Online assessment and enhancement of auditory perception for speech sound errors

Protocol NCT04858035

Document generated 03/17/2021

4.3 Statistical Design and Power: Study 3

In the proposed multiple-baseline across-subjects design with randomization, participants' progress will be tracked continuously across baseline and treatment phases using percent accuracy on the two perceptual tasks described in the main proposal. Accuracy in baseline versus treatment phases of the study will be compared using Koehler and Levin's randomization test [101] as implemented in the R package SCRT [102], with visual inspection as a secondary analysis. Visual inspection will compare the baseline and treatment phases for each participant to identify any differences in level, slope, and variability. In addition, trends will be compared across participants to assess whether changes in performance are temporally linked to the introduction of treatment, to evaluate the immediacy of any such changes, and to examine consistency across participants [119]. The SCRT algorithm will evaluate whether the observed difference in accuracy between baseline and treatment phases is unlikely to occur ($p < .05$) relative to the distribution of differences obtained when a null effect is assumed and the labels “baseline” and “treatment” are randomly assigned in all permutations permitted by the specified multiple-baseline design. If there is a statistically significant effect of treatment, the observed difference will be extreme relative to this distribution. In particular, for a p-value of .05, we would expect to see no more than 5% of estimates from the simulated distribution with absolute values equal to or greater than our observed estimate. In addition to the randomization test, we will calculate standardized effect sizes for each participant (Busk & Serlin's d_2 [103]), comparing their performance in the pre-treatment baseline versus post-treatment maintenance phase and dividing by the pooled standard deviation. These will be used to evaluate, at both individual and group levels, whether the effect of treatment on perceptual accuracy exceeds 1.0, a benchmark commonly accepted to represent the minimum clinically significant effect size [99]. The same effect size measure will be used to compare production accuracy in the first three baseline sessions versus the three post-treatment maintenance phases, to test whether gains in perception transfer to production.

Power for the randomization test was estimated using the method proposed by Bouwmeester & Jongerling [111] and their associated app built with the Shiny package [112] in R [89]. Per their research, power in a multiple-baseline across-subjects randomization study is influenced by (a) the number of participants, (b) the within-subject effect size, (c) the number of observations per phase, (d) the level of autocorrelation between observations within participants, and (e) the range of possible start points for the intervention, and whether this range overlaps across participants. Power of 80% was observed in a calculation assuming (a) $n = 10$ participants, (b) a within-subject effect size of 1.0 (which we treat as the minimum clinically relevant difference as in [99]), (c) 15 observations per participant, (d) autocorrelation of .25, based on pilot data, (e) at least 4 possible start points, with overlapping start points across participants, and (f) additional characteristics of the baseline phase including unequal numbers of observations between baseline and intervention phases, correlation between scores across baseline and intervention phases, and greater variance in the intervention than the baseline phase.