**Supplementary Materials**

 **S1. Explanation of outliers in MB-CDIs and Segmentation ERP difference score.**

The MB-CDI comprehension scores at both 12 months and 15 months were significantly skewed and contained several influential cases. The box and Q-Q plots for the two variables are shown below



**Figure S1.1.** Box and Q-Q plots for MB-CDI vocabulary comprehension score at 12 months.



**Figure S1.2.** Box and Q-Q plots for MB-CDI vocabulary comprehension score at 15 months.

Efforts to normalise the data proved futile. In every case, the outliers occurred at the high end of the distribution. For instance, in the case of the 15-month-old data, there are an unusually large number of data points (N = 7, 5 females, 2 males) outside of the interquartile range (and greater than 340 words). From the MB-CDI norms (Fenson et al., 2007), this means that 3 children were above the 99th percentile for their gender, and a further 4 were above the 95% percentile (although these are norms for children acquiring American English). Since we are dealing with a sample of 91 children at this age point, this shows that we have more positive outliers than would be expected were vocabulary comprehension to follow a normal distribution. Parental report runs the risk of introducing error in the estimation of children’s vocabulary due to parental differences in their understanding of what it means to know a word. Therefore we are cautious about these values because they are outside the normal range of the sample.

There were also two extreme values in our measure of segmentation, as shown in Figure S1.3.



**Figure S1.3.** Box and Q-Q plots for MB-CDI vocabulary comprehension score at 15 months.

Here there are two extreme values that are well outside of the interquartile range and which are noticeable deviations on the Q-Q plot.

Since we hypothesised a linear negative relationship between our segmentation measure and children’s vocabulary, the relationship is likely to be significantly affected by extreme values. Our non-parametric correlations indicated the existence of this negative relationship, but entering the variables into a multiple linear regression, which assumes normally distributed data, is problematic because the MB-CDI data are not normally distributed and in each case our variables of interest contain extreme values. Removing all extreme values is a possibility, but opens us up to claims of overfitting our data. However, it is valuable to consider how these extreme values do ultimately affect our analyses. Figure S1.4 shows the scatter plot between the measure of segmentation and children’s vocabulary comprehension at 15 months.



**Figure S1.4.** Scatter plot of children’s segmentation (*x*-axis) and vocabulary comprehension at 15 months (*y*-axis).

In Figure S1.4, the extreme values for the children’s vocabulary scores are in the solid-line ellipsis, and the extreme values for the segmentation score are in the dashed-line ellipsis. When these values are retained and entered into a multiple linear regression segmentation does not significantly contribute to vocabulary comprehension at 15-months over and above vocabulary at 9 months, session age at 9 months, SES, and gender (*β =* -1.57, se(*β*) = 1.05, *t* = -1.5, *p* = .14). However, when these extreme values are removed, segmentation is a significant predictor (*β =* -2.21, se(*β*) =.98, *t* = -2.25, *p* = .027). Therefore, without these extreme values, we obtain a very similar result to our binary logistic regressions.

**S2. Comparisons across ERP groupings on language, age, gender, SES and number of accepted trials in experiment.**

***Language Comprehension at 9 months.***

The parents of 5 of the 103 infants included in the final analysis did not return the MB-CDI. The average number of words the children in each group understood was: Negative responders (*M* = 35.58, *SD* = 34.73, N = 33), Positive Responders (*M* = 31.48, *SD =* 36.86, N = 31), and Intermediate responders (*M* = 33.48, *SD* = 32.42). The data were not normally distributed, and so were analysed using a non-parametric Kruskal-Wallis test. The results suggested no significant differences between the groups (χ2 = .743, df = 2, *p* = .69). We also checked to see whether group membership resulted in comparatively low or high vocabulary at 9 months; however, a chi-squared analysis showed that ERP group membership did not predict comparatively high or low vocabulary at 9 months (χ2 = .79, df = 2, *p* = .67).

***Language Production at 9 months.***

The average number of words the children in each group produced, as measured by the MB-CDI was: Negative responders (*M* = 1.3, *SD* = 1.7, N = 33), Positive Responders (*M* = 1.81, *SD =* 4.34, N = 31), and Intermediate responders (*M* = 1.71, *SD* = 2.58). The data were not normally distributed and so were analysed using a non-parametric Kruskal-Wallis test. The results suggested no significant differences between the groups (χ2 = .445, df = 2, *p* = .80).

***Age***

The average age of the children in each group was: Negative responders (*M* = 290.7 days, *SD* = 7.48), Positive responders (*M* = 291.8 days, *SD* = 7.05), Intermediate responders (*M* = 291.3 days, *SD* = 6.18). A 3-way univariate ANOVA comparing the three groups revealed no significant differences in their age [*F*(2, 100) = .47, *p* = .63, η2p = .009].

***Gender***

The distribution of gender in each group were as follows: Negative responders (Female = 17, Male = 17), Positive Responders (Female = 14, Male = 20), Neutral Responders (Female = 15, Male = 20). A Chi-Square analysis of frequencies revealed that this distribution did not significantly differ from chance (χ2 = 1.76, df = 2, *p* = .414, two-tailed test).

***SES***

Socioeconomic status was estimated from parental education levels, which were coded on a 6 point scale (0 = some High School, 1 = Finished High School, 2 = Trade/Vocational Certificate, 3 = Diploma, 4 = Bachelor Degree, 5 = Masters Degree, 6 = PhD). The distributional statistics are presented in Table S2.1

Table S2.1

*Distributional statistics for SES of Primary and Secondary Caregivers.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Mean (SD) | Median | Mode | Range |
| Primary Caregiver | 3.88 (1.1) | 4 | 4 | [0, 6] |
| Secondary Caregiver | 3.42 (1.5) | 4 | 4 | [0, 6] |

Kruskal-Wallis tests comparing the three segmentation groups on the education of their primary and secondary caregivers separately revealed no differences across the groups [Primary Caregivers: χ2 = .016, df = 2, *p* = .92; Secondary Caregivers: χ2 = .91, df = 2, *p* = .63]. It is important to note, however, that the Spearman rank order correlations between the primary and secondary caregivers’ education was significant but not overly high (ρ = .371, *p* < .001, N = 102). In our analyses we therefore combined the vales of the two for each infant to represent an overall SES value. Once again, there was no difference between the three segmentation groups on this combined score (χ2 = .423, df = 2, *p* = .81).

***Number of accepted trials.***

In our design, while the number of DF and nDF trails were equal (N = 120), nDF trails were divided into 60 IF and 60 UF trials. Since the groups differed in their response to the IF condition, with the Negative responder group showing a familiarity effect relative to the UF condition, it is important to determine whether there were differences in the number of accepted trials for each condition by group. Consistent with the greater number of DF trails, each group had a greater number of accepted DF trials than IF and UF: Negative responders (DF: *M* = 51.7, *SD* = 20.01; IF: *M* = 25.41, *SD* = 9.7; UF = 26.68, *SD* = 9.34), Positive responders (DF: *M* = 54.59, *SD* = 18.34; IF: *M* = 27.74, *SD* = 10; UF = 29.44, *SD* = 9.55), Neutral responders (DF: *M* = 57.71, *SD* = 19.45; IF: *M* = 28.94, *SD* = 9.63; UF = 30.11, *SD* = 9.53). A 3 (condition) X 3 (group) mixed ANOVA revealed a significant effect for condition [*F*(1.3, 129.6) = 506.83, *p* < .001, η2p = .835], which was driven by the fact that the number of accepted trials in the DF condition was overall significant higher than the IF and UF conditions (both *p*s < .001). The number of accepted trials in the UF condition (*M* = 28.74) was overall significant higher than the number of accepted trials in the IF condition (*M* = 27.36), *p* = .02. The main effect of group was not significant [F(2, 100) = 1.6, *p* = .35, η2p = .021], suggesting that overall the number of accepted trials in each group did not differ. The condition X group interaction was not significant [*F*(4, 200) = .46, *p* = .77, η2p = .009], suggesting that there was no difference across groups in the number of accepted trials across the three conditions.

***Language comprehension at 15 months.***

At 15 months there were 92 completed MB-CDI. The average number of words the children in each group understood was: Negative responders (*M* = 159.29, *SD* = 78.15, N = 31), Positive Responders (*M* = 138.93, *SD =* 93.92, N = 30), and Intermediate responders (*M* = 159.45, *SD* = 91.94). The data were not normally distributed, and so were analysed using a non-parametric Kruskal-Wallis test. The results suggested no significant differences between the groups (χ2 = 3.53, df = 2, *p* = .17). However, when we used group membership to predict high versus low vocabulary, we found a significant association (χ2 = 7.7, df = 2, *p* = .021, ϕ = .289, *r* = -.28, *p =* .007). Table S2.2 shows that membership Negative Responders group was associated with membership in the high vocab group at 15 months, membership in the Positive Responders groups was associated with membership in the low vocabulary group, and membership in the Intermediate Responders group showed no obvious preference.

Table S2.2

*Distribution of Responder Group By Vocabulary Group at 15-months.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  | Group |  |  |
|  | Negative | Intermediate | Positive | Total |
| High Vocab | 20 | 17 | 9 | 46 |
| Low Vocab | 11 | 14 | 21 | 46 |
| Total | 31 | 31 | 30 | 92 |

**S3. Replication of group-level analyses (with every other trial)**

To check the influence of different number of trials across the conditions (DF 120, IF and UF 60), trials of the DF conditions were exported such that odd trials and even trials were separated. Odd trials were used to calculate group membership while even trials were used in analyses. Those participants who had less than 10 trials accepted in either odd or even DF condition were excluded from this analysis (n=1). As such a total of 102 (52 male, 50 female) infants aged 9.5 months (mean = 291 days, range = 276-324) were included in these analyses.

A 3 (condition) x 4 (quadrant) way repeated-measures ANOVA showed that, as in the main analyses, there was no main effect of condition over the 250-500ms time window [*F*(2, 202) = .686, *p* = .505, η2p = .007], and nor was there a condition by quadrant interaction [*F*(4.22, 426.52) = .683, *p* = .612, η2p = .007]. To separate the whole cohort into three equal groups using the odd trials, those who had a negative going mean amplitude difference that exceeded -4.03µv were classified as a ‘negative responder’ (N=34), those who had a positive going mean amplitude difference that exceeded 2.29µv were classified as a ‘positive responder’ (N=34), and those who fell between these two parameters were classified as an ‘intermediate responder’ (N=34).

**Negative Responders.** A 3 (condition) x 4 (quadrant) way repeated-measures ANOVA revealed a significant main effect of condition [*F*(2, 66) = 5.45, *p* = .009, η2p = .142], but the condition by quadrant interaction was not significant [*F*(3.81, 125.76) = 1.58, *p* = .19, η2p = .046]. Bonferroni post hoc comparisons showed that the DF condition was significantly more negative than the UF condition (*p* = .002), and the IF condition was marginally more negative than the UF condition (*p* = .072). The DF and IF conditions did not differ.

Therefore, the main finding of a relative negativity for familiar versus unfamiliar words replicates. If we simply compare mean amplitudes in the left frontal quadrant we find the same result: DF and UF are more negative than the UF condition [DF versus UF: *t*(33) = -5.65, *p* < .001; IF versus UF: *t*(33) = -3.00, *p* = .005], whereas the DF and IF do not differ [*t*(33) = -1.4, *p* = .17].

**Positive Responders.** A 3 (condition) x 4 (quadrant) way repeated-measures ANOVA revealed a marginal main effect of condition [*F*(2, 66) = 2.56, *p* = .085, η2p = .072], and a non-significant condition by quadrant interaction was not significant [*F*(3.64, 120.13) = 1.17, *p* = .33, η2p = .034]. Bonferroni post hoc comparisons showed that the DF condition was marginally more positive than the UF condition (*p* = .071). No other contrasts were significant.

Once again, if we compare the conditions over the left frontal region, we find that the DF condition is significantly more positive than the UF condition [DF versus UF: *t*(33) = 2.3, *p* = .029]. The IF condition was also significantly more positive than the UF condition [*t*(33) = 2.28, *p* = .029]. The DF and IF conditions did not differ [*t*(33) = .102, *p* = .92]. Overall, the finding of a relative positivity for familiar versus unfamiliar words replicates.

**Intermediate Responders.** A 3 (condition) x 4 (quadrant) way repeated-measures ANOVA revealed no main effect of condition and no condition X region interaction (*F*s < 1). No comparisons over the left frontal region were significant.

Overall, these additional analyses reveal consistency in the results when we use a subset of the DF trails to create group membership. Any differences are largely due to the reduced power associated with not using half of our DF trials. As one final source of evidence, we reported the split half reliability of the odd and even DF trials relative to the UF trials. The results showed there was considerable consistency (Spearman-Brown coefficient = .71, Guttman Split-Half = .69).