined the 60 excluded tokens. When these were assigned DS01's codes and then SR02's codes, the number of errors rose to 143 and 151, but the distribution changed little: deletion accounted for 69.9% and 64.9% of errors, substitution for 20.3% and 25.2%, and epenthesis for 4.9% and 4.0%, while deletion still made up 84.9% and 81.7% of coda errors. The exclusion therefore did not change the answer to RQ1.

The raters' open comments illustrate these patterns. For coda targets, the raters described reduced or missing final consonants, for example "no s sound" for *clasp*, "weak or no k" for *bank*, "weak d" for *band*, "very weak l" for *help* and "weak t" for *paint*, which shows that some reductions were gradient rather than categorical. For onset targets, the comments were mostly positive (e.g., "sp sound is clear and recognizable" for *sport*); the exceptions

were substitutions, such as *glass* heard as *class* and *three* heard as *tree*, and the epenthesis noted as “the learner added a vowel” in *grow*. Disagreement between the raters was greatest for tokens of intermediate intelligibility (mean rating 2.0-2.5), where they agreed on the error category in only 39.8% of cases, compared with 98.9% for tokens rated 4.

The perception-production relationship (RQ2)

At the learner level, perception scores and mean rated intelligibility were positively correlated, $r(9) = .81$, $p = .003$, 95% CI [.40, .95] (Figure 2). The estimate, however, was sensitive to one learner (L09), who had by far the lowest perception score (10/24) and the lowest intelligibility (2.12). Without this learner the correlation fell to $r(8) = .56$, $p = .089$, whereas removing any other learner left it between .78 and .85. The rank-based correlation for all learners was $\rho = .55$, $p = .077$. H1 was therefore supported in direction, but the strength of the relationship is uncertain.

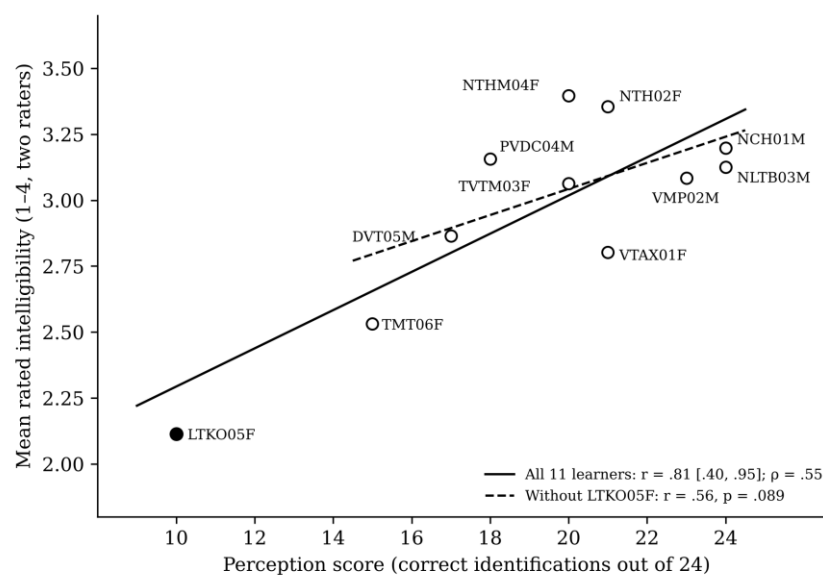


Figure 2. Perception score and mean rated intelligibility of the 11 learners (solid line: all learners; dashed line: without the lowest-scoring learner)

At the contrast level, the relationships were weak (Table 10). Spearman’s ρ ranged from -.08 to .50; three contrasts showed moderate associations, five weak and four negligible, and none was significant before or after Holm correction. Because contrast-level perception had only three possible values and several contrasts were near ceiling, the relationship was interpreted mainly through the joint display in Table 11.

Table 10. Contrast-level Spearman correlations between perception and production (n = 11)

Contrast	ρ	p (Holm p)	Strength
A3 sky - mask	.50	.115 (1.00)	Moderate
D2 twin - paint	.50	.115 (1.00)	Moderate
D3 fly - ask	.44	.177 (1.00)	Moderate
D1 three - health	.29	.388 (1.00)	Weak
A1 list - stay	.18	.591 (1.00)	Weak
C3 help - held	.17	.610 (1.00)	Weak
C1 test - task	.15	.658 (1.00)	Weak
B1 glass - grass	.15	.661 (1.00)	Weak
C2 bank - band	.09	.801 (1.00)	Negligible
B2 play - pray	-.03	.928 (1.00)	Negligible
A2 sport - clasp	-.05	.882 (1.00)	Negligible
B3 grow - slow	-.08	.826 (1.00)	Negligible

Note. Perception = number of correct responses to the two items of each contrast (0-2). Values in parentheses are Holm-adjusted p-values.

Table 11. Joint display of perception and production by contrast

	Relatively well produced ($M \geq 2.94$)	Relatively poorly produced ($M < 2.94$)
Relatively well perceived ($\geq 79.5\%$)	A3 sky-mask (95.5%, 3.19); B1 glass-grass (90.9%, 2.95); B3 grow-slow (90.9%, 3.51); D3 fly-ask (81.8%, 3.45)	C1 test-task (95.5%, 2.69); D2 twin-paint (90.9%, 2.72)
Relatively poorly perceived ($< 79.5\%$)	A1 list-stay (68.2%, 3.47); B2 play-pray (68.2%, 3.59)	A2 sport-clasp (68.2%, 2.25); C2 bank-band (68.2%, 2.92); C3 help-held (77.3%, 2.28); D1 three-health (72.7%, 2.63)

Note. Values are the 11 learners' mean identification accuracy and mean rated intelligibility; cut-offs are the medians of the 12 contrasts.

Eight contrasts were aligned. Four were relatively well perceived and well produced (A3, B1, B3, D3), all with their difficulty mainly in onset position. Four were poorly perceived

and poorly produced (A2, C2, C3, D1), including the least intelligible targets (*clasp*, *help*, *three*) and both /θ/ targets. Four contrasts, however, were mismatched. C1 (*test-task*) and D2 (*twin-paint*) were identified very accurately (95.5% and 90.9%) yet produced poorly (2.69 and 2.72): all eleven learners noticed *task* correctly, but every token analyzed acoustically showed deletion in its coda /sk/. Conversely, A1 (*list-stay*) and B2 (*play-pray*) were among the least accurately perceived contrasts (68.2%) yet among the most intelligibly produced (3.47 and 3.59).

DISCUSSION

Cluster errors and syllable position

The error profile follows the Vietnamese syllable closely. Deletion dominated because an English coda cluster violates two Vietnamese constraints at once, and deleting one consonant restores the familiar CVC shape without adding a syllable. This is consistent with the CAH [30] and with earlier Vietnamese studies [8, 9, 10], although the proportion of deletion here (67.5%) is lower than the 85-87% reported previously, possibly because the present targets were limited to two-consonant clusters and because acoustic verification identified substitutions that listeners had not detected. The concentration of errors in codas, and the within-learner contrast between intact onset /sp/ and /sk/ and reduced coda /sp/ and /sk/, support the MDH prediction that the more marked coda position is harder [21]. Within codas, heterorganic clusters such as /sp/, /sk/ and /lp/ were reduced far more often than homorganic /st/, /nd/ and /ŋk/.

Substitution was conditioned by segment identity rather than position: /θ/ was replaced by [t] in both *three* and *health*, a pattern consistent with equivalence classification of /θ/ with the Vietnamese stop associated with the spelling *th* [32]. Its effect on intelligibility depended on functional load and lexical competition [26]: /g/ → [k] in *glass* produced the real word *class*, whereas *health* realized as [helt] remained recoverable. Epenthesis was rare and restricted to onsets, in contrast to Arabic, Japanese and Spanish learners [17, 19, 20]. Because Vietnamese freely permits closed syllables, deletion offers a familiar repair in codas, while in onsets a vowel can preserve both consonants at the beginning of the word.

The positional pattern also bears on intelligibility. Jenkins's Lingua Franca Core treats onset deletion as the most serious threat to intelligibility (as reviewed in Levis [3]), but in

these data onset clusters were rarely simplified, and the loss of intelligibility came mainly from coda deletion. This agrees with evidence that final consonant errors are particularly damaging for Vietnamese speakers [3] and suggests that coda clusters should not be treated as a low priority for this population.

The perception-production relationship

At the learner level, learners who identified clusters more accurately tended to produce them more intelligibly, in line with the SLM [32] and with earlier findings of a positive link [34, 38]. The close alignment of the two tasks, which used the same target words in sentence contexts, may have favored this link [38]. However, the dependence of the estimate on a single learner and the non-significant Spearman coefficient mean that H1 is supported only tentatively.

At the contrast level, the weak and non-uniform relationships are more consistent with the co-evolution view of SLM-r [33] than with a strict perception-first view. The C1 *test-task* and D2 *twin-paint* cases show that accurate perception does not guarantee accurate production when the target structure is not permitted in the L1: learners heard the coda in *task* and *paint* but reduced it in production, echoing the finding that word position constrains production more than perception [36]. The A1 *list-stay* and B2 *play-pray* cases show the reverse asymmetry, as reported for Japanese learners' /r/-/l/ [39] and for Mandarin consonants [40]. These cases should be interpreted carefully, because their perception scores were low only relatively (68.2%, well above chance) and because ratings by listeners who knew the targets may overstate production accuracy. Overall, the results are consistent with the view that accurate perception is necessary, but is not always sufficient, for accurate production of English clusters.

Theoretical implications

Taken together, the three frameworks account for different parts of the data. The CAH explains *where* errors occur: in structures, such as coda clusters and /θ/, that have no counterpart in Vietnamese. The MDH explains *how much* difficulty arises: within the set of L1-L2 differences, the more marked coda position and heterorganic clusters were the most error-prone. Neither framework, however, predicts the specific repair chosen; the dominance of deletion over epenthesis appears to reflect an L1-specific preference shaped

by the availability of closed syllables in Vietnamese [3, 24]. The SLM accounts for errors that originate in perception, such as the shared perceptual difficulty with *clasp* and the /θ/ targets, whereas SLM-r is needed to explain contrasts that were perceived accurately yet produced poorly. The data therefore suggest that, for consonant clusters, perception is one source of difficulty among several, alongside phonotactic transfer and articulatory constraints at the end of the syllable.

Methodological implications

The study also shows why listener judgments of cluster errors should be verified. The two raters agreed well on overall intelligibility but only moderately on error categories, and the acoustic analysis matched each rater equally often and neither rater in a third of the disputed tokens. DS01 rated more strictly, attending to whether each target was accurately pronounced, whereas SR02, who had more experience with Vietnamese learners, gave more weight to whether the word as a whole was understood; both perspectives are legitimate, but they lead to different codes for gradient realizations such as weakened final stops. Combining independent ratings with explicit acoustic criteria, as in the present design, makes error classification more transparent and allows readers to judge how much of the error profile rests on disputed tokens. The sensitivity analysis showed that the main conclusions for RQ1 did not depend on how the unresolved tokens were treated.

Pedagogical implications

The findings suggest practical pedagogical implications for English teachers. First, instruction for this population should focus more on coda clusters, especially heterorganic codas such as /-sp/, /-sk/ and /-lp/, and to /θ/. Second, because accurate perception did not always translate into accurate production, perception training (e.g., minimal-pair identification) should be combined with explicit articulatory practice and feedback. Third, practice should move from isolated words to sentences, where learners are more likely to revert to the Vietnamese syllable shape.

For curriculum designers, the contrast sets used in this study offer a simple diagnostic tool: comparing the same cluster in onset and coda position (e.g., *sport* and *clasp*) quickly shows whether a learner's difficulty is positional, and the four perception–production patterns in Table 11 indicate whether a learner needs perceptual training, production

practice or both. Because the raters agreed least on tokens of intermediate intelligibility, teachers who assess pronunciation should also be aware that impressionistic judgments of cluster errors can vary considerably between listeners, and that recordings or simple acoustic displays can support more consistent feedback.

Limitations and future research

The study has several limitations. The production sample was a small volunteer subsample whose perception scores were above the cohort mean, and the learner-level correlation depended on one learner; future studies should sample learners across the range of perception ability. Regional dialect was not controlled. The raters knew the target words, and per-contrast reliability was poor for four contrasts. The acoustic analysis was not blind to the raters' codes, half of the disputed tokens were not analyzed acoustically, and after /s/ or /l/ an unreleased stop may be difficult to distinguish from deletion. Finally, production was elicited through reading tasks, so error rates in natural, spontaneous conversations may differ. Overall, larger samples from different learners' levels, data collection from freer speaking activities, and intelligibility ratings would allow these findings to be tested more rigorously.

Future research could extend the design in three directions. First, the same position-controlled contrasts could be used with three-consonant clusters and with English majors, to test whether the onset–coda asymmetry persists at higher proficiency. Second, a training study could test whether perception training, production training or a combination improves coda clusters most effectively. Third, acoustic measures could be applied to all tokens rather than to disputed tokens only, which would allow gradient reductions such as weakened final stops to be quantified.

CONCLUSION

This study examined English two-consonant cluster errors and the perception-production correlation among Vietnamese non-English-major freshmen. Errors were systematic and consistent with Vietnamese phonotactics: deletion dominated, especially in codas, substitution affected segments absent from Vietnamese, and epenthesis was rare. Perception and production were positively related across learners, but the relationship was

weak and uneven across contrasts, and several contrasts were perceived well but produced poorly or the reverse.

The main contribution of the study is position-controlled evidence, obtained from the same learners and the same target words, that the difficulty of English two-consonant clusters for these learners lies chiefly in coda position, and that accurate perception of a cluster does not guarantee its intelligible production. The study also shows that acoustic verification can change the classification of listener-coded errors, particularly by revealing substitutions that listeners did not detect.

REFERENCES

1. Crystal, D. *English as a global language*. 2nd ed. Cambridge University Press; 2003.
2. Seidlhofer, B. *Understanding English as a lingua franca*. Oxford University Press; 2011.
3. Levis, J. M. *Intelligibility, oral communication, and the teaching of pronunciation*. Cambridge University Press; 2018. <https://doi.org/10.1017/9781108241564>
4. Derwing, T. M., Munro, M. J. *Pronunciation fundamentals: Evidence-based perspectives for L2 teaching and research*. John Benjamins; 2015. <https://doi.org/10.1075/llt.42>
5. Thompson, L. C. *A Vietnamese grammar*. University of Washington Press; 1965.
6. Kirby, J. P. Vietnamese (Hanoi Vietnamese). *Journal of the International Phonetic Association*. 2011;41(3):381–392. <https://doi.org/10.1017/S0025100311000181>
7. Roach, P. *English phonetics and phonology*. 4th ed. Cambridge University Press; 2009.
8. Nguyen, N. Interlanguage phonology and the pronunciation of English final consonant clusters by native speakers of Vietnamese [master's thesis]. Ohio University; 2008.
9. Tran, T. U. Major problems in pronouncing English: A case study at the University of Dalat. *Tạp chí Khoa học Ngôn ngữ và Văn hóa*. 2019;3(3):418–427. <https://doi.org/10.63506/jilc.0303.189>

10. Nguyen, T. H. Strategies to correct errors in pronouncing final consonant clusters by second-year English major students at Hong Duc University. *Hong Duc University Journal of Science*. 2022;12(E7):12–22. <https://vjol.info.vn/HDU/article/view/68029>
11. Tran, T. K. L., Nguyen, A. T. Common mistakes in pronouncing English consonant clusters: A case study of Vietnamese learners. *CTU Journal of Innovation and Sustainable Development*. 2022;14(3):32–39. <https://doi.org/10.22144/ctu.jen.2022.040>
12. Phi, T. T. T. Problems in the pronunciation of English consonant clusters: A case study of English-majored students at a Vietnamese university. *International Journal of Social Science and Human Research*. 2025;8(9):7343–7349. <https://doi.org/10.47191/ijsshr/v8-i9-77>
13. Cruttenden, A. *Gimson's pronunciation of English*. 7th ed. Hodder Education; 2008.
14. Dinh, L. T., Nguyen, V. H. *Cơ cấu ngữ âm tiếng Việt* [Structure of Vietnamese phonetics]. Nhà xuất bản Giáo dục; 1998.
15. Doan, T. T. *Ngữ âm tiếng Việt* [Vietnamese phonetics]. Nhà xuất bản Đại học Quốc gia Hà Nội; 1999.
16. Hwa-Froelich, D. A., Hodson, B. W., Edwards, H. T. Characteristics of Vietnamese phonology. *American Journal of Speech-Language Pathology*. 2002;11(3):264–273. [https://doi.org/10.1044/1058-0360\(2002/028\)](https://doi.org/10.1044/1058-0360(2002/028))
17. Al-Yami, E. M., Al-Athwary, A. A. H. Phonological analysis of errors in the consonant cluster system encountered by Saudi EFL learners. *Theory and Practice in Language Studies*. 2021;11(10):1237–1248. <https://doi.org/10.17507/tpls.1110.11>
18. Bouchhioua, N. Epenthesis in the production of English consonant clusters by Tunisian EFL learners. *Applied Linguistics Research Journal*. 2019;3(4):33–44.
19. Carlisle, R. S. The influence of environment on vowel epenthesis in Spanish/English interphonology. *Applied Linguistics*. 1991;12(1):76–95. <https://doi.org/10.1093/applin/12.1.76>

20. Tajima, K., Erickson, D., Nagao, K. Production of syllable structure in a second language: Factors affecting vowel epenthesis in Japanese-accented English. *IULC Working Papers*. 2002;2(2):77–92. <https://scholarworks.iu.edu/journals/index.php/iulcwp/issue/view/1697>
21. Eckman, F. R. Markedness and the contrastive analysis hypothesis. *Language Learning*. 1977;27(2):315–330. <https://doi.org/10.1111/j.1467-1770.1977.tb00124.x>
22. Major, R. C. Markedness in second language acquisition of consonant clusters. In: Bayley, R., Preston, D. R., editors. *Second language acquisition and linguistic variation*. John Benjamins; 1996. p. 75–96. <https://doi.org/10.1075/sibil.10.04maj>
23. Benson, B. The markedness differential hypothesis: Implications for Vietnamese speakers of English. In: Eckman, F., Moravcsik, E., Wirth, J., editors. *Markedness*. Springer; 1986. p. 271–289.
24. Sato, C. J. Phonological processes in second language acquisition: Another look at interlanguage syllable structure. *Language Learning*. 1984;34(4):43–57. <https://doi.org/10.1111/j.1467-1770.1984.tb00351.x>
25. Munro, M. J., Derwing, T. M. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning*. 1995;45(1):73–97. <https://doi.org/10.1111/j.1467-1770.1995.tb00963.x>
26. Munro, M. J., Derwing, T. M. The functional load principle in ESL pronunciation instruction: An exploratory study. *System*. 2006;34(4):520–531. <https://doi.org/10.1016/j.system.2006.09.004>
27. Niu, Z., Hao, Y., Liu, S. Assessing pronunciation proficiency of Chinese learners of English: Development and validation of an intelligibility rating scale. *TESOL Communications*. 2022;1(1):27–47. <https://doi.org/10.46451/tc.20220103>
28. Ha, C. T. Common pronunciation problems of Vietnamese learners of English. *VNU Journal of Science: Foreign Languages*. 2005;21(1):35–46. <https://js.vnu.edu.vn/FS/article/view/2072>

29. Nguyen, T. T. H. English consonant clusters in the pronunciation of Vietnamese learners: A study on two students at Ha Long University. *Tạp chí Thiết bị Giáo dục*. 2024;314(1):197–199.
30. Lado, R. *Linguistics across cultures: Applied linguistics for language teachers*. University of Michigan Press; 1957.
31. Anderson, J. The markedness differential hypothesis and syllable structure difficulty. In: Ioup, G., Weinberger, S. H., editors. *Interlanguage phonology: The acquisition of a second language sound system*. Newbury House; 1987. p. 279–291.
32. Flege, J. E. Second language speech learning: Theory, findings, and problems. In: Strange, W., editor. *Speech perception and linguistic experience: Issues in cross-language research*. York Press; 1995. p. 233–277.
33. Flege, J. E., Bohn, O.-S. The revised Speech Learning Model (SLM-r). In: Wayland, R., editor. *Second language speech learning: Theoretical and empirical progress*. Cambridge University Press; 2021. p. 3–83.
34. Rauber, A. S., Escudero, P., Bion, R. A. H., Baptista, B. O. The interrelation between the perception and production of English vowels by native speakers of Brazilian Portuguese. In: *Proceedings of Interspeech 2005*. 2005. p. 2913–2916. <https://doi.org/10.21437/Interspeech.2005-769>
35. Cho, M.-H., Lee, S. The perception-production link varies with stages of L2 development and vowel properties. *Poznań Studies in Contemporary Linguistics*. 2023;59(2):317–342. <https://doi.org/10.1515/psicl-2023-2003>
36. Bin Hadjah, N. A. K., Hamzah, M. H. Production and perception of English consonants by Yemeni EFL learners. *Indonesian Journal of Applied Linguistics*. 2023;12(3):719–738. <https://doi.org/10.17509/ijal.v12i3.47474>
37. Kim, D., Clayards, M. Individual differences in the link between perception and production and the mechanisms of phonetic imitation. *Language, Cognition and Neuroscience*. 2019;34(6):769–786. <https://doi.org/10.1080/23273798.2019.1582787>

38. Llompart, M. On the effects of task focus and processing level on the perception–production link in second-language speech learning. *Studies in Second Language Acquisition*. 2023;46(1):214–226.
39. Sheldon, A., Strange, W. The acquisition of /r/ and /l/ by Japanese learners of English: Evidence that speech production can precede speech perception. *Applied Psycholinguistics*. 1982;3(3):243–261. <https://doi.org/10.1017/S0142716400001417>
40. Wang, X. L2 speech learning: Perception, production & training. *Cadernos de Linguística*. 2020;1(1):1–22. <https://doi.org/10.25189/2675-4916.2020.v1.n1.id280>
41. Creswell, J. W., Plano Clark, V. L. *Designing and conducting mixed methods research*. 3rd ed. SAGE Publications; 2018.
42. Cronbach, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika*. 1951;16(3):297–334. <https://doi.org/10.1007/BF02310555>
43. Nunnally, J. C., Bernstein, I. H. *Psychometric theory*. 3rd ed. McGraw-Hill; 1994.
44. Boersma, P., Weenink, D. Praat: Doing phonetics by computer (Version 6.4.67) [Computer software]. 2026. <https://www.praat.org>
45. Koo, T. K., Li, M. Y. A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*. 2016;15(2):155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
46. Landis, J. R., Koch, G. G. The measurement of observer agreement for categorical data. *Biometrics*. 1977;33(1):159–174. <https://doi.org/10.2307/2529310>
47. Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*. 1979;6(2):65–70.
48. Schober, P., Boer, C., Schwarte, L. A. Correlation coefficients: Appropriate use and interpretation. *Anesthesia & Analgesia*. 2018;126(5):1763–1768. <https://doi.org/10.1213/ANE.0000000000002864>