Supplementary materials — "Divergence point analyses of visual world data: applications to bilingual research"

S1. Demographic profiles of the language groups

Table S1. Demographic profiles of L1 German, L1 Spanish and L1 English participants. Standard deviations are shown in parentheses. Proficiency was measured with self-ratings and with the Goethe Institute Placement Test (Goethe Institute, 2010). By-participant self-ratings averaged across the four skills were significantly correlated with Goethe scores (r = 0.59, p < 0.05). Values in parentheses show standard deviations.

	L1 German (n = 74)	*L1 Spanish (n = 48)*	*L1 English (n = 48)*
Mean age [years]	25 (6)	31 (6)	32 (8)
Female participants [count]	47	33	29
Right-handed participants [count]	67	45	42
Mean age of German acquisition [years]	-	21 (8)	19 (6)
Mean Goethe score [%]	-	68 (16)	71 (14)
Mean German self-rated proficiency [%]	-	73 (6)	72 (7)
Listening	-	*79 (13)*	*77 (11)*
Speaking	-	*77 (12)*	*77 (15)*
Reading	-	*71 (13)*	*69 (14)*
Writing	-	*65 (15)*	*64 (14)*

S2. Comparison of different statistical tests in the bootstrap approach

The choice of the statistical test within the bootstrap should not affect its validity, since we ultimately base our inferences on the bootstrap distribution, rather than on the individual tests. However, the choice of test may influence the mean and/or variance of the bootstrap distribution, thus affecting the location of the divergence point and/or the width of its confidence interval. To investigate this possibility, we evaluated the bootstrap procedure using different tests.

The first test was a logistic generalized linear mixed-effects model (GLMM), which appropriately accounted for the binomial nature of the data. Given that our data include multiple observations per participant and item, we used the maximal random-effects structure supported by the data by including varying intercepts for participants and items (Barr et al., 2013; Matuschek et al., 2017; see **code §S2**). To assess the role of the random effects (i.e. to assess whether a simpler model would yield similar estimates), we also tested a logistic model without a random effects structure (GLM).

We also evaluated two simpler tests. The first was a one-sample t-test of fixation proportions, as described in the manuscript. In non-parametric approaches, t-tests are often employed because they are conceptually straightforward and computationally light (e.g. Groppe et al., 2011; Maris & Oostenveld, 2007; Efron & Tibshirani,1986; Hesterberg, 2015; Reingold & Sheridan, 2014). When a dataset does not contain extreme mean fixation proportions (e.g. clustered around 0% or 100%), a paired or one-sample t-test is a reasonable approximation of a logistic model (see Gelman & Hill, 2006). In prediction studies such as ours, which measure fixations prior to the appearance of a target word, participants are unlikely to be fully certain about the identity of the target and thus the data will not contain a large number of extreme values. In addition, since we were only interested in the difference between the target and competitor, fixation proportions at the divergence point between these two regions were clustered around 50%, where a linear model (and thus a paired or one-sample t-test) most closely approximates a logistic model.

Our second t-value approach was a linear model of weighted empirical logits (Barr, 2008; Veríssimo & Clahsen, 2014). This test was chosen because the empirical logit transform better reflects the binary nature of the data. Following Agresti (2002), empirical logits were computed by taking, for each participant at each timepoint, the log of the number of times the target was fixated across items (y) divided by the total remaining number of looks to either the target or competitor ($n - y$). Each of these terms was adjusted by 0.5 to avoid infinite values due to taking the log of zero or $y/0$:

$$elogit_{target} = log(\frac{y_{target} + 0.5}{n_{target} - y_{target} + 0.5})$$

We added weights to the model to allow cases with variance closer to zero to inform the statistical model more than cases with higher variance, which means that logits representing proportions near 0% or 100% are less informative than those near 50% (Barr, 2008; McCullagh & Nelder, 1989). Upweighting cases with lower variance also means that the by-item information lost through aggregation is to some extent encoded in the model, because the number of observations in each empirical logit will also reduce that logit's variance. Weights were added by taking the inverse of the variance for each empirical logit, in line with McCullagh and Nelder (1989) and Gart and Zweifel (1967):

$$weight_{target} = \frac{1}{v_{target}}, \text{ where } v_{target} = \frac{1}{y_{target} + 0.5} + \frac{1}{n_{target} - y_{target} + 0.5}$$

Figure S2 presents a comparison between tests. As can be seen, the choice of test does affect the divergence point estimates and their confidence intervals, although these are mostly consistent across tests. While these results justify the use of t-tests on fixation proportions in the main manuscript, researchers should consider which test is most appropriate for their own data.

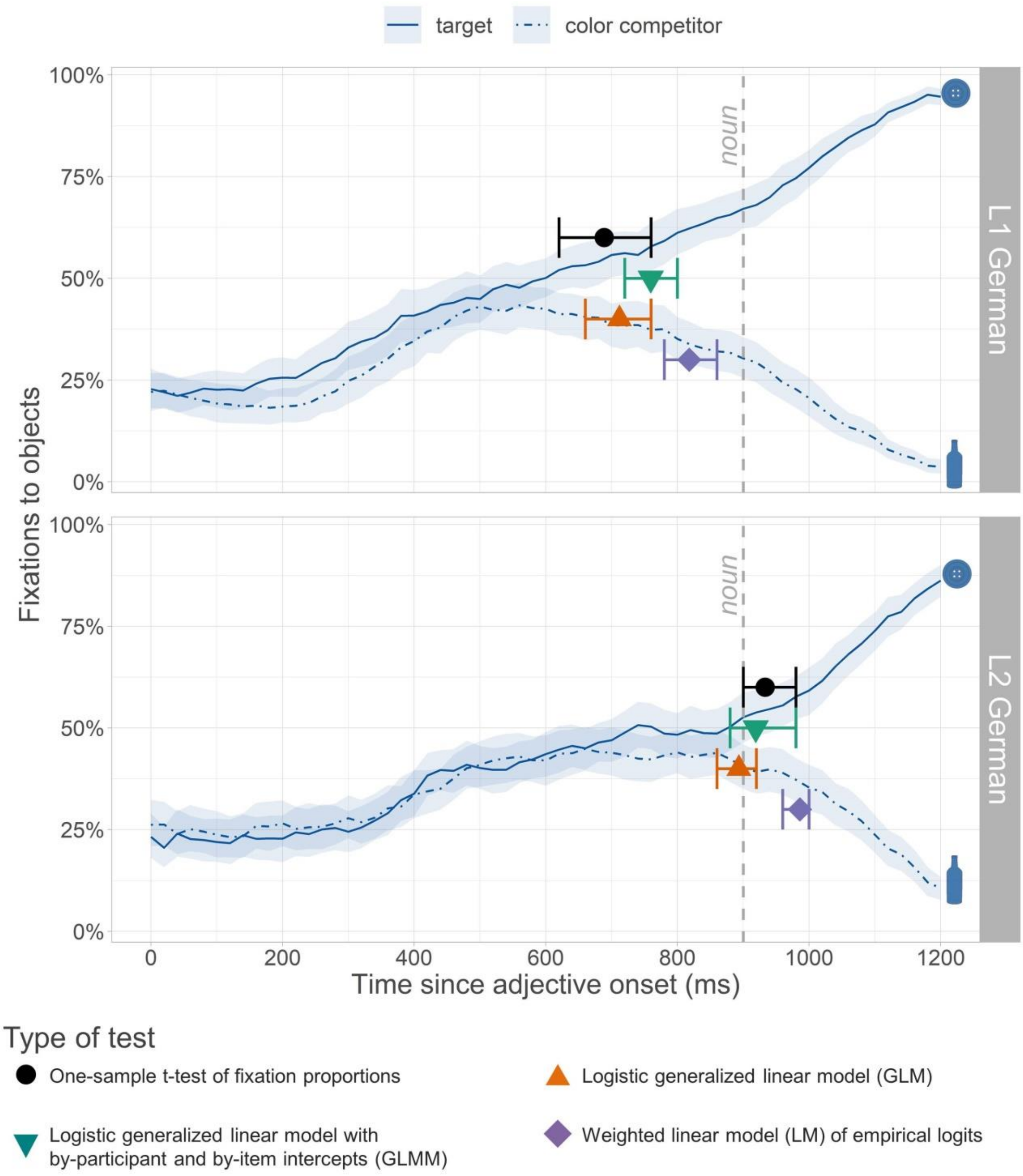

Figure S2. Results of four different tests comparing target vs. competitor fixations in the bootstrapping procedure. The black-colored onsets from the one-sample t-test of fixation proportions correspond to the onsets presented in the main manuscript. The t-test yields onsets that are similar to those of the GLMM (green), which is the test that most appropriately models the binomial nature and nested structure of the data. The GLM (orange) without random effects structure yields a different estimate to the GLMM, illustrating how not accounting for nested variance can affect the result of a logistic model.

S3. Comparison of bootstrap- and GAMM-derived divergence points

As discussed in the main article, there are other methods that allow the estimation of divergence points. One of these methods includes generalized additive mixed models (GAMMs). In this section, we compare the divergence points estimated via a GAMM with the divergence points estimated via the bootstrapping approach in the native (L1) vs. non-native (L2) speakers of German. The code for the GAMMs can be found in section **§S3** of our script. For each group, we fit a GAMM to the binomial outcome of fixations to the target vs. competitor, with a fixed non-linear predictor for time and an interaction term to model the trend over time for the factor “region of interest” (target/competitor). We used the maximal random effects structure supported by the data, i.e. an optimal model (Bates et al., 2018; Matuschek et al., 2017).

The model-predicted differences between fixations to the target vs. competitor over time are plotted together with 95% confidence intervals in **Figure S3A**. To determine the divergence point of looks to the target vs. competitor, we found the first time point at which the lower bound of the 95% confidence interval is greater than zero. Because we are interested in the beginning of a *sustained* preference for the target, we specified that the lower bound of the confidence interval had to remain above zero for at least 200 ms. This is equivalent to the 10 consecutive test statistics criterion used in the bootstrap, i.e. given our 20 bins, 10 consecutive timepoints correspond to a 200 ms time period (this criterion was adapted from Sheridan and Reingold, 2012; Reingold & Sheridan, 2014). The onset point for each language group is indicated with red dashed lines in **Figure S3A**.

Panel B in **Figure S3** shows that the GAMM-derived divergence points are earlier than the mean bootstrap-derived divergence points, but they fall close to the lower bound of the 95% confidence interval derived via the bootstrap. The difference between divergence points may stem from the fact that the GAMM confidence intervals are continuous. This means that estimates may correspond to timepoints that do not necessarily appear in the original data. For example, the GAMM-derived onset estimate for the L1 German group is 629 ms, but there are only fixation samples at 620 ms and 640 ms in the original data. By contrast, the bootstrap approach only uses the timestamps provided in the original data. This is not necessarily a weakness of either approach, but a potential source of differences between them. It should also be noted that our method for extracting divergence point estimates from the GAMM difference curves may be too simplistic (see code **§S3**). It is provided here to allow for a preliminary comparison between methods.

An important difference between the estimates from both methods is that the bootstrap-derived estimate additionally provides a measure of the temporal uncertainty associated with the onset time, expressed as a 95% confidence interval. Estimating this uncertainty means that we have an estimate of the sampling distribution for each divergence point, which enables statistical inference. These inferences include assessing whether a divergence point reflects a predictive effect (i.e. its confidence interval is earlier than the onset of the noun corresponding with the target object), or whether the divergence points differ between the language groups. This information cannot be obtained from fitting a GAMM once. To produce a measure of uncertainty for the GAMM-derived onset, bootstrapping could be used.

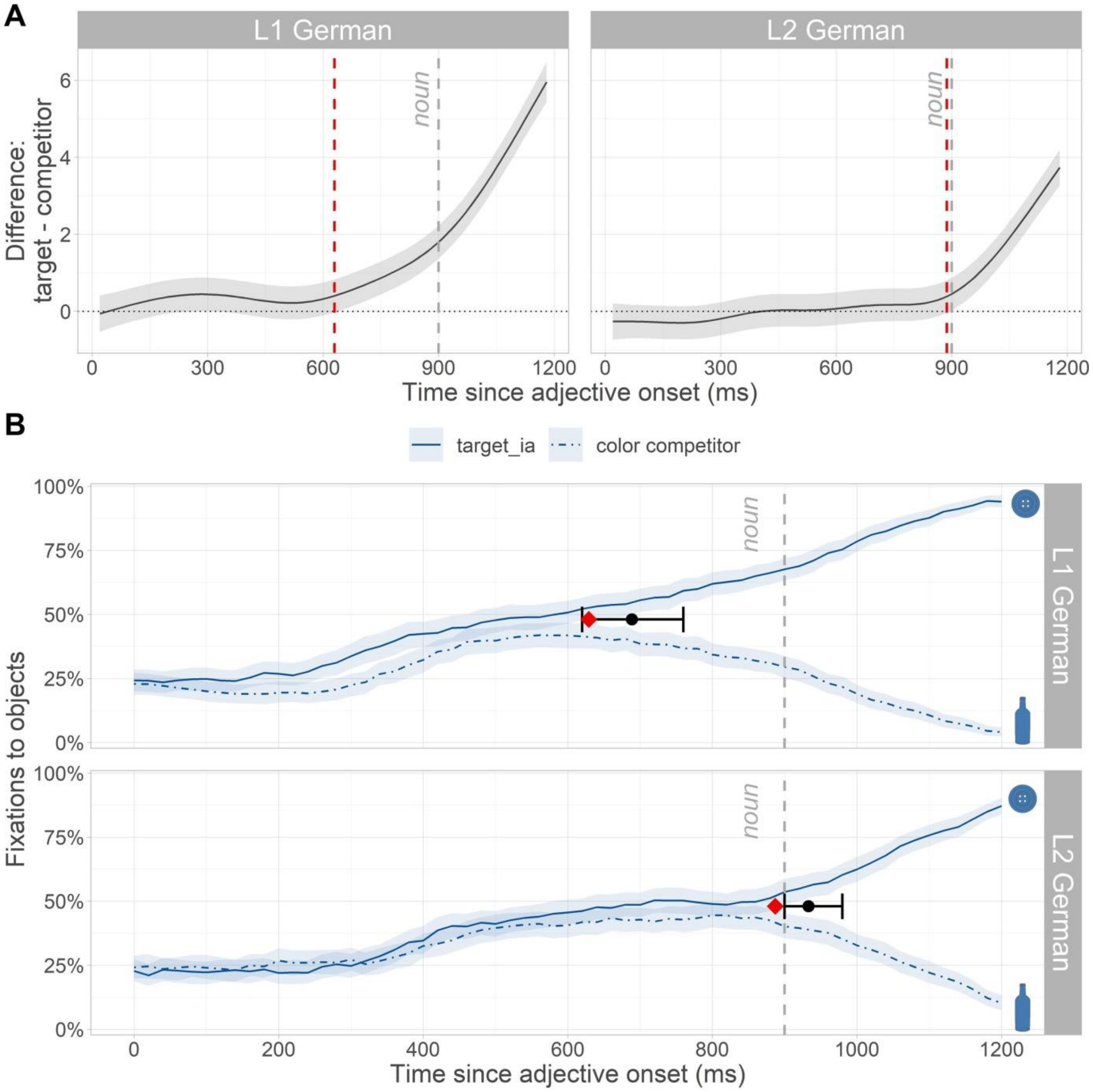

Figure S3. (A) Difference curves with 95% confidence intervals estimated via GAMMs with an optimal random effects structure showing the difference in fixations to the target vs. competitor after adjective onset. The red dashed line indicates the first timepoint in a run of 10 consecutive points at which the lower bound of the confidence interval is greater than zero. The onset of the target noun is displayed 200 ms shifted to the right, to account for the time taken to program and launch an eye movement (Hallett, 1986; Salverda et al., 2014) **(B)** Comparison of GAMM- vs. bootstrap-derived divergence points. The curves show mean fixation proportions with 95% confidence intervals for the target vs. competitor objects. Black points show the bootstrap-derived divergence point estimates with their 95% confidence interval. Red point estimates show the GAMM-derived divergence point estimates, which correspond with the red dashed lines in panel (A). The GAMM-derived estimates are consistent with the bootstrap estimates, as they fall close to the lower bound of the 95% confidence interval of the bootstrap-derived estimates.

S4. Null hypothesis tests of the bootstrapped estimates

In **Section 4.2** of the manuscript, we computed distributions of the difference in divergence point between native (L1) and non-native (L2) German speakers, and between the Spanish and English groups. We determined the statistical significance of the difference between groups by determining whether the 95% confidence interval contained zero. However, if desired, a p-value for the difference distributions can additionally be calculated by creating a bootstrap distribution of the null hypothesis and finding the probability that the bootstrapped divergence point would be found in this null distribution (Efron & Tibshirani, 1993).

To do this, the original fixation data from L1 and L2 speakers (for example) are pooled, their group labels are randomly reassigned, and a difference in divergence point is estimated. The procedure is repeated many times to create a distribution of divergence points that could be expected if there were no true difference between the groups. The p-value is then the proportion of samples from this null distribution that are larger than the bootstrapped divergence point.

In **code §4**, we find that the p-value for the L1-L2 between-group comparison is 0.00. At an alpha of 0.05, this means the difference in divergence points between the L1 and L2 groups is significant, consistent with the inference based on the 95% confidence interval of the bootstrap distribution. For the Spanish-English between-group comparison, the p-value is not significant: 0.79. This is consistent with the inference based on the 95% confidence interval of the bootstrap distribution.

However, we note in **code §4** that the random reassignment of labels has been conducted in such a way that paired comparisons were not possible when conducting statistical tests at each time point. Thus, the statistical test in creating the null distribution differs from that used in the main bootstrapping procedure, and may result in a different null distribution to that which we might see with identical tests. Inferences about statistical significance should therefore not be based solely on p-values computed via this approach, but rather in combination with the 95% CIs from the bootstrapping approach described in the manuscript.