Sentence Reading Task Results

Andrés-Rojas (2022)

Overview

The purpose of this document is to provide a statistical analysis of the data of three participants segmented so far (Andrés Rojas, 2022). The primary research question investigated was whether there was a difference in rhotic production (both taps and trills) between a pre-test and post-test. The following analysis is for the 3 participants in the sentence reading task only. Overall, I took the following steps in the analysis:

1. I imported and tidyed the data. The scripts that I used is available in this github repo under `scripts/01_tidy_srt.R`. In addition to the script to tidy the data, I also wrote a function `run_t_test` to help run paired t-tests per participant that can be used later (and to more or less automate further analysis in this style). This function's source code can be found in this repo under `scripts/00_helpers.R`
2. I ran the scripts and analyzed the data to within this document, please see the results below, and made all analysis available in this Github repository (https://github.com/kparrish92/andres_rojas_consult).

Results

Overall, it looks like the training worked well for P3E, but not as much for the other two. Table 1 shows the mean duration of taps per participant during the pre and post tests, and Table 2 shows the same information for trills. As you explained, if training is effective, tap duration should be lower at post test. This was reflected in the means of both P1E and P3E, but not P5E. For trills, an increase in duration would suggest that the training was effective: this was only the case for P3E.

In addition to mean comparisons, I did a series of paired t-tests. This test determines whether a mean difference can, within a 95% confidence interval, not be equal to zero by taking into account the variation in the response durations submitted to the model. That is, if there was greater deviation in the overall sample, a mean difference between pre and post tests would more likely be due to chance than if the values were more consistent. In general, t-tests are reported using this format in APA:

t(df) = t value, p = p value

Df is degrees of freedom and reflects the total number of tokens -1 per comparison. This metric determines the shape sampling distribution. The t-value and p-value are typically used to assess statistical "significant", which refers to a non-zero difference between two objects of study in a frequentist framework. It's actually possible to calculate a p-value given the t-value and degrees of freedom. The rules of thumb here are that a p-value of less than .05 is considered to be good evidence that there is a non-zero difference between two objects of interest, and typically corresponds to a t-value of greater than 2.

In addition to this reporting, R also determines the mean difference between two objects of interest and generates a 95% confidence interval. This measurement can be helpful to report also, since the t-test alone

Table 1: Average duration of tap productions during pre and post-tests per participant

time	P1E	P3E	P5E	P6E	P7E	P8E
PRE	35.33	60.46	23.32	35.26	63.72	28.63
POST	29.21	24.46	24.45	21.86	23.40	23.51

Table 2: Average duration of trill productions during pre and post-tests per participant

time	P1E	P3E	P5E	P6E	P7E	P8E
PRE	98.41	38.81	30.67	56.59	53.92	64.36
POST	66.21	146.35	36.02	69.76	80.47	109.10

Table 3: Results of Paired t-tests of taps comparing duration during pre and post tests

Participant	df	p_val	t_val	estimate	ci_lo	ci_hi
P1E	66	0.019	-2.399	-6.149	-11.266	-1.032
P3E	58	0.000	-7.378	-36.102	-45.896	-26.307
P5E	60	0.246	1.171	1.590	-1.126	4.306
P6E	61	0.000	-7.514	-13.758	-17.419	-10.097
P7E	56	0.000	-17.909	-41.211	-45.820	-36.601
P8E	67	0.000	-4.138	-5.118	-7.586	-2.649

does not give information about a magnitude of an effect nor its directionality In this case, we are of course interested in the direction of the effect, and not just whether there is a difference between groups.

Having this in mind, take a look at the results of the first 3 participants. A total of 6 paired t-test were carried out, in which each participant's productions of the tap at pre-test and post-test were compared followed by their trill productions at both times (2 t-tests per participant - one for each segment).

First, table 3 shows the results of each t-test in which tap duration production is compared from pre to post test. The results suggest that the duration of taps by participant P1E fell by 6.15 (95% CI 1.03 - 11.27), as determined by a paired t-test ($t(66) = -2.399$, $p < .05$). The results of the same test procedure suggested a decrease of 36.1ms (95% CI 26.31 - 45.9; $t(58) = -7.378$, $p < .005$) in participant P3E, while participant P5E actually showed evidence of an increased duration in rhotic production for taps of 1.59 (95% CI -4.31 - 1.13), as determined by a paired t-test ($t(60) = 1.171$, $p < .246$). Importantly, the results of the paired t-test for P5E were inconclusive, suggesting that there was no evidence that training was effective for this participant. P6E's tap duration decreased by 13.76ms (95% CI 10.1 - 17.42), as determined by a paired t-test ($t(61) = -7.514$, $p < .005$).

New P7E's tap duration decreased by 41.21ms (95% CI 36.6 - 45.82), as determined by a paired t-test ($t(56) = -17.909$, $p < .005$).

P8E's tap duration decreased by 5.12ms (95% CI 2.65 - 7.59), as determined by a paired t-test ($t(67) = -4.138$, $p < .005$).

Table 4 shows an analogous analysis in trill production in which pre-test duration was compared to post-test duration. P1E duration fell by 31.02 (95% CI 19.01 - 43.03; $t(47) = -5.195$, $p < .005$). On the other hand, P3E showed evidence of an increase in duration of 109ms (95% CI 97.69 - 120.31; $t(46) = 19.398$, $p < .005$) while P5E's trill duration also increased from pre to post test by 5.83ms (95% CI -12.86 - -1.21), as determined

Table 4: Results of Paired t-tests of trills comparing duration during pre and post tests

Participant	df	p_val	t_val	estimate	ci_lo	ci_hi
P1E	47	0.000	-5.195	-31.021	-43.034	-19.007
P3E	46	0.000	19.398	109.000	97.689	120.311
P5E	45	0.102	1.667	5.826	-1.212	12.864
P6E	48	0.001	3.612	14.592	6.469	22.714
P7E	38	0.000	5.617	25.154	16.087	34.220
P8E	49	0.000	12.448	44.740	37.517	51.963

by a paired t-test ($t(45) = 1.667$, $p < .1$). Again, P5E did not show evidence of distinct performance in trill production from pre to post tests. P6E's trill duration increased from pre to post test by 14.59ms (95% CI 6.47 - 22.71), as determined by a paired t-test ($t(48) = 3.612$, $p < .005$).

New

P7E's trill duration increased from pre to post test by 25.15ms (95% CI 16.09 - 34.22), as determined by a paired t-test ($t(38) = 5.617$, $p < .005$).

P8E trill duration increased from pre to post test by 44.74ms (95% CI 37.52 - 51.96), as determined by a paired t-test ($t(49) = 12.448$, $p < .005$).

Plots

Here are a few boxplots by participant. These plots show the distribution of responses per participant in both taps (Figure 1) and trills (Figure 2) between the pre and post tests.

Figure 1: Productions of taps per participant during the pre and post test

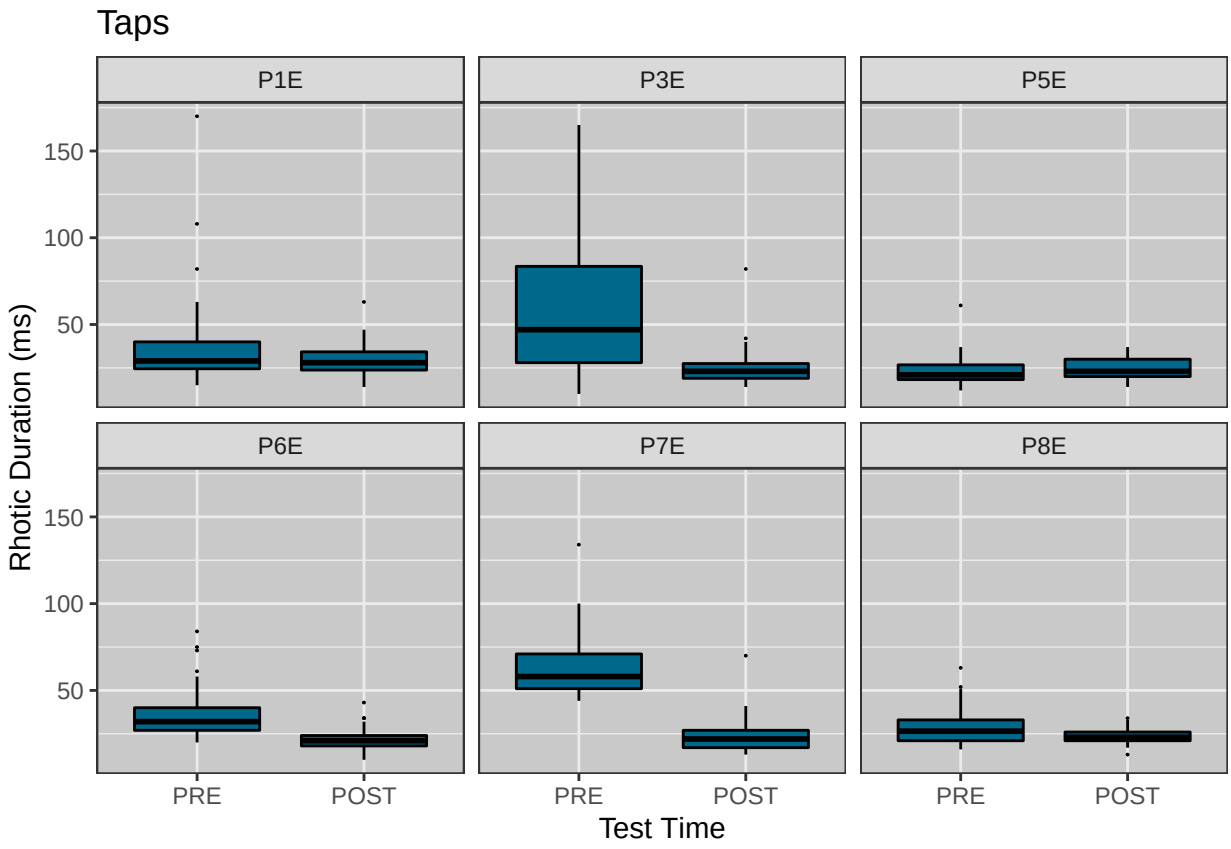
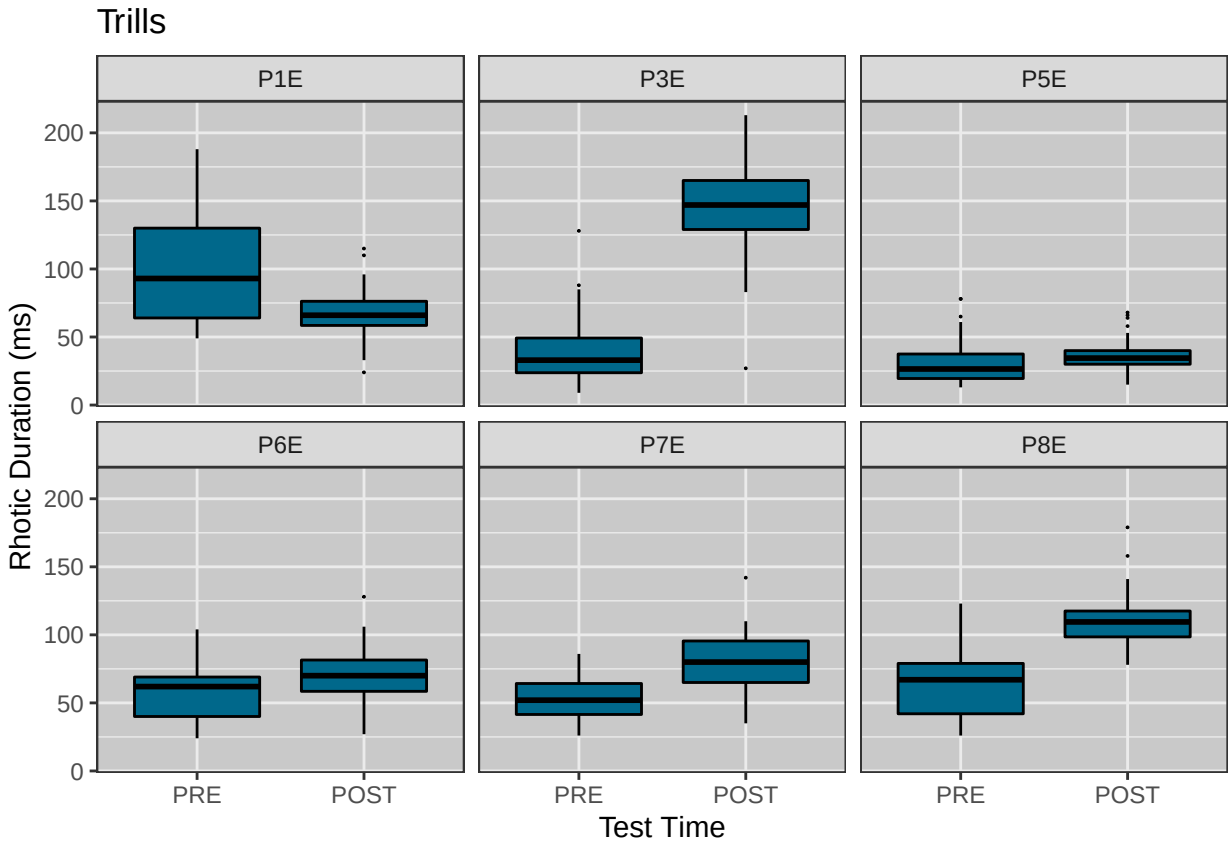


Figure 2: Productions of trills per participant during the pre and post test



Conclusions

It looks like there is good evidence that training worked for P3E, while P1E did show a decrease in duration for taps, but not an increase for trills, though this effect was small. P5E did not appear to respond to training, as pre and post tests were inconclusive. Overall, adding more participants should reveal whether there is a group trend and paint a clearer picture in terms of how much variation in the effectiveness of training exists.