**Comments/Questions on MEG Infant Data Analysis**

**Date:** 6/15/20

1. **Table 1**
   1. The age row could be split into 2M and 6M rows. Is head size also age dependent, i.e. measured at 2M or 6M? It could also be separated by age since there is a mixture of infants with 2M and 6M (21 with both) scans.
   * Values are for total sample of infants ranging in age BETWEEN 2-6 mos. age. I could split things up but it would just lead to visual clutter, thus I am showing demographics grouped by gender/sex.
   1. Not really a fan of adjusting the p-values for multiple comparisons since this is not done consistently
   * Please explain why MP corrections should not always be used?
   1. ~~What does the \* mean in the last column never to 0.079?~~
2. Histogram of age in file ***AUC\_55\_lbfgs\_N72\_EDA\_1.htm*** suggests 2M, 4M, and 6M scans
   * As per paradigm design specification children between 2-6 months were sampled. FWIW this is a product of initial grant writing to different funding sources on same resources.
3. Page 2
   1. ~~AUC looks “normal enough” in histograms to not need non-parametric tests. The TOST tests are parametric, so it would be more consistent to use all parametric for AUC~~
   * Good observation! I believe the slight rt. skew is the result of the shift in auditory sensitivity to speech sound contrasts we spoke about. My response, in general, is why not use the most suitable test for the question at hand, in this case one which makes no assumptions about normality?
   1. The 2M and 6M groups are not independent as 21 infants are in both. The same comparisons can be done using techniques that account for this overlap: generated estimating equations (GEE) and linear mixed models (LMM)
   * The non-independence in the 2- & 6-mos samples is why I opted to report LME4 results only for the repeated measures cohort. I ran GLM models on the total sample and got the same results as the mixed model, but with rather ugly AIC/BIC metrics. If geeglm provides more power in this situation I am willing to try. But since I’ve NEVER used it, and if there’s nothing fundamentally wrong with modeling for just n=21 cohort then I’d prefer the dumb way out. Unless you would like to do it? )))
     1. I tend to prefer GEE when the sample size is big enough because it makes fewer assumptions. In R, use the package *geepack* and function *geeglm*. The subject ID is provided via the *id* argument to indicate which observations (rows) below to the same subject
     2. For mixed models, the *lme4* package is the easiest to use with the *lmer* function. In the model formula, the string *+ (1|id)* would be added to indicate a random intercept per subject
     3. There is also the *nmle* package with *lme* for LMMs. In this case there is a separate formula to specific the random effect terms, in this case as *~ 1|id*.
4. **Table 2** 
   1. Since the underlying hypothesis is that there would be change with age, a null hypothesis of 0 makes more sense than the TOST approach, which would be more appropriate if you hypothesized no change
   2. I suggest calculating a single 95% CI for each age comparison per condition, which can be used to test the null of 0 change, as well as ± 0.5 SDs. However, doing the latter could cause problems, because if ± 0.5 SDs make sense as equivalence bounds (i.e., any change within this range is effectively no change for practical purposes), that suggests you need to see changes >0.5 SDs (excluded by the 95% CI) to conclude there is a meaningful change in age.
   3. The point estimates and CIs can be calculated using the GEE or LMM approach described above
   * Ugh…I don’t know how to do this stuff man! My understanding of non-equivalence TOST procedure as per Lakens article (<https://bit.ly/3hzYtxH>) is that’s analogous to / same as zero-mean-difference testing, where the interval can either be a priori specified from existing data CI or Cohen’s d values.
5. **Table 3**
   1. How is this “canonical correlation analysis (CCA)” as mentioned on page 3?
   * I stand corrected. It’s not CCA, it’s just the out put from paired correlations.
   1. Is M3L measured at one timepoint (which one) or multiple?
   * M3L and Vocabulary size are amongst the handful of behavioral correlate CDI measurements that folks like to work with. The CDI battery including M3L is collected at 3 mos. intervals between 18-30 months.
   1. Vocabulary size is measured at 18-30 months, so better to separate into each timepoint
   * Are you suggesting we go with 5 tables? That’s easy enough if necessary to show simple correlations at all, otherwise better go in supplementals.
   1. The results, particularly statistical significance, do not fully match between **Table 3** and the file **AUC\_55\_lbfgs\_N72\_CCA.htm**
   2. Some of the correlation coefficients marked as statistically significant are quite small. In **AUC\_55\_lbfgs\_N72\_CCA.htm** the N=288, much higher than the number of subjects. How were these calculated?
   * Looks like I botched this up copying values over from Jamovi to my word processor; my bad. The correlation coefficients were computed on data for 72 x 4 subjects-by-classification conditions, thus N=288.
6. **Table 4**
   1. ~~As with~~ **~~Table 1~~**~~, I don’t recommend the p-value adjustments.~~
   2. Either way, the statement “Kruskall-Wallis testing confirmed boys and girls were sampled from
   3. families with homogeneous demographic attributes” is not correct, as lack of statistical significance does not imply lack of a difference, or even an important difference, particularly with such small sample sizes
   * Fair enough, but I still need to communicate to some reviewers that any differences between boys and girls in this sample is negligible, how do you suggest going about doing that?
   1. What is ΔHead size?
   * Should be clarified as relative percent change in head size between 2-6months.
7. **Table 5**
   1. Is **Table 6** calculated using only on the N=21 with both 2M and 6M scans? The table heading says “paired” but the corresponding text on Page 5 says “independent samples”. It seems like **Table 2** is calculated using the independent samples t-test. A paired t-test approach would appropriately handle the repeated measures if restricted to the N=21 with both.
   * As per HTML descriptives for TOST procedures, both were computed using independent samples testing assuming unequal and equal variances for table 2 and 6, respectively.
   1. Conclusions cannot be drawn of non-significant results: from pages 5-6 “nonsignificant Welch t-test again ΔU…confirmed that sensitivity to phonological VOT change detection is indeed more nuanced by 6-months”
   * Again I would direct you to Lakens article (<https://bit.ly/3hzYtxH>) which defines non-equivalence as non-significant t-test (Welch) against either lower or upper limit around the null difference value, that is at least that’s my understanding. As such, TOST results are said to be reported by the larger of the two p-values for limits.
8. Modeling
   1. What is AUC here? Is it from the MMN or ERFdeviant condition?
   * In what you’re seeing, it is AUC data for all four classification conditions/treatments. I wasn’t sure if it’s kosher to pick and choose from amongst treatment levels, so I included all of them and designated a fixed effects term for condition with 4-levels.
   1. The models contain 14 subject-level predictors but are based on only 21 subjects. This should be scaled back substantially. Perhaps start with univariable models to get a feel for which predictors are most relevant so a small subset can be considered simultaneously
   * We’re going to have to discuss this b/c I cannot determine the rule of thumb for which and how many single variable(s) I’m expected to include.
   1. ~~I think you need to include follow-up time as another predictor (fixed effect) in the model (see h. below\_~~
   2. From my experience with Kubi, the vocabulary scores were pretty skewed, so I used the following transformation
      1. Arcsin(sqrt(vocabsize/679)) – the arcsin transformation, typically used for proportions
   * What is 679 here? Also the vocabulary scores here are not raw values but standardized scores, also most likely the case in the KUBI dataset unless you got your hands on the raw vocabulary scores.
     1. Also, we dropped the 30M observations as more and more infants approached the upper bound of the vocabulary list (i.e., saturation)
   * I’ve seen/heard this issue dealt with any different ways. By this logic, why include the 18M observations, which are clearly floor effects. I would argue, naively, that saturation at 30M is a reflection of the general inadequacy of the CDI test. There’s no way kids are at a ceiling for vocabulary size at this age.
   1. In Kubi the LMMs included random intercepts and random slopes
      1. To indicate this using *lme4*, use *lmer* with *+ (1|id) + ((tp-18)|id)*
      2. To indicate this using *nlme*, use *lme* with the argument *random=~1+(tp-18)|id*
      3. *tp* is the age in months of the CDI test, so *tp-18* goes from 0 to 12
   * I will have to confirm this, but I am pretty sure |id random term is implied because (i) the software ‘knows’ which column is the ‘id’ i.e. labeled/treated as special status variable, and (ii) AFEX package is used to wrap around LME4 mixed module and I don’t believe it would leave out that term. LME should error out.
   1. We also considered models with and without interactions with time. The model without the interaction assessed whether infants had consistently higher or lower vocab scores over time based on some predictor
   * I am not sure what time refers to here. I have CDI age, and also the subjects age in days between 2-6 months of age for MEG exams, to me that is a more interesting ‘time’ covariate. TBS I am not interested in ‘fishing’ for results, but if you know how to formulate the model, I’d be interested in seeing it.
   1. The model with an interaction with time assessed whether the trajectory of vocabulary growth (slope) got steeper or less steep based on some predictor. The ceiling inherent in the vocabulary score makes this type of model more difficult to use because the slope cannot be steep for a long time before the ceiling is reached, forcing the effective slope to be less steep
   * See previous response.
   1. The model specifications in the files **AUC\_55\_lbfgs\_M3l\_MLM.htm** and **AUC\_55\_lbfgs\_VOCAB\_MLM.htm** include the random effects specified by *+ (1|cdi\_age)*. That isn’t correct. There needs to at a minimum be a term *+(1|id)* to account for repeated measurements by subject (random intercept per subject). Age is better included in the model as a fixed-effect term rather than as a random effect
   * TBD c.f. response to sub-e above.