Children’s Articulation Practice in SpeechLP Across Two Months

Shail Silver; Alissa Fox, M.S., CCC-SLP; Mari Gottlieb, M.A., CCC-SLP; Matthew Siu; Jabali Ndungu

Affiliation: SpeechLP (all authors)

Retrospective observational study · November 3, 2026

Abstract

Purpose. To describe practice and changes in accepted-card accuracy across two months among returning SpeechLP profiles.

Method. This retrospective observational study used routine records from January 22 through October 7, 2026. The cohort comprised 101 profiles with full 60-day observation opportunity and matching sound targets in consecutive 30-day periods. Accepted-card proportions were averaged across matched sounds within profiles, then equally across profiles. Confidence intervals used 10,000 whole-profile bootstrap resamples. Small cells and complementary cells were suppressed. There was no study-specific consent or independent ethics review; the company confirmed permission under the app’s terms and privacy policy.

Result. Across 660 matched profile-and-sound pairs, mean accepted-card accuracy increased from 69.4% to 74.9% (+5.5 percentage points; 95% CI 0.4 to 10.4). Of these pairs, 278 increased, 177 decreased and 205 were unchanged. Mean sessions per profile declined from 6.73 to 3.44; mean practice days declined from 4.70 to 2.71.

Conclusion. Accepted-card accuracy was higher in the second month in this selected returning cohort. Scores may reflect repeated utterances, auditory modeling and recognition/scoring changes; app version was not adjusted for. Unvalidated scoring and the absence of a comparison group prevent attribution to practice or interpretation as clinical articulation improvement.

Disclosure. These findings are preliminary and descriptive. Acceptance rates reflect the app’s automated speech recognition decisions, which have not been validated against SLP ratings, and may also reflect changes in app scoring over time. They should not be interpreted as clinical articulation accuracy, within-platform growth, or evidence of treatment efficacy.

Keywords: speech sound disorders; articulation; mobile applications; automated speech recognition; retrospective observational study; home practice

Introduction

Speech sound acquisition is developmental, and some errors are typical at younger ages (American Speech-Language-Hearing Association [ASHA], n.d.). A review of 15 U.S. studies covering 18,907 children found that most English consonants were acquired by age 5;0 under the review's 90% acquisition criterion, with /ɹ/, /θ/, /ð/ and /ʒ/ among the later-acquired sounds (Crowe & McLeod, 2020). These population norms are not individual diagnostic deadlines. ASHA reports estimated speech sound disorder prevalence of 2.1%–23% at ages 4–6 and persistent speech sound disorder in about 3.6% of children at age 8; early speech sound difficulties are also associated with later literacy problems (ASHA, n.d.). The app records do not establish whether profiles had a clinically diagnosed speech sound disorder.

Clinical treatment commonly progresses through establishment, generalization to more complex contexts and maintenance with self-monitoring (ASHA, n.d.). Card-level app scores describe structured-practice performance; they do not establish a child's treatment stage or demonstrate generalization or maintenance. Because 55 of the 101 profiles were recorded as ages 2–5, typical development is one possible contributor to observed changes.

Following a sound over time requires distinguishing a change in that sound from a change in the targets a child practices. A pooled score across different targets in different months can reflect the target mix as well as performance. This study therefore compares identical sound labels within the same profile across two fixed periods.

The study addresses three questions: how much these profiles practiced in each month, how their average accuracy on matched sounds changed, and how individual profile-and-sound changes were distributed.

Methods

Study design and observation window

This is exploratory retrospective product research using an existing cached extract of practice events. The source observation window runs from January 22, 2026, inclusive, to October 8, 2026, exclusive, in America/Toronto time. The month comparison and its sound-direction analyses were developed after inspecting the existing practice data; the investigation was not prospectively registered.

Ethics and data protection

This retrospective study used routine practice-result records from independent family accounts, including timestamps, target labels, recorded profile characteristics and card outcomes. No speech recordings were accessed. Source extracts contained account identifiers, profile names and email addresses; hashed profile keys remained relinkable through retained source records. The company confirmed that parents accepted the applicable terms and privacy policy and that these covered secondary analysis and aggregate publication. Separate study-specific consent was not obtained; historical signup screens and policies were available, but individual acceptance logs were unavailable. Only aggregate results are reported, with suppression of cells based on fewer than five profiles and complementary suppression to prevent reconstruction. The

study received company internal review only; no independent IRB/REB approval, exemption determination or consent waiver was obtained.

Eligible records and profile construction

Eligible practice records exclude identified internal, staff, demonstration and test accounts; clinician-role accounts; screening and coverage games; missing app-version records; version 3.x records; and cards with ambiguous or multiple target-sound settings.

A profile is constructed from a family account and its normalized child-profile label. It represents a recorded profile rather than an independently verified unique child. The selected monthly cohort has 101 distinct family accounts, with one matched profile per account.

Cohort selection and month boundaries

For each profile, the earliest observed eligible card anchors elapsed time zero. Month 1 includes timestamps from zero up to, but not including, day 30. Month 2 includes day 30 up to, but not including, day 60. Each day is a 24-hour elapsed interval. A profile must have the entire 60-day interval within the fixed source observation window. These are consecutive elapsed windows, not named calendar months, and the first observed card need not represent first-ever app use. Month 1 is therefore an observed practice period, not a pre-practice baseline.

The primary cohort comprised the 101 recorded profiles with complete 60-day observation opportunity and at least one identical sound target scored in both periods. These are described as returning profiles: each had some practice in Month 2, not necessarily regular practice throughout either month. Profile selection is shown in Figure 1.

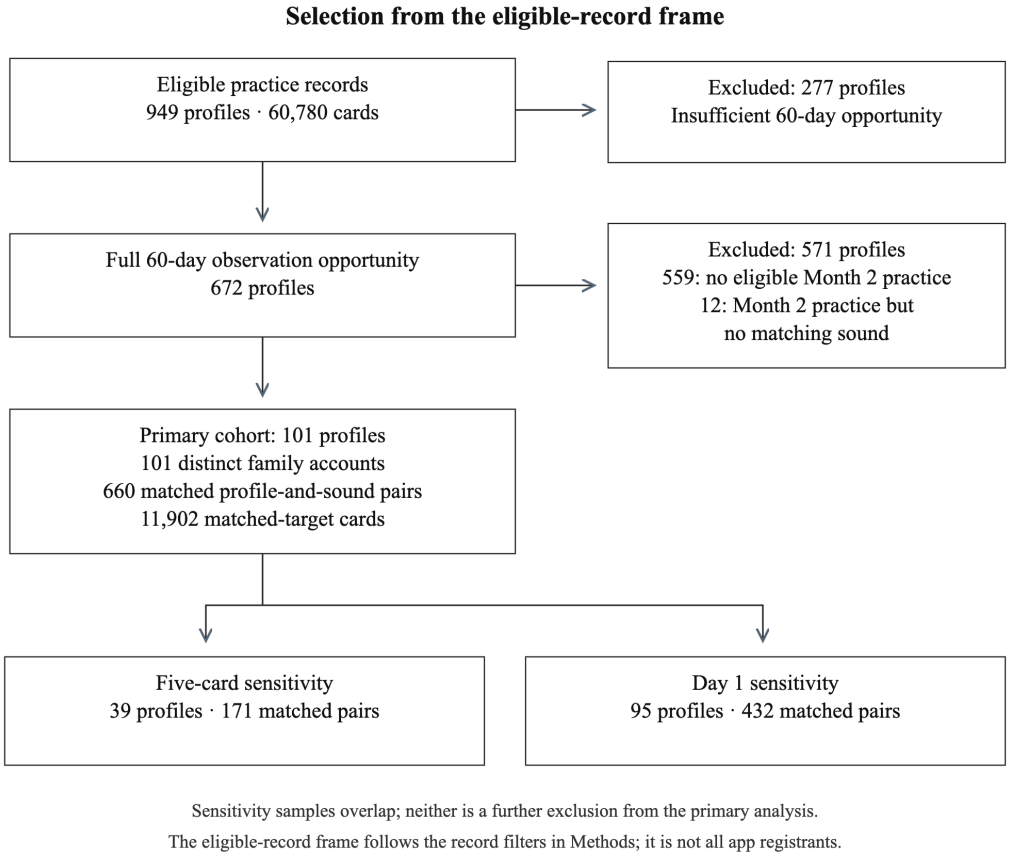


Figure 1. Profile selection after card-eligibility filtering. Profiles excluded at each stage are counted once. The five-card and Day 1 sensitivity samples overlap and are not additional exclusions from the primary cohort.

Table 1. Analyzed sample and matched accuracy records.

Analytic measure	Count
Child profiles included in the study	101
Matched profile-and-sound pairs	660
Cards used in matched accuracy	11902

Practice-card task and recorded outcome

In the studied pre-3.x iOS app, each practice card showed a pictured target word and automatically played a spoken model. Later 1.x and 2.x screens allowed replay without a programmed limit while the card was active, but an early 1.0 screen had no replay control, so replay availability cannot be assumed for every studied version. Speech recognition evaluated the word against the card's selected target-sound settings. Games differed in when they advanced after an unsuccessful encounter: Perry Pilot advanced after the second unsuccessful obstacle encounter, whereas Chicken Little advanced after one missed collision. The encounter counter did not count spoken attempts. A child could produce several utterances while a card remained active, so the two-count setting did not impose a two-utterance limit. A successful evaluation advanced the game, and a card was recorded as accepted if any evaluation during it succeeded (Figure 2).

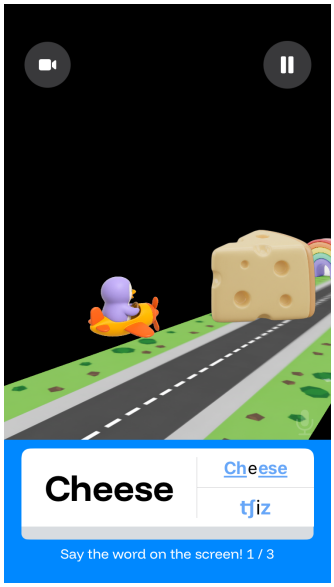


Figure 2. Archived pre-3.x iOS practice screen with a pictured target (cheese) and word and phonetic display. Illustrative only; the exact build captured is unverified, and the screen may not represent every studied release.

Scoring versions

Speech recognition and acceptance scoring changed over the study period. Records from multiple eligible app versions were pooled without adjustment for app version or recognition/scoring changes.

Practice measures

Practice cards are eligible scored card events, not counts of individual speech attempts. Sessions are grouped by recorded session identifier within a profile, split after a gap exceeding 30 minutes, and ordered by their first eligible card. A session is assigned to a month by that start time; cards are assigned by their own timestamps. No selected session crossed the day-30 or day-60 boundary.

Practice days count distinct device-local dates with eligible cards, including the first observed practice date. All 17,675 cards in the selected cohort have recorded usable timezones. Totals of practice days are sums of profile-specific active dates, rather than unique dates for the cohort. Distinct sounds describe target coverage in each month; matched sounds are the intersection observed for that profile in both months. Reported age is taken from the first eligible card and is not independently verified.

Accuracy outcomes and weighting

The primary outcome, accepted-card accuracy, is the proportion of scored cards marked accepted by the app scoring system. A card was accepted if any evaluation during it succeeded, so acceptance could follow repeated utterances and auditory modeling. Individual production attempts and model replays were not counted. Accepted-card accuracy does not measure first-attempt accuracy, spontaneous naming or clinician-rated production.

For each matched profile-and-sound pair, the Month 1 accepted-card proportion is compared with the Month 2 proportion. The primary comparison requires at least one card in each month. Within a profile, matched sounds receive equal weight regardless of card count. The primary cohort mean then gives each of the 101 profiles equal weight. This same set of matched sounds is used in both periods for each profile. Different words, sound positions, games and attempts may nevertheless occur within a matched sound.

A secondary summary gives each profile-and-sound pair equal weight. Profiles contributing more matched sounds have more influence in that summary. Pair counts therefore represent correlated observations from 101 profiles, not 660 independent children. For a sound-specific mean, each profile practicing that sound in both periods contributes one pair.

Direction classifications and confidence intervals

Throughout this report, pp denotes percentage points.

Higher, lower and unchanged classifications use exact differences between the recorded proportions. Any positive difference is classified as higher, any negative difference as lower and an exact zero as unchanged. No minimum clinically meaningful change is imposed.

Nominal 95% percentile intervals for mean paired changes use 10,000 bootstrap resamples of whole profiles with seed 20261014. Resampling whole profiles preserves the matched month pairing and dependence between sounds contributed by the same profile; for the pair-weighted mean, each resampled profile retains its number of sounds. Subgroups use the same profile draws for both periods, and each between-group contrast resamples both groups at their observed sizes. Intervals are exploratory and uncorrected for multiple sound comparisons. Implementation details are given in Supplementary Methods S1.

Sound results with fewer than five matched profiles are suppressed. Additional sound information is withheld where necessary to prevent reconstruction from totals and other displayed cells: the displayed sound row with the smallest profile count is withheld under a fixed rule that does not depend on its result. Recorded-age counts use complementary suppression for the same reason.

Subgroup and sensitivity analyses

The primary estimate is the profile-weighted mean (Table 4). Pair-weighted means and direction counts are secondary summaries (Tables 4–7). Regular-user and first-month frequency comparisons are exploratory subgroup analyses; the five-card minimum and Day 1 analyses are sensitivity analyses (Supplementary Results S5–S7).

A regular-user subgroup contains 47 of the 101 primary profiles with at least ten total eligible sessions across the entire fixed source window. This definition can include sessions outside the first 60 days; it is not ten sessions in Month 1 or ten sessions on each sound.

The minimum-card sensitivity requires at least five cards per matched sound in each month. It retains 39 of the primary profiles and 171 pairs. The regular-user subset within that sensitivity contains 32 profiles and 156 pairs. The primary cohort remains the full 101 profiles; the higher minimum changes which profiles and targets qualify.

Exploratory first-month frequency comparisons

First-month practice was grouped post hoc at 1–4 versus 5+ and 1–9 versus 10+ sessions or distinct practice days. These overlapping partitions use the same profile-equal matched-sound monthly scores as the primary result. The cutoffs were specified after an earlier 1–5, 6–9 and 10+ session grouping, which is not reported here, rather than selected by searching for favorable outcomes. Practice days include the first observed date; session starts assign sessions to Month 1. Complete subgroup results appear in Supplementary Results S6, with the carried-over 10+ session interval explained in Supplementary Methods S2.

Day 1 sensitivity: matching across three periods

Day 1 is the first 24 elapsed hours after first observed eligible practice, [0,1) days. Month 1 remains [0,30) days and Month 2 [30,60). The sensitivity retains only sound targets observed in all three periods and uses those same sounds for each profile's three means. It includes 95 profiles and 432 profile-and-sound pairs. Cards were pooled within each sound and period, matched sounds were averaged within each profile, and profiles were weighted equally. Because the sample and sound set are narrower, its Month 1 and Month 2 scores differ from the primary 101-profile analysis and are not interchangeable with them. Day 1 contains observed practice, not a pre-practice baseline or necessarily first-ever use. The median Day 1 card count is two per matched sound; 327 of 432 pairs have fewer than five Day 1 cards.

This post hoc sensitivity was specified after the month-average results were reviewed. It uses the existing five- and ten-session/day cutoffs without a new threshold search and the same whole-profile bootstrap. The observations remain retrospective, unadjusted and uncorrected for multiple comparisons.

Results

Cohort and recorded ages

The eligible-record frame contained 949 profiles and 60,780 card records. Of these profiles, 277 lacked full 60-day observation opportunity, leaving 672. A further 559 had no eligible Month 2 practice and 12 had Month 2 practice without a matching sound target, leaving 101 profiles for the primary analysis (Figure 1). These sequential exclusions follow the card-eligibility filters in Methods; the 949 profiles are not all app registrants.

The primary sample contains 101 matched profiles from 101 family accounts. Together, they contribute 660 matched sound pairs and 17,675 eligible cards across the two periods. Accuracy comparisons use 11,902 of these cards: 7,600 in Month 1 and 4,302 in Month 2. The remaining cards concern sounds that are not matched for that profile across both periods and are included only in practice descriptions.

Table 2. Recorded ages among the 101 primary profiles.

Recorded age band	Profiles
2-5	55
6-8	33
9-12	Suppressed
Unknown or out of range	Suppressed

Age rows marked suppressed are not separately disclosed. Suppression of an additional age row prevents calculating the small cell from the cohort total.

Practice in Month 1 and Month 2

Observed practice amounts were lower in Month 2 for this returning cohort. Across the same 101 profiles, sessions totaled 680 in Month 1 and 347 in Month 2. Mean sessions per profile were 6.73 and 3.44, respectively. Median sessions declined from five to two. Mean practice days per profile declined from 4.70 to 2.71, and the median declined from four to one.

Profiles practiced a median of 15 distinct sounds in Month 1 and eight in Month 2. Each profile had a median of four matched sounds contributing to the accuracy comparison, with an interquartile range of two to nine. Declining target coverage does not change the targets used for its paired accuracy outcome, because only identical matched sounds are averaged in both periods.

Table 3. Practice amounts across the same 101 profiles. IQR denotes the 25th to 75th percentile.

Measure	Month 1 total	Month 1 mean per profile	Month 1 median (IQR)	Month 2 total	Month 2 mean per profile	Month 2 median (IQR)
All eligible cards	12,557	124.3	53 (20 to 150)	5,118	50.7	20 (8 to 49)
Sessions	680	6.7	5 (2 to 9)	347	3.4	2 (1 to 3)
Practice days	475	4.7	4 (2 to 6)	274	2.7	1 (1 to 3)
Distinct sounds	1,415	14.0	15 (8 to 21)	908	9.0	8 (4 to 13)
Matched sounds	660	6.5	4 (2 to 9)	660	6.5	4 (2 to 9)
Cards used in matched accuracy	7,600	75.2	23 (8 to 90)	4,302	42.6	16 (4 to 38)

Practice-day and distinct-sound totals sum profile-specific counts. The 660 matched-sound total represents 660 profile-and-sound pairs, with the same pairs in both months. All eligible cards and matched-accuracy cards have different denominators.

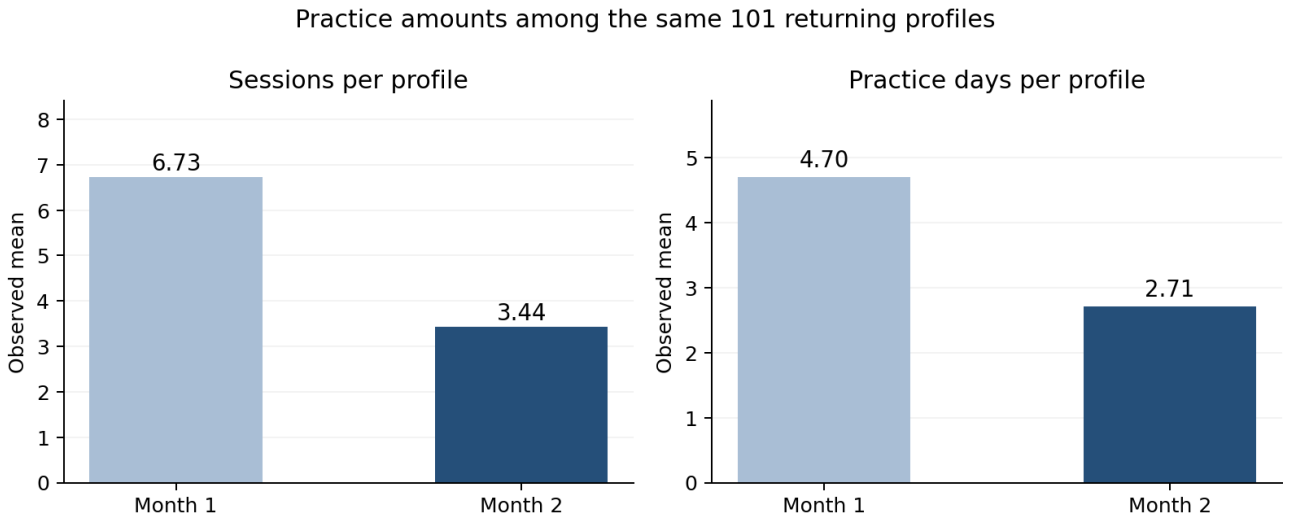


Figure 3. Mean observed sessions and practice days per profile across the two months.

Primary accuracy across matched sounds

The primary mean increased from 69.4% in Month 1 to 74.9% in Month 2: +5.5 percentage points (95% CI 0.4 to 10.4). The median change in a profile's mean was +2.0 points. Giving each sound pair equal weight produced a mean increase of +5.1 points; its median pair change was zero.

Table 4. Overall paired accuracy in the primary 101-profile cohort.

Weighting	Profiles	Sound pairs	Month 1 (%)	Month 2 (%)	Mean change, pp (95% CI)	Median change, pp
Each profile equally	101	660	69.4	74.9	+5.5 (0.4 to 10.4)	+2.0
Each sound pair equally	101	660	68.2	73.3	+5.1 (1.1 to 9.2)	+0.0

The different means reflect different weighting of profiles with many matched sounds. The profile-weighted estimate is the primary result.

Higher and lower individual scores

Of the 101 profiles, 54 (53.5%) had higher Month 2 averages, 34 (33.7%) had lower averages and 13 (12.9%) were unchanged. At the individual sound level, 278 of 660 comparisons (42.1%) increased, 177 (26.8%) decreased and 205 (31.1%) were unchanged. Among the 455 sound comparisons that changed, 61.1% increased. This percentage excludes the 205 unchanged pairs.

Table 5. Directions of observed paired change in the primary cohort.

Unit counted	Comparisons	Higher, n (%)	Lower, n (%)	Same, n (%)
Profile mean	101	54 (53.5)	34 (33.7)	13 (12.9)
Individual profile and sound	660	278 (42.1)	177 (26.8)	205 (31.1)

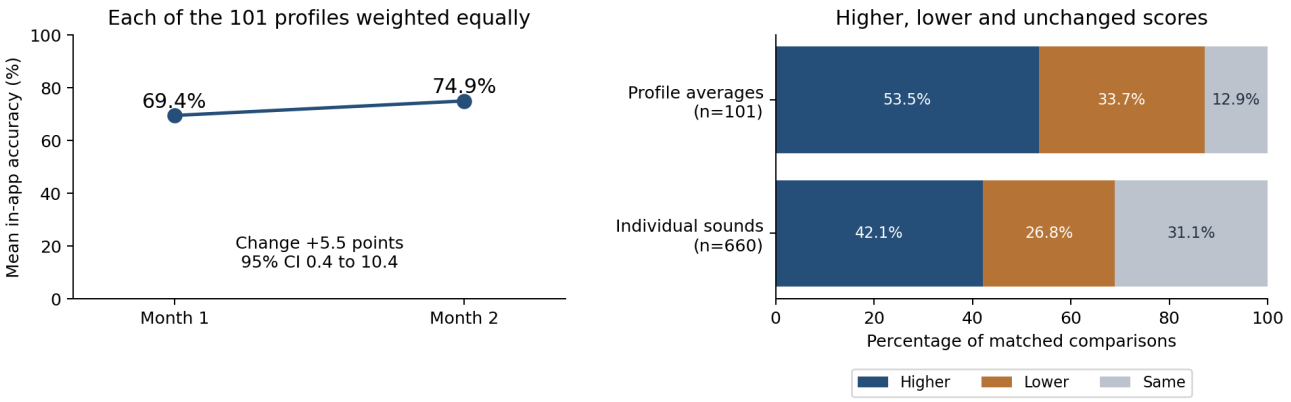


Figure 4. Primary profile-weighted means and distributions of paired change. Individual sound counts and profile-average counts describe different units.

Complete sound results

Sixteen of 21 displayed sound means were higher in Month 2 and 5 were lower. All 24 consonant targets appear in fixed phonetic order; unvoiced TH and ZH have fewer than five matched profiles and are suppressed, and Y /j/ is withheld under complementary suppression so that the suppressed counts cannot be recovered from the 660-pair total. These are directions of aggregate sound means, not the percentage of children improving and not 21 independent tests of treatment effects. No blend target qualified for the same-profile monthly matching in this cohort.

Table 6. Month 1 and Month 2 accuracy for every consonant in the 101-profile primary cohort. Card counts combine the two matched periods.

Sound	Matched profiles	Compared cards	Month 1 (%)	Month 2 (%)	Change, pp (95% CI)
/p/	40	820	72.7	67.6	-5.1 (-15.7 to 5.4)
/b/	40	895	71.3	76.1	+4.7 (-3.9 to 14.3)
/t/	38	700	64.7	75.8	+11.1 (1.4 to 21.0)
/d/	41	814	75.6	76.3	+0.6 (-10.8 to 11.5)
/k/	39	557	69.5	83.9	+14.4 (3.2 to 25.4)
/g/	26	384	60.7	70.9	+10.3 (-6.0 to 26.0)
/f/	24	317	61.5	54.2	-7.3 (-26.0 to 10.7)
/v/	21	202	52.1	62.5	+10.4 (-12.6 to 33.7)
TH /θ/ (thin)	Suppressed	Suppressed	Suppressed	Suppressed	Suppressed
TH /ð/ (this)	8	99	26.9	31.2	+4.3 (-14.2 to 27.7)
/s/	42	964	78.3	84.3	+5.9 (-1.0 to 13.5)
/z/	33	338	67.6	83.6	+16.0 (3.2 to 29.2)
SH /ʃ/	32	726	62.6	69.3	+6.7 (-8.8 to 22.2)
ZH /ʒ/ (measure)	Suppressed	Suppressed	Suppressed	Suppressed	Suppressed
CH /tʃ/	23	479	59.3	70.4	+11.1 (-2.6 to 25.9)
J /dʒ/	20	333	63.1	56.9	-6.2 (-25.3 to 12.8)
/m/	29	431	81.9	86.1	+4.2 (-7.6 to 15.8)
/n/	46	743	77.4	84.8	+7.4 (-0.3 to 15.6)
NG /ŋ/	23	225	62.0	73.2	+11.2 (-0.2 to 23.7)
/h/	21	204	89.7	87.0	-2.7 (-17.2 to 12.1)
/l/	38	939	60.3	67.4	+7.0 (-2.4 to 16.1)
R /ɹ/	39	1,421	57.9	54.6	-3.3 (-17.8 to 11.5)
/w/	23	134	77.8	82.4	+4.6 (-17.1 to 26.4)
Y /j/	Withheld	Withheld	Withheld	Withheld	Withheld

Values are rounded from the unrounded paired estimates. Confidence intervals are nominal; individual sound samples differ and overlap. Subtracting rounded means may differ slightly from the displayed change. Suppressed rows have fewer than five matched profiles; the withheld row has at least five and is hidden only to protect the suppressed counts.

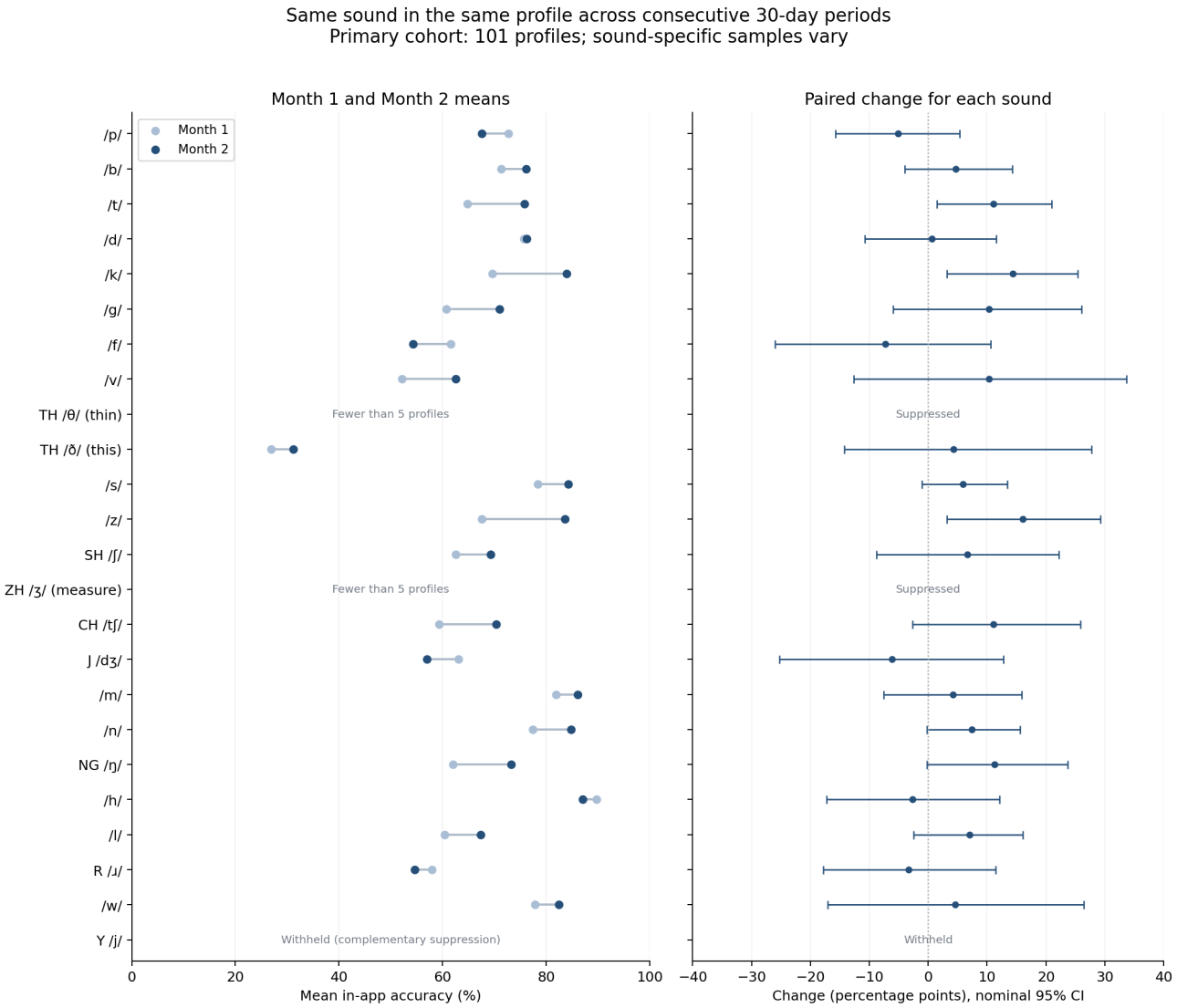


Figure 5. Mean accuracy and paired changes for all 24 sound rows. Sparse rows remain visible with suppressed estimates.

Table 7. Individual change directions for every consonant in the primary cohort.

Sound	Matched profiles	Higher	Lower	Same
/p/	40	18 (45.0%)	17 (42.5%)	5 (12.5%)
/b/	40	16 (40.0%)	14 (35.0%)	10 (25.0%)
/t/	38	17 (44.7%)	8 (21.1%)	13 (34.2%)
/d/	41	19 (46.3%)	10 (24.4%)	12 (29.3%)
/k/	39	23 (59.0%)	8 (20.5%)	8 (20.5%)
/g/	26	14 (53.8%)	6 (23.1%)	6 (23.1%)
/f/	24	8 (33.3%)	9 (37.5%)	7 (29.2%)
/v/	21	9 (42.9%)	7 (33.3%)	5 (23.8%)
TH /θ/ (thin)	Suppressed	Suppressed	Suppressed	Suppressed
TH /ð/ (this)	8	2 (25.0%)	4 (50.0%)	2 (25.0%)
/s/	42	19 (45.2%)	9 (21.4%)	14 (33.3%)
/z/	33	16 (48.5%)	4 (12.1%)	13 (39.4%)
SH /ʃ/	32	13 (40.6%)	6 (18.8%)	13 (40.6%)
ZH /ʒ/ (measure)	Suppressed	Suppressed	Suppressed	Suppressed
CH /tʃ/	23	10 (43.5%)	7 (30.4%)	6 (26.1%)
J /dʒ/	20	6 (30.0%)	7 (35.0%)	7 (35.0%)
/m/	29	9 (31.0%)	7 (24.1%)	13 (44.8%)
/n/	46	21 (45.7%)	7 (15.2%)	18 (39.1%)
NG /ŋ/	23	10 (43.5%)	3 (13.0%)	10 (43.5%)
/h/	21	3 (14.3%)	5 (23.8%)	13 (61.9%)
/l/	38	19 (50.0%)	12 (31.6%)	7 (18.4%)
R /ɹ/	39	15 (38.5%)	19 (48.7%)	5 (12.8%)
/w/	23	6 (26.1%)	4 (17.4%)	13 (56.5%)
Y /j/	Withheld	Withheld	Withheld	Withheld

Subgroup and sensitivity findings

The five-card sensitivity retained 39 profiles and 171 pairs, with a profile-weighted change of +6.3 percentage points (95% CI 1.4 to 11.3). Complete regular-user and five-card results appear in Supplementary Results S5. First-month frequency comparisons did not show a consistent pattern of greater gain with more practice; all four between-group gain intervals included zero (Supplementary Results S6). The Day 1 sensitivity retained 95 profiles and 432 pairs; its different sound set and sparse first-day scores limit comparisons with the primary analysis (Supplementary Results S7).

Discussion

In this selected cohort of 101 returning profiles, average accepted-card accuracy on identical sound targets increased from the first observed month to the second. More individual sound pairs increased than decreased, although almost one third were unchanged. Just over half of profiles had higher overall matched-sound means.

The sound results vary. Most reportable mean changes are positive, but several are negative, and the nominal intervals include substantial uncertainty for many targets. The zero median change among all 660 pairs also shows that the average gain and the typical individual sound comparison answer different questions.

Recorded practice amounts declined in Month 2 while paired accuracy increased. This does not imply that less practice produces higher scores. Cards, sessions and practice days are proxies for practice exposure rather than counts of production attempts, and exposure was not controlled. Higher later scores alongside less recorded practice could reflect retained skills, learning, target familiarity, selective return, changed task difficulty, outside practice, concurrent therapy, changes in speech recognition and acceptance scoring, or maturation, which these records cannot separate.

Speech production involves motor learning, and practice amount, distribution, variability and feedback schedule may affect retention and transfer (Maas et al., 2008). Speech Motor Chaining applies these principles through structured practice progressing from syllables to self-generated sentences and randomized practice (Preston et al., 2019). Evidence about intensity is intervention-specific: in a randomized study of preschool children receiving the multiple oppositions approach, treatment three times weekly produced greater phonological gains than treatment once weekly when cumulative intervention intensity was held constant (Allen, 2013).

Automated feedback has also been investigated in structured speech practice. In a five-participant single-case experiment involving older children and young adults, listener-rated rhoticity in practiced utterances was higher immediately after 30 minutes of automated Speech Motor

Chaining practice than immediately before it. Automated practice followed clinician-led prepractice, and agreement between the system and clinician judgments was evaluated (Benway & Preston, 2024). That finding does not validate SpeechLP's acceptance model, which was not compared with clinician ratings in this cohort.

Generalization requires assessment beyond practiced items. ASHA distinguishes single-word performance from connected speech. In a three-child home-delivered articulation study, transfer across tested settings exceeded transfer to untrained words (Costello & Bosler, 1976). Self-monitoring has been studied as a way to support generalization, although its relationship with conversational gains is variable (Koegel et al., 1986; Shriberg & Kwiatkowski, 1990). In a four-child Swedish single-case study of a self-supervised home speech app, slight improvement on trained words was observed in two participants, without improvement on untrained words (Strömbergsson et al., 2026). A meta-analysis found generalization to untreated sounds following complexity-based phonological treatment (Maggu et al., 2021); this does not establish transfer from every form of app practice.

The five-card sensitivity retains more rising than falling pairs, but substantially reduces the sample and changes the weighting-sensitive mean estimate. The Day 1 sensitivity produced a different pattern of group gains, but all four between-group intervals included zero, and many Day 1 scores rested on few cards. Some primary period scores are based on a single card and can take only zero or 100%, making individual differences noisy. Raising the card minimum improves measurement density while selecting different targets and users. Neither result replaces the other.

The present study measured acceptance on practiced app cards and did not assess untrained words, connected speech or communication outside the app, so it provides no evidence of carryover. Three lines of research would address these gaps. Scoring validation would compare app acceptance with blinded ratings from several SLPs across ages, sounds and speech characteristics, reporting uncertainty in the clinical reference itself (Meyer & Munson, 2021). Prospective comparisons with predefined outcomes, a comparison group and controlled practice exposure would test whether practice amount or feedback schedule affects learning. Generalization studies would pair in-app scores with clinician-rated probes of untrained words and connected speech and with functional measures such as the Intelligibility in Context Scale (McLeod et al., 2012).

Limitations

Selection. The sample requires recorded practice and matching sound coverage in both observed periods and is not a random sample of app users. Its results describe the analyzed profiles and do not generalize to all children starting the app or to profiles without observations in both periods. Full 60-day observation opportunity means the source window was long enough, not that every practice event inside or outside the app was recorded. Profile identity and age are derived from app records rather than independent child verification.

Outcome measurement. In-app acceptance can occur after repeated utterances or game retries. Individual production attempts, audio-model replays, additional prompting, clinical production quality and transfer to conversational speech were not independently measured. Matching a sound label does not standardize the word, sound position, game or difficulty. Mean changes can therefore reflect task and scoring differences alongside changes in performance.

Scoring validity and version changes. The app's acceptance model was not compared with SLP perceptual ratings for this cohort. Model-scored accuracy may differ from clinician-scored accuracy, including for distortions of sounds such as /ɹ/ and /s/. Speech recognition and acceptance scoring changed over time. Without app-version adjustment, these changes could have contributed to the observed Month 1-to-Month 2 difference alongside changes in participant performance. Their contribution was not quantified.

Development and maturation. Recorded ages include many children in a period of active consonant acquisition, including later-developing sounds (Crowe & McLeod, 2020). Without a comparison group, typical development over the observation period cannot be separated from practice-associated change.

Causal inference. The analysis contains no randomized comparison or untreated control group, and concurrent therapy and outside practice were not measured. Month 1 includes practice, so neither period is a pre-practice baseline. The regular-user and first-month frequency comparisons are observational; between-group differences in gain, including the +4.2-point difference favoring 5+ first-month sessions (95% CI −5.7 to 14.6), do not show that more practice caused greater improvement. Sounds and groups starting near the top of the scale have little room to rise, and groups selected or observed at a low first score tend to move toward the mean on remeasurement; both effects can produce differences in gains that are unrelated to practice. No causal model was fitted.

Multiplicity and sparse data. Intervals are exploratory and uncorrected for multiple sound comparisons, and small sound samples are suppressed. In the Day 1 sensitivity, matching across three periods narrows the profile sample and sound coverage, and many first-day sound scores rest on one or a few cards; the large ten-day gain comes from nine profiles starting at a lower Day 1 mean, and its direct between-group interval includes zero. These sensitivity results are not confirmatory and do not replace the primary month-average estimates.

Conclusion

In 101 returning profiles, accepted-card accuracy on identical sound targets was higher in the second 30-day period than the first (+5.5 percentage points; 95% CI 0.4 to 10.4). Change varied across profiles and sounds and, without validated scoring, adjustment for recognition/

scoring changes or a comparison group, cannot be attributed to practice or interpreted as clinical improvement. The defined cohort and complete sound results provide a reference for prospective evaluation.

A primary goal in publishing these findings is to seek university research partners to collaborate on future research.

Author contributions

Formal analysis: Shail Silver performed the initial analysis, with feedback from Matthew Siu and Jabali Ndungu.

Data and statistical review: Shail Silver, Matthew Siu and Jabali Ndungu reviewed the data analysis, statistical methods and numerical results.

Clinical review: Mari Gottlieb and Alissa Fox reviewed the clinical terminology, interpretation and relevance to speech-language pathology practice.

Funding

SpeechLP funded this study.

Competing interests

Shail Silver, Alissa Fox and Mari Gottlieb are cofounders of SpeechLP, which develops the app studied; Alissa Fox and Mari Gottlieb also serve as co-CEOs. Matthew Siu and Jabali Ndungu are software engineers with SpeechLP. All authors receive compensation from SpeechLP through equity, pay, or both.

AI assistance

Generative AI tools, including Claude and ChatGPT/Codex, assisted with manuscript drafting and editing, analysis-code preparation and computational reconstructions of the reported values from source records (Supplementary Methods S3). Exact model versions were not completely documented. The authors are responsible for the content and references.

Data availability

Aggregate results for Tables 1–7 and Supplementary Tables S1–S8, including all 24 sound rows with small-cell and complementary suppression, accompany this manuscript as aggregate-tables.csv, aggregate-tables.json, sound-results.csv and study-results.json. The last file includes unrounded estimates and distributions. These four aggregate data files are licensed under Creative Commons Attribution 4.0 International (CC BY 4.0; https://creativecommons.org/licenses/by/4.0/). Participant-level and source records are not publicly available because they contain account identifiers; analysis scripts are not publicly available. Repository DOI for this preprint: 10.5281/zenodo.23932485. Supplementary Methods describe the bootstrap implementation, the carried-over interval and computational reconstruction.

Research contact

Alissa Fox, M.S., CCC-SLP, SpeechLP: alissa@speechlp.com

References

Allen, M. M. (2013). Intervention efficacy and intensity for children with speech sound disorder. *Journal of Speech, Language, and Hearing Research, 56*(3), 865–877. https://doi.org/10.1044/1092-4388(2012/11-0076)

American Speech-Language-Hearing Association. (n.d.). *Speech sound disorders: Articulation and phonology* [Practice Portal]. Retrieved October 25, 2026, from https://www.asha.org/practice-portal/clinical-topics/articulation-and-phonology/

Benway, N. R., & Preston, J. L. (2024). Artificial intelligence–assisted speech therapy for /ɹ/: A single-case experimental study. *American Journal of Speech-Language Pathology, 33*(5), 2461–2486. https://doi.org/10.1044/2024_AJSLP-23-00448

Costello, J., & Bosler, S. (1976). Generalization and articulation instruction. *Journal of Speech and Hearing Disorders, 41*(3), 359–373. https://doi.org/10.1044/jshd.4103.359

Crowe, K., & McLeod, S. (2020). Children's English consonant acquisition in the United States: A review. *American Journal of Speech-Language Pathology, 29*(4), 2155–2169. https://doi.org/10.1044/2020_AJSLP-19-00168

Koegel, L. K., Koegel, R. L., & Costello Ingham, J. (1986). Programming rapid generalization of correct articulation through self-monitoring procedures. *Journal of Speech and Hearing Disorders, 51*(1), 24–32. https://doi.org/10.1044/jshd.5101.24

Maas, E., Robin, D. A., Austermann Hula, S. N., Freedman, S. E., Wulf, G., Ballard, K. J., & Schmidt, R. A. (2008). Principles of motor learning in treatment of motor speech disorders. *American Journal of Speech-Language Pathology, 17*(3), 277–298. https://doi.org/10.1044/1058-0360(2008/025)

Maggu, A. R., Kager, R., To, C. K. S., Kwan, J. S. K., & Wong, P. C. M. (2021). Effect of complexity on speech sound development: Evidence from meta-analysis review of treatment-based studies. *Frontiers in Psychology, 12*, 651900. https://doi.org/10.3389/fpsyg.2021.651900

McLeod, S., Harrison, L. J., & McCormack, J. (2012). The Intelligibility in Context Scale: Validity and reliability of a subjective rating measure. *Journal of Speech, Language, and Hearing Research, 55*(2), 648–656. https://doi.org/10.1044/1092-4388(2011/10-0130)

Meyer, M. K., & Munson, B. (2021). Clinical experience and categorical perception of children's speech. *International Journal of Language & Communication Disorders, 56*(2), 374–388. https://doi.org/10.1111/1460-6984.12610

Preston, J. L., Leece, M. C., & Storto, J. (2019). Tutorial: Speech motor chaining treatment for school-age children with speech sound disorders. *Language, Speech, and Hearing Services in Schools, 50*(3), 343–355. https://doi.org/10.1044/2018_LSHSS-18-0081

Shriberg, L. D., & Kwiatkowski, J. (1990). Self-monitoring and generalization in preschool speech-delayed children. *Language, Speech, and Hearing Services in Schools, 21*(3), 157–170. https://doi.org/10.1044/0161-1461.2103.157

Strömbergsson, S., Edlund, E., Pettersson, M., Phan, N., & Kurimo, M. (2026). Self-supervised app-based speech training for children with speech sound disorder: A single-case experimental design study. *International Journal of Language & Communication Disorders, 61*(1), e70163. https://doi.org/10.1111/1460-6984.70163