Frontiers paper

Tara McAllister Byun and Heather Campbell

June 24, 2016

Individual results: Effect sizes

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

## [1] Brooklyn Connor   Jack     Madison
## 11 Levels: Adrian Brooklyn Connor Emily Emma Gabriel Hailey ... Madison
```

Effect sizes representing change in $\hat{p}_{correct}$ for the treated variant, vocalic /r/, are reported for all participants in Table 3. The first column shows participants' mean $\hat{p}_{correct}$ in the baseline period, averaged across all vocalic /r/ items from all sessions. The second column shows the equivalent mean across the three midpoint sessions, and the third shows the three maintenance sessions. The next three columns report three standardized effect sizes: ES_{Phase1} compares baseline versus midpoint scores, ES_{Phase2} compares midpoint versus maintenance scores, and ES_{all} compares baseline versus maintenance, reflecting overall gains across both phases of treatment. Participants are blocked by the order in which they received treatment (traditional-first or biofeedback-first). **The effect sizes in Table 3 show a wide range of variability in overall response to treatment across individuals. Averaging effect sizes across all participants yields a mean of ###4.61, suggesting that on average, participants' response to the combined biofeedback and traditional treatment package was positive and exceeded the minimum value considered clinically significant. Individual patterns of response will be examined in detail in the next section.**

	subject	bl1_m	bl1_sd	mp1_m	mp1_sd	mn1_m	mn1_sd	ESPhase1	ESPhase2	ESall
1	Adrian	45.65	8.08	86.97	3.00	77.96	4.43	6.78	-2.38	4.96
2	Brooklyn	24.34	1.69	25.61	1.92	26.00	3.43	0.70	0.14	0.62
4	Emily	3.18	0.81	6.21	1.13	28.19	4.52	3.09	6.68	7.71
7	Hailey	3.02	1.80	3.48	1.51	2.46	1.45	0.28	-0.69	-0.35
8	Hannah	4.51	1.81	3.65	1.17	4.88	1.29	-0.54	1.00	0.23
3	Connor	3.49	0.34	5.02	1.02	3.90	0.26	2.20	-1.50	1.33
5	Emma	65.81	7.17	92.88	3.61	86.87	2.97	4.35	-1.81	3.45
6	Gabriel	27.39	3.33	27.79	2.50	31.06	7.06	0.13	0.62	0.71
9	Jack	15.86	0.26	15.95	2.92	20.36	4.90	0.04	1.10	1.30
10	Liam	34.08	5.80	63.46	2.62	73.99	3.41	6.13	3.46	8.00
11	Madison	47.73	2.54	85.98	2.78	96.04	0.81	14.59	4.92	22.72

Individual results: Visual inspection

Figures 2-4 represent each participant's pattern of change in accuracy ($\hat{p}_{correct}$) over time, which can be visually inspected to corroborate the effect sizes reported in Table 3. Discussion of differences in relative response to biofeedback versus traditional treatment will be deferred until the next section, which reports the results of individual randomization tests. In the single-subject plots in Figures 2-4, each child is represented by

two boxes. The top box reflects performance on probe measures administered before and after each biofeedback treatment session and additionally reports probe scores from the pre-treatment baseline and post-treatment maintenance periods (shaded gray). The lower box reflects performance during traditional treatment sessions. Recall that sessions were randomly assigned to feature biofeedback or traditional treatment, but in a blocked fashion so that a participant experienced no more than two consecutive sessions of the same type. For each treatment session, a black circle represents performance on the pre-treatment probe measure, and a red star represents performance on the post-treatment probe. The y-axis represents $\hat{p}_{correct}$ aggregated across all vocalic /r/ items within a session; that is, the total number of "correct /r/" ratings as a percentage of the total number of ratings collected. Thus, the distance between the circle and the star within a session provide an index of the participant's progress (or lack thereof) during that treatment session. Finally, a dashed horizontal line tracks the participant's mean $\hat{p}_{correct}$ from the baseline interval, so that subsequent scores can be compared to the baseline mean.

The mean number of probe words on which $\hat{p}_{correct}$ scores are based was 17.54 (SD = 6.01) for pre- and post-treatment probes and ###NaN (SD = NA) for baseline and maintenance probes. The number of ratings collected in connection with a given probe session (i.e., the denominator in $\hat{p}_{correct}$) was roughly nine times the number of items in that probe and often was larger.

For convenience, the single-subject graphs have been grouped into three sets of 3-4 participants who demonstrated broadly similar patterns of response to treatment. Figure 2 depicts four participants for whom visual inspection provided strong evidence of a response to at least one type of treatment, and Figure 3 shows three participants for whom visual inspection provided moderate support of a response to treatment. Finally, Figure 3 shows four participants who showed no reliable evidence of a response to treatment. The participants in each group are discussed in detail below. **Comment on baseline stability: Which participants met the criterion of <10% mean session-to-session variability over the baseline interval? Which participants have demonstrations of noneffect in the baseline phase?** Those participants who met the criterion of <10% mean session-to-session variability (calculated as the baseline standard deviation divided by the baseline mean) were Brooklyn, Connor, Jack, Madison.

Visual inspection of data from 10;2 year-old Adrian shows a large change in level between the baseline phase and all subsequent phases of treatment, with minimal overlap between observations in the baseline phase and any subsequent phase. This change occurred immediately after the initiation of treatment. Adrian's progress in the first phase, which featured traditional articulatory treatment, yielded a large ES_{Phase1} of 6.78. His performance declined slightly in the second phase, which featured biofeedback treatment, yielding an ES_{Phase2} of -2.38 from midpoint to maintenance. However, his performance still remained substantially above baseline levels, with a final ES_{All} of 4.96. **Is unstable baseline a threat to internal validity?**

Emma, age 13;8, also showed a sizable change in level between the baseline phase and all subsequent phases of the study. This change was evident in post-treatment word probes within the first three sessions of treatment; from the fourth session on, there was virtually no overlap of data points with the baseline phase. The first phase, which featured biofeedback treatment, yielded an ES_{Phase1} of 4.35. Emma's gains remained mostly stable through the second phase of treatment and the maintenance period, yielding an ES_{All} of 3.45. Overlap with the baseline phase was minimal from the midpoint session on. **Is unstable/rising baseline a threat to validity?**

Participants Liam, age 9;6, and Madison, age 12;10, exhibited similar trajectories of progress, although Madison showed higher overall accuracy throughout the study. Like Emma, both boys began in the biofeedback treatment condition and began to show gains within the first three sessions of treatment; they sustained their gains during midpoint probes and made additional improvements in the second phase of treatment. They likewise showed no overlap between the baseline phase and data points in the midpoint phase, second phase of treatment, or maintenance phase. In Madison's case, there also was very little overlap between the baseline phase and the first phase of treatment. For both boys, a larger effect size was calculated for the first phase of treatment, in which biofeedback was provided, although this could be attributed to a ceiling effect in Madison's case. Large overall effect sizes of 8 and 22.72 were observed for Liam and Madison, respectively. **Demonstration of noneffect: trend in Liam's baseline data? What is going on with Madison's tx2 pre-probe? Might there be missing data?**

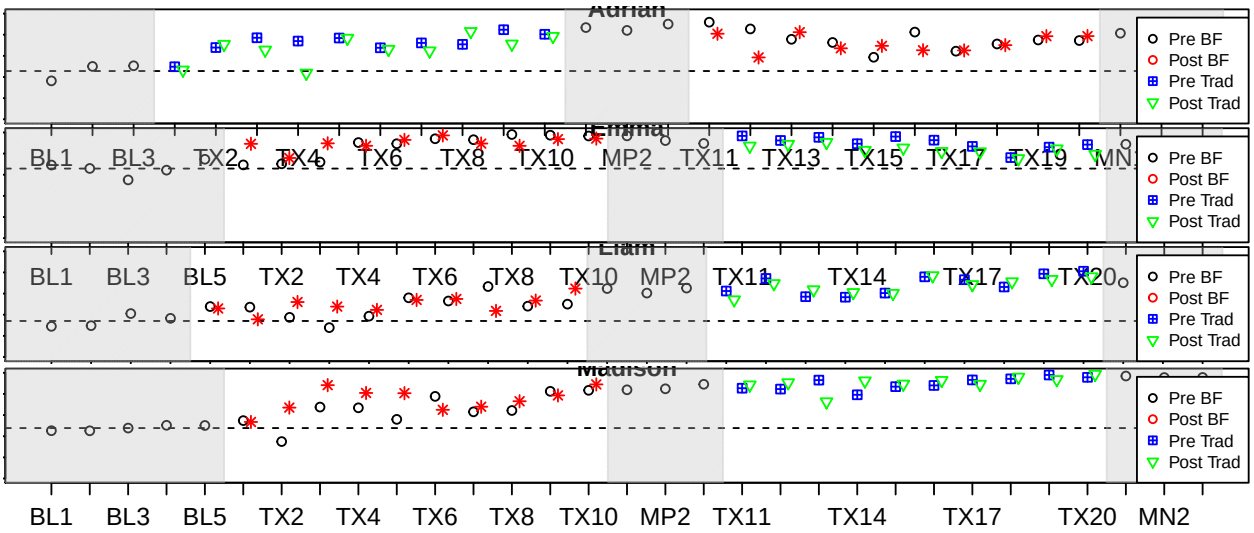


Figure 2. Longitudinal plots of $\hat{p}_{correct}$ for participants with large positive effect sizes. Dashed line represents mean across baseline sessions. BL = Baseline, Tx = Treatment, MN = Maintenance.

Figure 3 shows three participants for whom visual inspection offered moderate evidence of a response to at least one type of treatment. Emily, age 9;3, showed no visual evidence of improvement in the first phase, which featured traditional treatment. (An ES_{Phase1} of 3.09 was computed for her, but given that the raw difference in means was only 3.03, this was judged to be an artifact of low variance during the baseline phase.) During the second phase of treatment, Emily showed a change in trend from an unchanging near-zero level of accuracy to a small but consistent increase. This effect is ambiguous to interpret, since it emerged only in the last three sessions of the study and thus could reflect a cumulative effect of treatment over the preceding two months. Interestingly, the rising trend that Emily exhibited in her last three sessions of treatment continued into the maintenance phase, suggesting that ongoing generalization gains might be anticipated. Her final effect size (ES_{All}) of 7.71 reflected strong overall gains.

Participant Jack, age 15;10, showed a trajectory that was in some ways a mirror image of Emily's. Both visual inspection and effect size showed no meaningful change during the first phase of treatment, which featured biofeedback. He also showed no change throughout most of the second treatment phase, but in the final few sessions, his perceptually rated accuracy took on a distinct upward slope. Unlike Emily, his gains tended to affect post-test but not pre-test probes, and accuracy was not sustained throughout the maintenance period. This suggests more short-term learning that was not yet robustly transfering to other contexts. Accordingly, the effect sizes calculated using midpoint and maintenance probe data suggested a weaker effect of treatment than the within-treatment probes (ES_{Phase1}: 0.04; ES_{Phase2}: 1.1; ES_{All}: 1.3). **What's going on with Jack's TX1 pre probe?**

Participant Gabriel, age 9;10, showed a small increase in the perceptually rated accuracy of rhotics produced during Phase 1, followed by a slightly larger increase during Phase 2. Ongoing overlap in data points between baseline and treatment phases prevents us from attaching a strong interpretation to these data, and gains were minimally sustained into the midpoint and maintenance probe intervals. All three effect sizes computed for Gabriel fell short of the threshold to be considered clinically significant (ES_{Phase1}: 0.13; ES_{Phase2}: 0.62; ES_{All}: 0.71). On the other hand, Gabriel's accuracy during both phases of treatment was much more variable than at baseline. This suggested that he was engaging some degree of exploration of new strategies for rhotic production, although he had not yet stabilized a production pattern that would reliably yield a perceptually accurate rhotic sound.

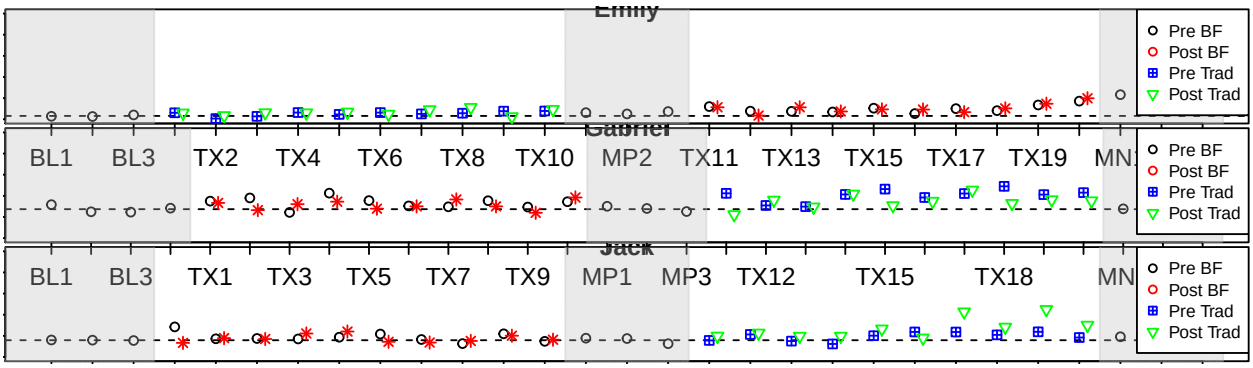


Figure 3. Longitudinal plots of $\hat{p}_{correct}$ for participants with small positive effect sizes. Dashed line represents mean across baseline sessions. BL = Baseline, Tx = Treatment, MN = Maintenance.

Finally, Figure 4 depicts the four participants for whom visual inspection of generalization probe data yielded no significant evidence of a response to either type of treatment. Three of these participants (Brooklyn, Hailey, and Hannah) began treatment in the traditional treatment condition, while one participant (Connor) received treatment in the opposite order. Across all of these participants, visual inspection of perceptual rating data yields a consistent picture of minimal change across all phases of the study. Effect sizes are generally in accordance with visual inspection; an exception is Connor's ES_{Phase1} of 2.2, but this value is inflated by minimal variance during the baseline phase. Note also that most participants in this group had near-zero perceptual accuracy ratings at baseline, contrasting with the more intermediate accuracy ratings given to participants in the other groups; we return to this topic in the discussion section.

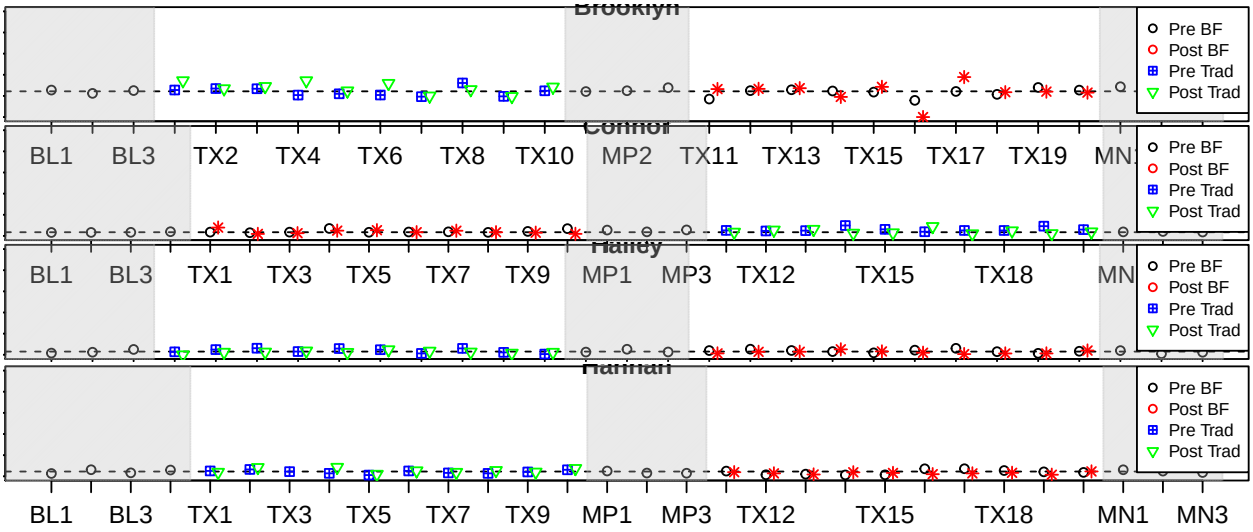


Figure 4. Longitudinal plots of $\hat{p}_{correct}$ for participants with null or negative effect size. Dashed line represents mean across baseline sessions. BL = Baseline, Tx = Treatment, MN = Maintenance.

Comparison of results across conditions: Effect sizes

Is there an order effect ($ES_{Phase1} > ES_{Phase2}$, or vice versa?)

-Boxplot depicting ES_{Phase1} across participants; boxplot depicting ES_{Phase2} across participants; boxplot depicting difference between ES_{Phase1} and ES_{Phase2} across participants. -Include t-test? (or leave hypothesis-testing to mixed model?)

Is there an effect of treatment condition ($ES_{BF} > ES_{TRAD}$, or vice versa?)

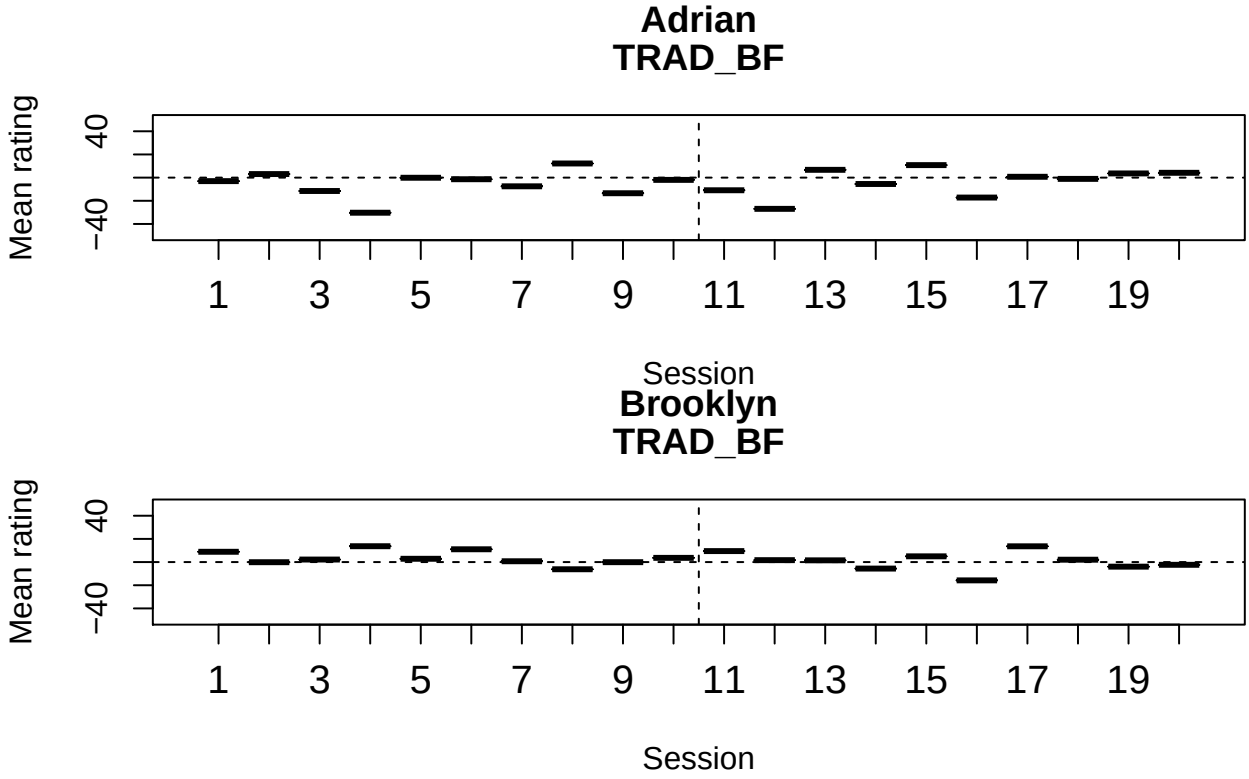
-Boxplot depicting ES_{BF} across participants; boxplot depicting ES_{TRAD} across participants; boxplot depicting difference between ES_{BF} and ES_{TRAD} across participants. -Include t-test?

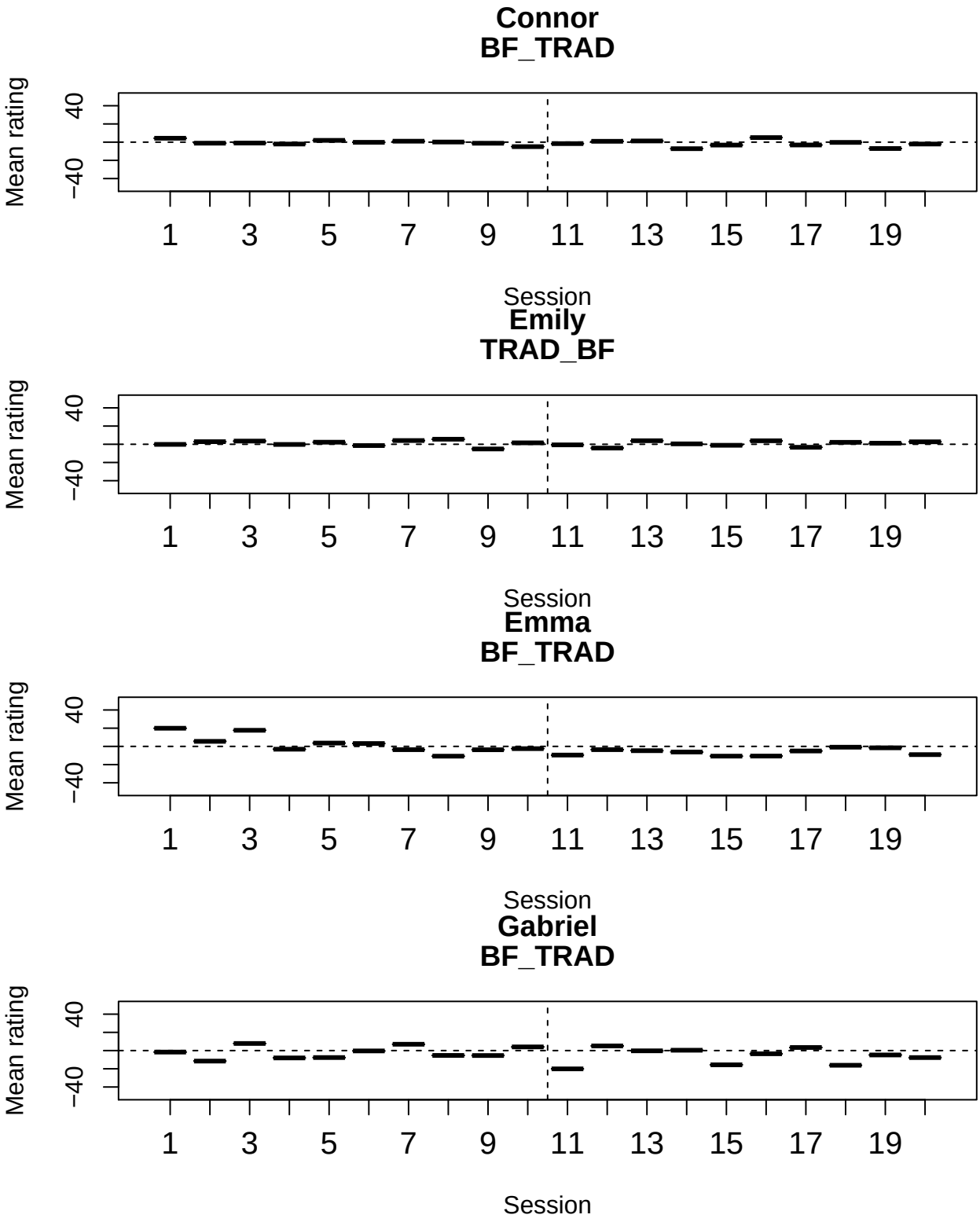
Does order of treatment delivery influence overall effect size (BF-TRAD > TRAD-BF, or vice versa?)

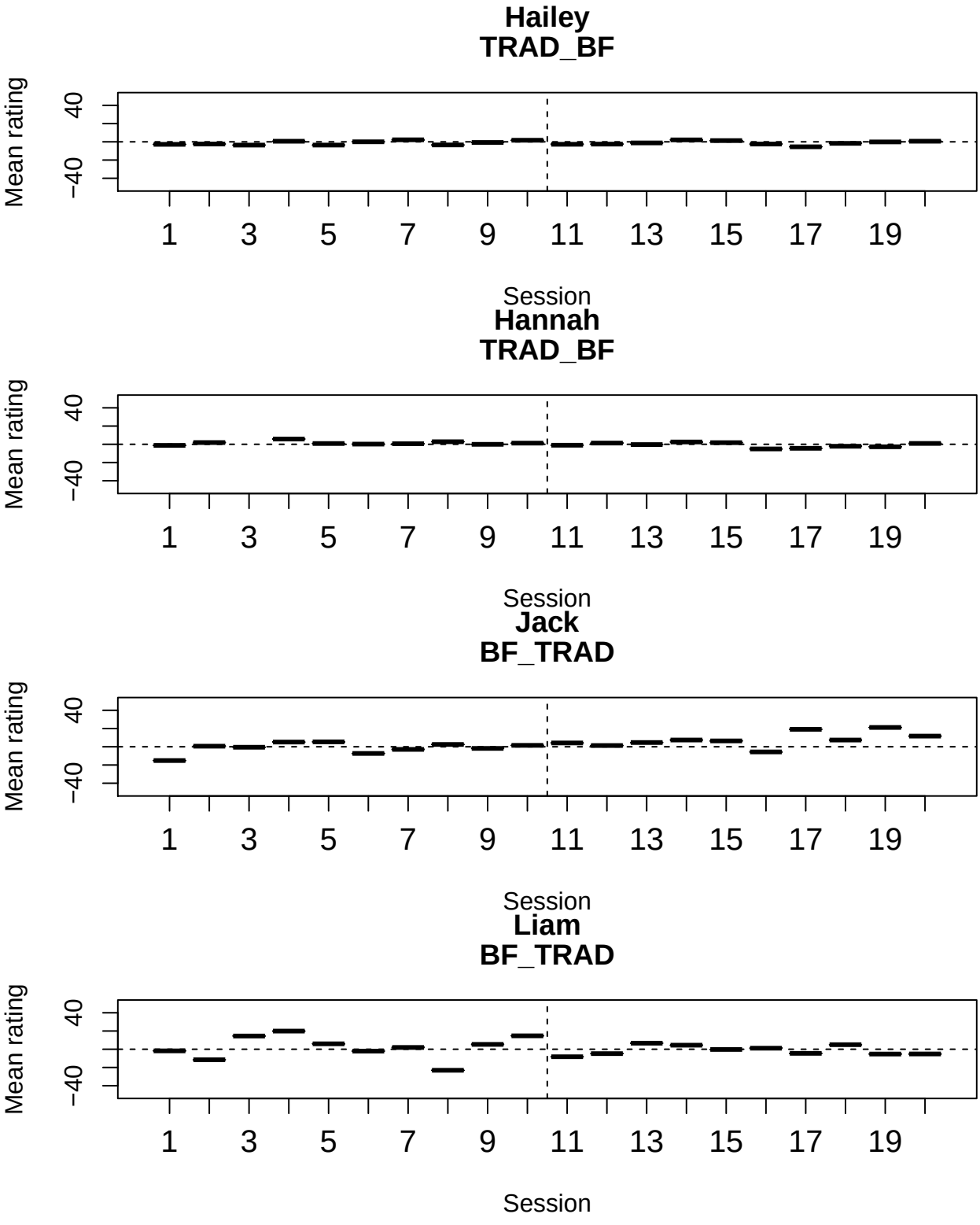
-Boxplot depicting ES_{All} across participants in BF-first condition; boxplot depicting ES_{All} across participants in TRAD-first condition -Include t-test?

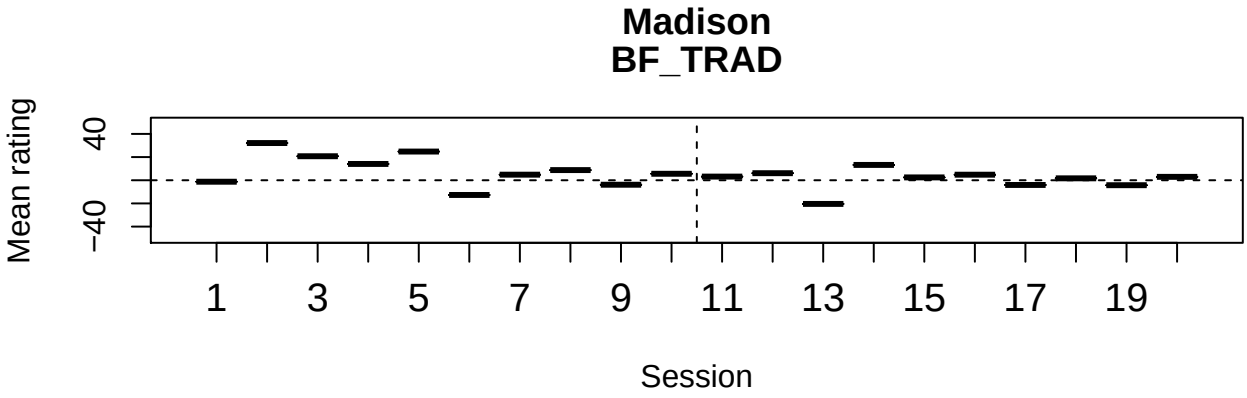
Comparison of results across conditions: Mixed logistic model

In the logistic mixed model described above, treatment condition (TRAD versus BF) was a significant predictor of /r/ production accuracy (($\beta =$, $SE =$, , $p =$), with the coefficient indicating that the reference level, traditional treatment, was associated with lower accuracy than post-treatment probes. Order of treatment delivery (biofeedback-first versus traditional-first) was not a significant predictor ($\beta =$, $SE =$, , $p =$), but the interaction between treatment type and treatment order was significant ($\beta =$, $SE =$, , $p =$). This interaction can be visualized in Figure 5, which plots accuracy in within-treatment probes (pre- and post-probes) during both biofeedback and traditional treatment conditions; they are pooled across subjects but partitioned by the order of treatment application (BF-first or TRAD-first). Figure 5 shows that accuracy scores tended to be higher in the BF-first order than the TRAD-first order, although this effect was not significant. Furthermore, the main effect of condition (TRAD > BF) appears to be driven heavily by the BF-first group. Complete regression results are reported in On-line Supplement A. FIX: Should use phase designation (phase 1 vs phase 2) instead phase order (BF_TRAD vs TRAD_BF)??




Connor
BF_TRAD
Mean rating
40
−40
1
3
5
7
9
11
13
15
17
19
Session
Emily
TRAD_BF
Mean rating
40
−40
1
3
5
7
9
11
13
15
17
19
Session
Emma
BF_TRAD
Mean rating
40
−40
1
3
5
7
9
11
13
15
17
19
Session
Gabriel
BF_TRAD
Mean rating
40
−40
1
3
5
7
9
11
13
15
17
19
Session


Hailey
TRAD_BF
Mean rating
40
−40
1
3
5
7
9
11
13
15
17
19
Session
Hannah
TRAD_BF
Mean rating
40
−40
1
3
5
7
9
11
13
15
17
19
Session
Jack
BF_TRAD
Mean rating
40
−40
1
3
5
7
9
11
13
15
17
19
Session
Liam
BF_TRAD
Mean rating
40
−40
1
3
5
7
9
11
13
15
17
19
Session



DISCUSSION

-Clearly there were both responders and non-responders. Can we identify individual characteristics that predict response to treatment? Comment on possibe subtypes/profiles. -Correlation between ES_{All} and age -Correlation between ES_{All} and baseline accuracy (participants who start out more accurate tend to gain more) -Correlation between ES_{All} and duration of previous treatment? -Correlation between ES_{All} and perceptual acuity?

-Regression yields main effect of treatment type (TRAD > BF) but also interaction between treatment type and treatment order. The majority of generalization gains observed seem to be associated with traditional treatment that followed biofeedback treatment. This could be a fluke (with the small number of subjects, random allocation of participants to TRAD-first vs BF-first condition could be skewing outcomes). On the other hand, this outcome is consistent with research and theories of motor learning.

-Patterns of relative response: No conclusive differences; examples of participants who responded to TRAD but not BF treatment (Jack, Adrian), to BF but not TRAD (Emily), or more to BF than to TRAD (Liam, Emma, Madison–but see ceiling effect)

-Discuss within-session gains versus generalization gains. Are these results consistent with the theoretically motivated claim that biofeedback should enhance initial acquisition of a target, while traditional treatment should promote generalization? May need to leave this for a subsequent paper. Comment on generalization and dose.